Web Crawling and Basic Text Analysis

Hongning Wang

CS@UVa
Abstraction of search engine architecture

Indexed corpus

Crawler

Doc Analyzer

Doc Representation

Indexer

Index

Ranking procedure

Feedback

Query Rep

(_Query)

Ranker

Evaluation

User

results
Web crawler

• A automatic program that systematically browses the web for the purpose of Web content indexing and updating
  • Synonyms: spider, robot, bot
How does it work

• In pseudo code

```python
Def Crawler(entry_point) {
    URL_list = [entry_point]
    while (len(URL_list) > 0) {
        URL = URL_list.pop();
        if (isVisited(URL) or !isLegal(URL) or !checkRobotsTxt(URL)) continue;
        HTML = URL.open();
        for (anchor in HTML.listOfAnchors()) {
            URL_list.append(anchor);
        }
        setVisited(URL);
        insertToIndex(HTML);
    }
}
```

Which page to visit next?

Is it visited already?

Or shall we visit it again?

Is the access granted?
Visiting strategy

• Breadth first
  – Uniformly explore from the entry page
  – Memorize all nodes on the previous level
  – As shown in pseudo code

• Depth first
  – Explore the web by branch
  – Biased crawling given the web is not a tree structure

• Focused crawling
  – Prioritize the new links by predefined strategies
Focused crawling

• Prioritize the visiting sequence of the web
  – The size of Web is too large for a crawler (even Google) to completely cover
  – Not all documents are equally important
  – Emphasize more on the high-quality documents
• Maximize weighted coverage

In 1999, no search engine indexed more than 16% of the Web
In 2005, large-scale search engines index no more than 40-70% of the indexable Web

![Graph showing weighted coverage till time t and importance of page p]

Weighted coverage till time \( t \)
Pages crawled till time \( t \)
Focused crawling

- Prioritize by in-degree \([\text{Cho et al. WWW'98}]\)
  - The page with the highest number of incoming hyperlinks from previously downloaded pages is downloaded next

- Prioritize by PageRank \([\text{Abiteboul et al. WWW'07, Cho and Uri VLDB'07}]\)
  - Breadth-first in early stage, then compute/approximate PageRank periodically
  - More consistent with search relevance \([\text{Fetterly et al. SIGIR'09}]\)
Focused crawling

• Prioritize by topical relevance
  – In vertical search, only crawl relevant pages [De et al. WWW’94]
    • E.g., restaurant search engine should only crawl restaurant pages
  – Estimate the similarity to current page by anchortext or text near anchor [Hersovici et al. WWW’98]
  – User given taxonomy or topical classifier [Chakrabarti et al. WWW’98]
Avoid duplicate visit

• Given web is a graph rather than a tree, avoid loop in crawling is important

• What to check
  – URL: must be normalized, not necessarily can avoid all duplication
    • http://dl.acm.org/event.cfm?id=RE160&CFID=516168213&CFTOKEN=99036335
    • http://dl.acm.org/event.cfm?id=RE160
  – Page: minor change might cause misfire
    • Timestamp, data center ID change in HTML

• How to check
  – trie or hash table
Politeness policy

• Crawlers can retrieve data much quicker and in greater depth than human searchers

• Costs of using Web crawlers
  – Network resources
  – Server overload

• Robots exclusion protocol
  – Examples: CNN, UVa
Robot exclusion protocol examples

• Exclude specific directories:
  
  User-agent: *
  Disallow: /tmp/
  Disallow: /cgi-bin/
  Disallow: /users/paranoid/

• Exclude a specific robot:
  
  User-agent: GoogleBot
  Disallow: /

• Allow a specific robot:
  
  User-agent: GoogleBot
  Disallow: 

  User-agent: *
  Disallow: /
Re-visit policy

• The Web is very dynamic; by the time a Web crawler has finished its crawling, many events could have happened, including creations, updates and deletions
  – Keep re-visiting the crawled pages
  – Maximize freshness and minimize age of documents in the collection

