CE306/CE706
Information Retrieval

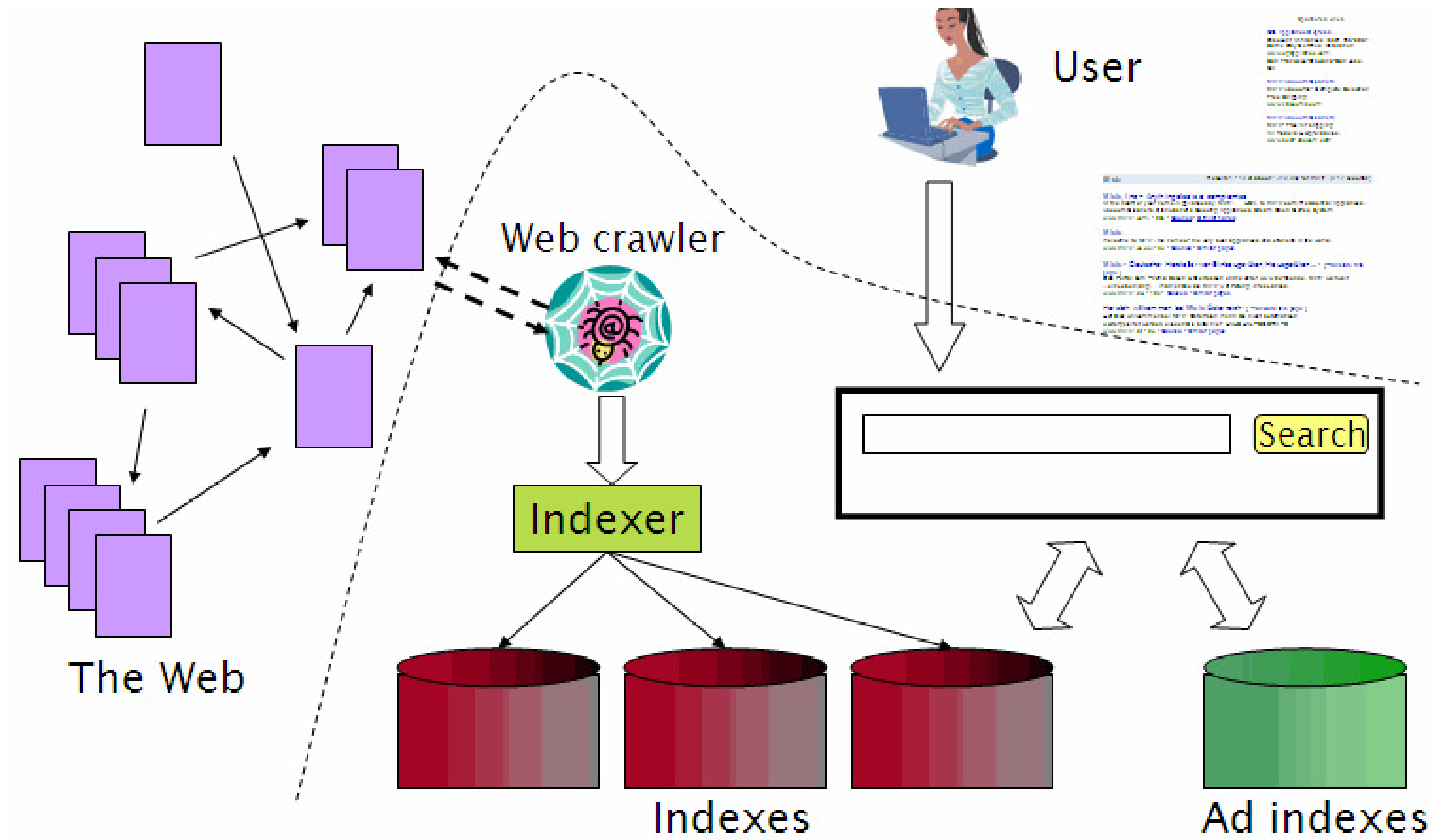
Web Search

Spring 2018

Brief Module Outline (Reminder)

- Motivation + introduction
- Processing pipeline / indexing + query processing
- Large-scale open-source search tools
- Information Retrieval models
- Evaluation
- Log analysis
- User profiles / personalisation / contextualisation
- IR applications, e.g. enterprise search

Web Search Overview



http://nlp.stanford.edu/IR-book/pdf/19web.pdf

Web Search - Motivation

- Web = biggest freely accessible data collection (growing by millions of pages every day)
- Lots of internal structure, e.g.:
  - Mark-up
  - Hyperlinks
  - Meta tags
  - Organisational structure of documents
- Intranets contain even more data than Internet! (enterprise search)

Web Search - General Challenges

- No underlying data model
- Web is very heterogenous:
  - Document size
  - Languages
  - Multimedia
- Quality of data
- Dynamics

Web Search - Basics

- Two ways to search:
  - Index database (e.g. Google, Bing)
  - Web directories (e.g. DMOZ, originally: Yahoo!)
- Typical components of a search engine:
  - Web crawler (spider, robot)
  - Indexing component
  - User interface
- Meta search engines do not need crawler or indexer (e.g. Yippy, Dogpile)

