Search Engine Architecture II
Primary Goals of Search Engines

• Effectiveness (quality): to retrieve the most relevant set of documents for a query
  – Process text and store text statistics to improve relevance

• Efficiency (speed): process queries from users as fast as possible
  – Use specialized data structures
Two Major Functions

- Search engine components support two major functions
  - The index process: building data structures that enable searching
  - The query process: using those data structures to produce a ranked list of documents for a user’s query
The Indexing Process

Text Acquisition

Text Transformation

Index Creation

Document data store

Index

E-mail, Web pages, News articles, Memos, Letters
The Query Process
User Interaction

• Providing the interface between users and the search engine

• Tasks
  – Accepting a user query and transforming it into index terms
  – Taking the ranked list of documents from the search engine and organizing it into the results shown to the user
    • Generating snippets to summarize documents
  – Refining a user query to better represent the information need
Query Input

• Query input: providing an interface and a parser for a query language
• Simple languages: some Boolean operations on keywords
• More complex languages may consider issues like proximity of keywords
• [Discussion] What is the language used in Google and Bing? Do you think using punctuations in queries a good idea?
Transformation

- Query transformation: improving the initial query before and after producing a document ranking
- Tokenizing, stopping, and stemming queries
- Spell checking and query suggestion using query logs
- Query expansion based on analysis of term occurrences in documents
- Relevance feedback based on term occurrences in documents that are identified as relevant by a user
Results Output

• Constructing the display of ranked documents
• Generating snippets
• Highlighting important words and passages
• Clustering documents to identify related groups
• Adding appropriate ads
• Translating multiple language documents into the common language
The Query Process

- User Interaction
- Ranking
- Evaluation
- Document data store
- Index
- Log Data
Ranking and Evaluation

• Ranking: taking the transformed query from the user interaction component and generating a ranked list of documents using scores based on a retrieval model
  – The core of a search engine
  – Both effectiveness and efficiency matter

• Evaluation: measuring and monitoring the effectiveness and efficiency
  – Recording and analyzing user behavior using log data
  – Evaluation results are used to tune and improve the ranking component
Scoring

• Assigning a score to a document according to the query and a retrieval model
• Some models are efficient: Boolean model
• Some models are costly: language model – compute the probability that a document is relevant to the query based on the words in the document
Performance Optimization

• Find an efficient way to compute the scores
• Using indexes term by term, term-at-a-time scoring
• Accessing all indexes for the query terms simultaneously, document-at-a-time
• Safe optimizations guarantee that the score calculated will be the same as the scores without optimization
• Unsafe optimization (approximation) may be faster in the tradeoff of accuracy
Distributed Processing

• Queries can be distributed to processors in a network
  – Distributed by a query broker

• Caching: indexes or even ranked document lists from previous queries are left in local memory
  – Reusing previous query results for popular queries
Evaluation Tasks

• Logs: clickthrough data and dwell time (time spent looking at a document)
  – Can be used for spell checking, query suggestions, query caching, mapping ads to users, …

• Ranking analysis: measuring the ranking method and comparing to alternatives
  – Top ranked documents rather than the whole ranked list are more important

• Performance analysis: monitoring and improving overall system performance
  – Response time, throughput, network usage
  – Simulation methods
Summary

- A high-level description of search engine software architecture
- The indexing process and the query process
- Building blocks and their functionalities
Data in Search Engines

- Documents
- Queries
- Logs: history of queries and answers, possibly including user interaction data, such as click-through data
- Search engine task can be tackled using individual types of data and their combinations
- [Discussion] Identify the types of data used in the components of search engine architecture
To-Do List

• Expand the figures of the query process to include the detailed functionalities

• A “more-like-this” query occurs when a user clicks on a particular document in the result list and tells the search engine to find documents that are similar to this one. Describe which low-level components are used to answer this type of query and the sequence they are used in

• Document filtering is an application that stores a large number of queries or user profiles and compares these profiles to every incoming document on a feed. Documents that are sufficiently similar to the profile are forwarded to that person via email or some other mechanism. Describe the architecture of a filtering engine and how it may differ from a search engine

• Read textbook Chapter 2.3.4-2.4