• Strategy
  – Uniform re-visiting
  – Proportional re-visiting
    • Visiting frequency is proportional to the page’s update frequency
Analyze crawled web pages

- What you care from the crawled web pages
Analyze crawled web pages

• What machine knows from the crawled web pages
Basic text analysis techniques

• Needs to analyze and index the crawled web pages
  – Extract informative content from HTML
  – Build machine accessible data representation
HTML parsing

• Generally difficult due to the free style of HTML

• Solutions
  – Shallow parsing
    • Remove all HTML tags
    • Only keep text between <title></title> and <p></p>
  – Automatic wrapper generation [Crescenzi et al. VLDB’01]
    • Wrapper: regular expression for HTML tags’ combination
    • Inductive reasoning from examples
  – Visual parsing [Yang and Zhang DAR’01]
    • Frequent pattern mining of visually similar HTML blocks
HTML parsing

- **jsoup**
  - Java-based HTML parser
    - scrape and parse HTML from a URL, file, or string to DOM tree

- Python version: **Beautiful Soup**

- CSS

- `e()`
How to represent a document

• Represent by a string?

<HEAD>Crowds in Liverpool to Mark 10th Anniversary of John Lennon's Death</HEAD>
<DateLINE>LIVERPOOL, England (AP) </DateLINE>
<TEXT>

Dozens of fans of rock legend and former Beatle John Lennon gathered in the snow on a windy Saturday for a ceremony marking the 10th anniversary of his death. Liverpool's mayor, Dorothy Gavin, led Lennon devotees who laid wreaths at the foot of a bronze statue of The Beatles in the city's Cavern Walks shopping center. The center was built on the original site of the Cavern Club, made famous when The Beatles played there in the 1960s, and has become a place of pilgrimage. "Give peace a chance," the title of one of singer-songwriter Lennon's greatest hits, was the theme for the day.

Lennon and his wife, Yoko Ono, were returning to their apartment in New York's Dakota apartment building after a recording session on Dec. 8, 1980, when Lennon was shot to death by Mark David Chapman, a deranged fan to whom Lennon had given his autograph only hours before. Lennon was 40. A spokesman for the Lennon family said Ms. Ono and the couple's son, Sean, were in Europe and would spend the anniversary privately.

Peebles said late in 1980 that Lennon had just recovered from a period when he had "gone off the rails" and his relationship with Ms. Ono had suffered. "But (when I saw him) they'd had the baby, Sean had been born, and everything was great."

— Bag-of-Words representation!
Tokenization

• Break a stream of text into meaningful units
  – Tokens: words, phrases, symbols
    • **Input**: It’s not straight-forward to perform so-called “tokenization.”
    • **Output(1)**: 'It’s', 'not', 'straight-forward', 'to', 'perform', 'so-called', ‘“tokenization.”’
    • **Output(2)**: 'It', '', 's', 'not', 'straight', '-', 'forward', 'to', 'perform', 'so', '-', 'called', ‘“’, 'tokenization', '.', ‘”’
  – Definition depends on language, corpus, or even context
Tokenization

• Solutions
  – Regular expression
    • \[w\]+: so-called -> ‘so’, ‘called’
    • \[S\]+: It’s -> ‘It’s’ instead of ‘It’, “s’
  – Statistical methods
    • Explore rich features to decide where is the boundary of a word
      – Apache OpenNLP (http://opennlp.apache.org/)
    • Online Demo
      – UIUC (http://cogcomp.cs.illinois.edu/curator/demo/index.html)
Full text indexing

• Bag-of-Words representation
  – Doc1: Information retrieval is helpful for everyone.
  – Doc2: Helpful information is retrieved for you.