Web Search - Typical Problems

- Query formulation
- Too many matches
- Coverage
- Dead links and duplicated pages

Google - Basics

- Original aim: improving search quality
- Exploits structure in hypertext
- Details in Brin & Page (1998)
- Link structure analysis (PageRank algorithm)
- Associates anchor text with documents
- Exploits visual presentation details
- Combines everything to rank results

PageRank - Motivation 1

- Ranking for a Web page based on the graph of the Web
- Comparable to citation in scientific literature
- Not all citations are equally important
- Page should have high rank if pages with high ranks point to it

PageRank - Motivation 2

- *Random surfer* on the Web
- Starts at random Web page
- Follows links to get to other pages
- After getting bored: jumps to other random pages
- PageRank can be seen as a probability that a *random surfer* visits that page

PageRank - Formula

- Definition of PageRank value for page a:

$$PageRank(a) = \frac{q}{N} + \frac{(1-q)}{N} * \sum_{i=1}^{k} \frac{PageRank(p_i)}{C(p_i)}$$

  - q: probability to jump randomly to a page
  - $(1-q)$: probability following a hyperlink from the current page
  - p_1 ... p_k pages with link to a
  - $C(p_i)$: number of outgoing links from i
  - N: number of Web pages

PageRank - Results

- PageRank calculated in iterative process
- PageRank is query-independent
- Higher quality ordering of search results
- PageRank makes spamming difficult
- Can be used for Web search as well as crawling

Google - Important Data Structures

- Repository of all documents
- Document index
- Lexicon
- Hit lists to record types of matches for each document and word: (1) *Fancy hits* for hits in URL, title, anchor text, meta tag; (2) *Plain hits* for every other match (includes information about position, capitalization etc.)
- Forward index for each doc. to list words and hit lists
- Inverted index to link each word to matching documents: (1) *Short index* for hit lists with title or anchor matches; (2) *Long index* for all hit lists

Google - Search Process

- Try to find matches in short index first (good matches)
- Otherwise try all matches (long index)
- Overall ranking of retrieved documents depends on:
  - Type of matches (e.g. title text, anchor, plain text large font)
  - Number of hits in a document
  - Proximity of query terms in document
  - PageRank value of document
- Literally *hundreds* of other parameters
- No particular factor has too much influence
- But: check out *Google Bombs*

Similar Idea - Clever

- Topic related search
- Find truly relevant Web pages for a query
- Query a standard search engine and rank the results
- Details in Chakrabarti et al. (1999)

How Clever Works

- Query submitted to search engine
- Result pages expanded by adding linked pages
- Calculation of *hub* and *authority* values based on Kleinberg's HITS algorithm (``Hypertext Induced Topic Search'')

Idea of HITS

- There are two types of interesting Web pages:
  - Authorities: relevant for a specific topic
  - Hubs: collections of links to authorities
- Analysis of hyperlinks finds hubs and authorities
- Iterative algorithm

HITS Algorithm (simplified)

- Retrieve set of Web pages for a query
- Initialize non-negative *authority weight* and *hub weight* for each page
- Iteratively increase
  - the *authority weight* of a page depending on the *hub weights* of all pages pointing to it
  - the *hub weight* of a page depending of the *authority weights* it is pointing to
- This procedure converges

PageRank (Google) vs. HITS (Clever)

- Similarities
  - Aim is to find truly relevant Web pages
  - Exploitation of link structure (very different to standard IR)
  - Related to citation analysis
  - Result: more resistant to spamming than standard approaches

PageRank (Google) vs. HITS (Clever)

- Differences:
  - Google assigns global rankings to pages
  - Clever first collects root set of document for a specific query, then ranks them
  - Google only looks in forward direction of links
  - Clever looks both ways introducing *hubs* in addition to *authorities*

Web Search - Some Current Challenges

- Spam
- Content Quality (e.g. *fake news*)
- Quality evaluation
- Web conventions
- Duplicated hosts
- Structure of documents
- Interactive IR, personalization, context, enterprise search, privacy, …
- Details in Baeza-Yates & Ribeiro-Neto (2011), Belkin (2008)

Web Search - Summary

- Explosion of Web triggered enormous interest in IR
- But significant differences between traditional IR data and Web
- Growing interest in hyperlinks and metadata
- Very lively research area
- Current trends: entity search, knowledge graphs, search assistance (search for Daniel Tunkelang's postings on LinkedIn and his blogs to be kept up to date)

Reading

- Chapter 3, Sections 4.4 and 4.5, Section 7.5 in Croft et al. (2015)
- S. Brin & L. Page ``The Anatomy of a Large-Scale Hypertextual Web Search Engine'', In Proceedings of the Seventh World Wide Web Conference (WWW7), Brisbane, 1998.
- S. Chakrabarti et al. ``Hypersearching the Web'', Scientific American, 1999.
- N. J. Belkin ``Some(what) grand challenges for information retrieval'', SIGIR Forum, 42(1), pages 47-54, 2008.