Informatics 225
Computer Science 221

Information Retrieval

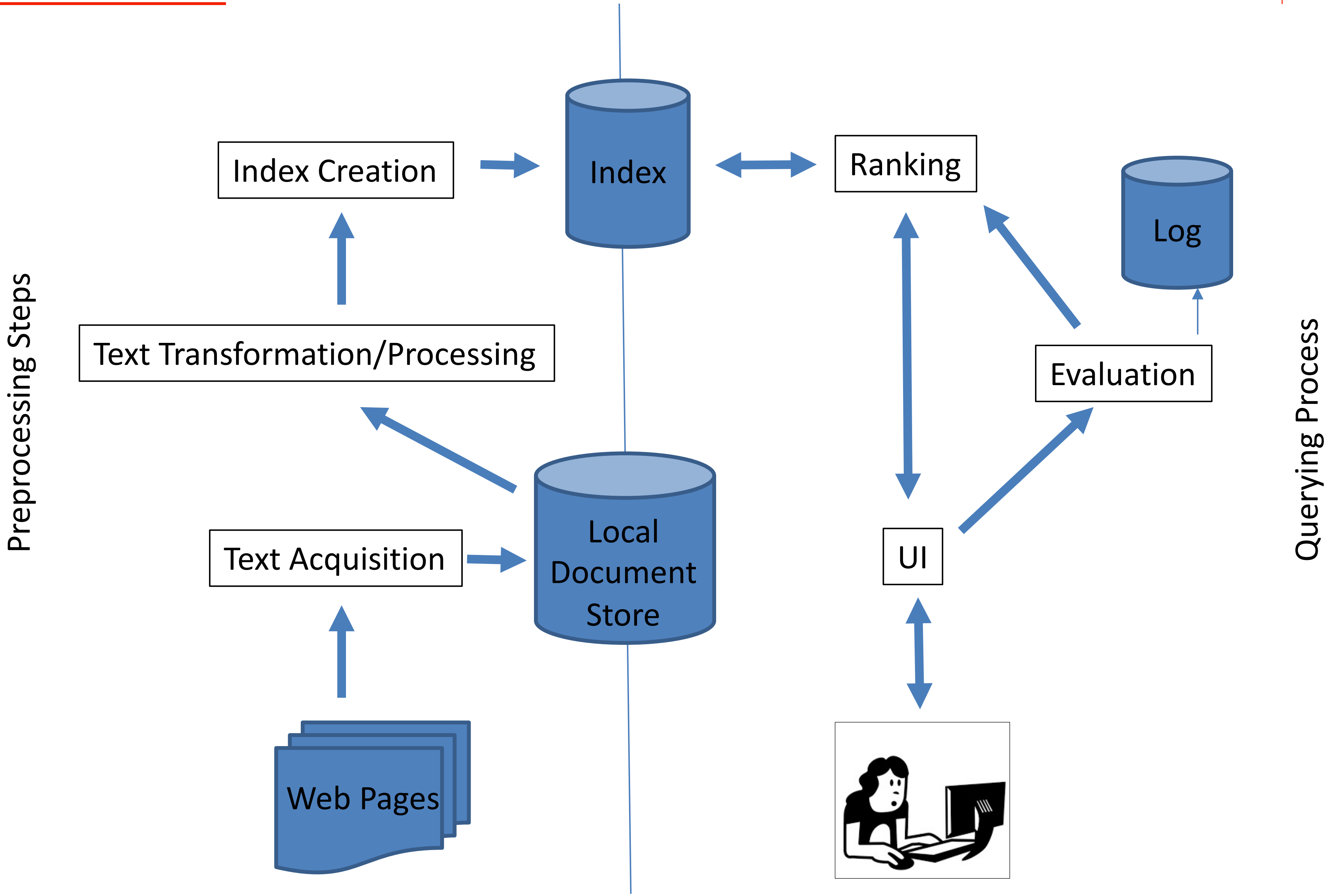
Lecture 4

These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Prof. Alberto Krone-Martins, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.