<table>
<thead>
<tr>
<th></th>
<th>information</th>
<th>retrieval</th>
<th>retrieved</th>
<th>is</th>
<th>helpful</th>
<th>for</th>
<th>you</th>
<th>everyone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doc2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Word-document adjacent matrix
Full text indexing

• Bag-of-Words representation
  – Assumption: word is independent from each other
  – Pros: simple
  – Cons: grammar and order are missing
  – The most frequently used document representation
    • Image, speech, gene sequence
Full text indexing

• Improved Bag-of-Words representation
  – N-grams: a contiguous sequence of n items from a given sequence of text
    • E.g., Information retrieval is helpful for everyone
    • Bigrams: ‘information_retrieval’, ‘retrieval_is’, ‘is_helpful’, ‘helpful_for’, ‘for_everyone’
  – Pros: capture local dependency and order
  – Cons: purely statistical view, increase vocabulary size $O(V^N)$
Full text indexing

• Index document with all the occurring word
  – Pros
    • Preserve all information in the text (hopefully)
    • Fully automatic
  – Cons
    • Vocabulary gap: cars v.s., car
    • Large storage: e.g., in N-grams $O(V^N)$
  – Solution
    • Construct controlled vocabulary
Statistical property of language

- **Zipf’s law**
  - Frequency of any word is inversely proportional to its rank in the frequency table.
  - Formally:
    \[
    f(k, s, N) = \frac{1}{k^s} \quad \text{where } s \text{ is a language-specific parameter.}
    \]

A plot of word frequency in Wikipedia (Nov 27, 2006)

In the Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences; the second-place word "of" accounts for slightly over 3.5% of words.
Zipf’s law tells us

• Head words may take large portion of occurrence, but they are semantically meaningless
  – E.g., the, a, an, we, do, to

• Tail words take major portion of vocabulary, but they rarely occur in documents
  – E.g., dextrosinistral

• The rest is most representative
  – To be included in the controlled vocabulary
Automatic text indexing

Remove non-informative words

Remove rare words

Figure 2.1. A plot of the hyperbolic curve relating \( f \), the frequency of occurrence and \( r \), the rank order (Adapted from Schultz, page 4221)
Normalization

- Convert different forms of a word to normalized form in the vocabulary
  - U.S.A -> USA, St. Louis -> Saint Louis

- Solution
  - Rule-based
    - Delete periods and hyphens
    - All in lower case
  - Dictionary-based
    - Construct equivalent class
      - Car -> “automobile, vehicle”
      - Mobile phone -> “cellphone”
Stemming

• Reduce inflected or derived words to their root form
  – Plurals, adverbs, inflected word forms
    • E.g., ladies -> lady, referring -> refer, forgotten -> forget
  – Bridge the vocabulary gap
  – Risk: lose precise meaning of the word
    • E.g., lay -> lie (a false statement? or be in a horizontal position?)
  – Solutions (for English)
    • Porter stemmer: pattern of vowel-consonant sequence
    • Krovetz Stemmer: morphological rules
Stopwords

- Useless words for query/document analysis
- Not all words are informative
- Remove such words to reduce vocabulary size
- No universal definition
- Risk: break the original meaning and structure of text

E.g., this is not a good option
- to be or not to be
- null

The OEC: Facts about the language

The OEC: Facts about the language
Abstraction of search engine architecture

Indexed corpus

1. Visiting strategy
2. Avoid duplicated visit
3. Re-visit policy

Crawler

Doc Analyzer

1. HTML parsing
2. Tokenization
3. Stemming/normalization
4. Stopword/controlled vocabulary filter

Doc Representation

BagOfWord representation!
Automatic text indexing

• In modern search engine
  – **No** stemming or stopword removal, since computation and storage are no longer the major concern
  – More advanced NLP techniques are applied
    • Named entity recognition
      – E.g., people, location and organization
    • Dependency parsing

Query: “to be or not to be”
What you should know

- Basic techniques for crawling
- Zipf’s law
- Procedures for automatic text indexing
- Bag-of-Words document representation
Reference I


Reference II