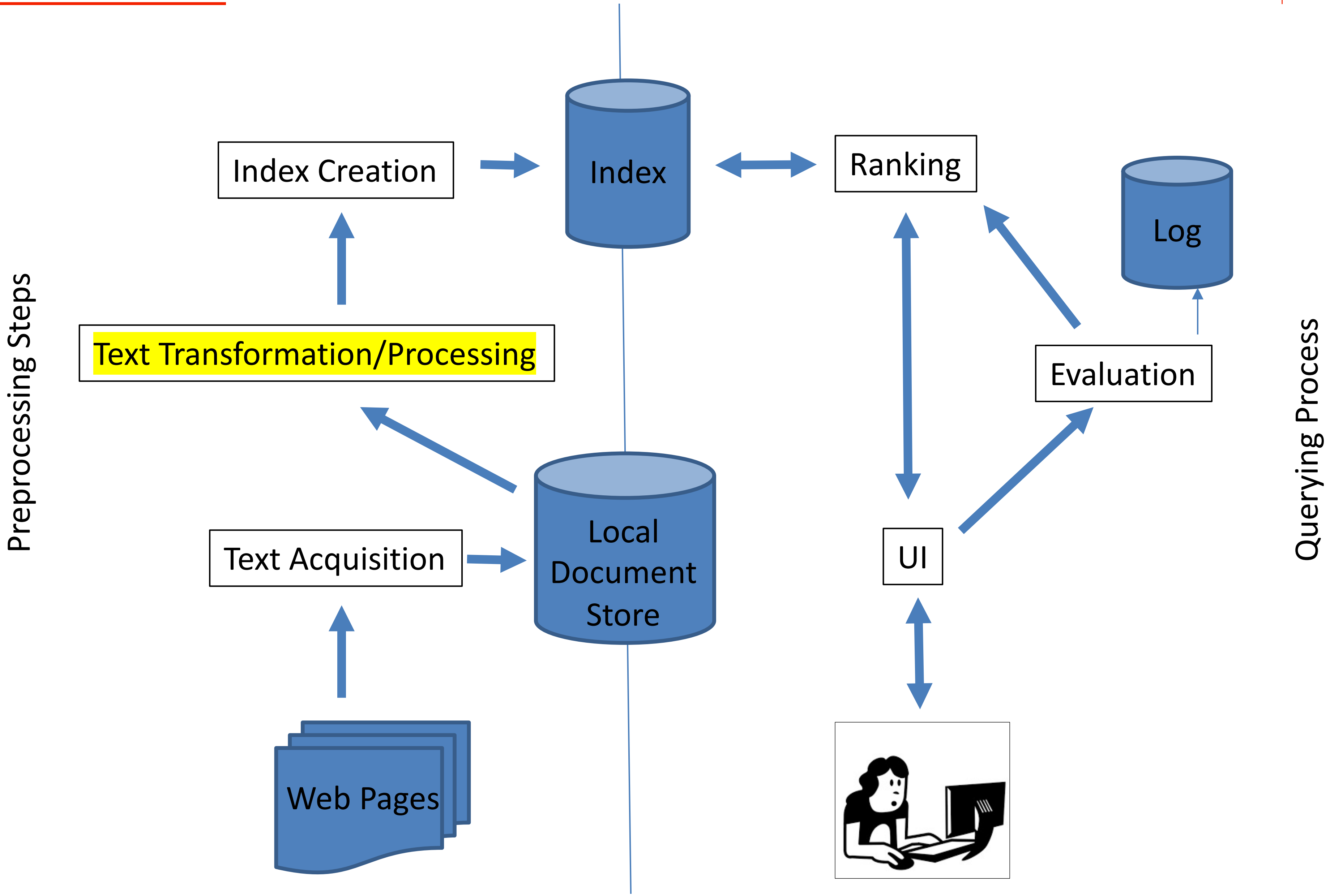
Tokenization
Part 1

Information Retrieval

Architecture : Tokenization?

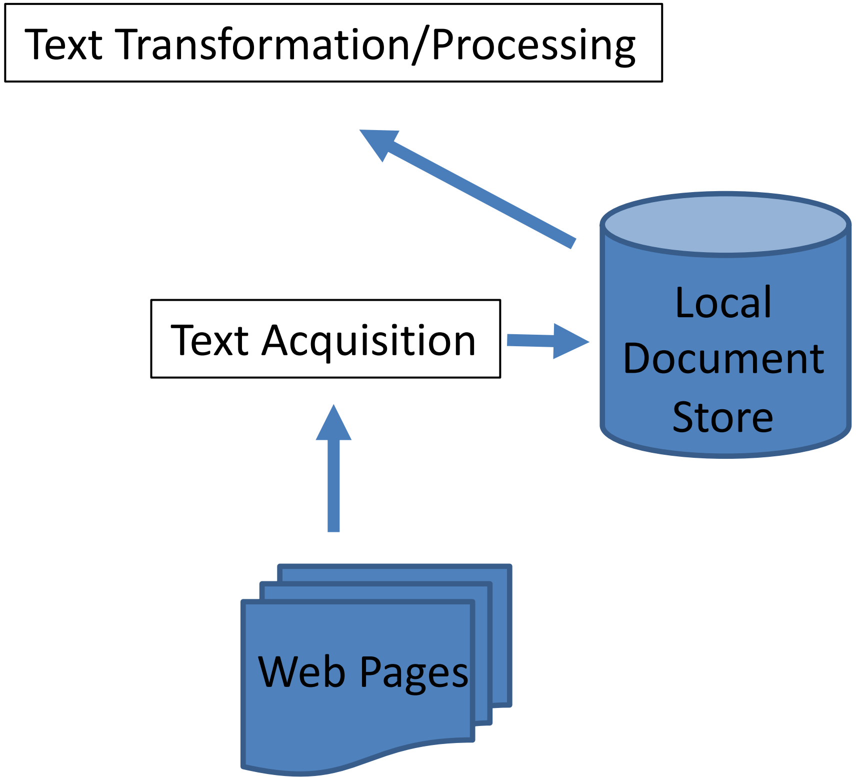


Architecture : Text Transformation and Processing



Corpus

- The complete set of documents that are available to be searched



Corpus

- The complete set of documents that are available to be searched

- In a websearch context:
  - The full set of pages that will be considered by the search engine

Text Processing

- The step/stage of processing the documents/pages in your corpus for use in your search engine

- Occurs prior to building your index to improve the efficiency/performance

Text Processing

- Typically 2 main steps in text processing:
  - **Tokenization (lexical analysis)**
    - Breaking the text into tokens
  - **Linguistic pre-processing**
    - Applying rules to the tokens to improve the efficiency of the index

Tokenization

- **What is tokenization?**

Tokenization

- Break the input into simple units of meaning
  - Can be a word or not! Depends on the choice of how to create tokens…
  - Character stream -> token stream
  - Called a tokenizer / lexer / scanner

- Compiler front-end
  - Tokenizer / lexer / scanner may hook up to a parser

- Preprocessor for information retrieval
  - Tokenizer / lexer / scanner feeds tokens to retrieval system

Identifying Tokens

- Divide on whitespace and throw away punctuation?

- What is a token? Depends…
  - Apostrophes
  - O’Neill
  - Aren’t

- Hyphen-handling
  - clear-headed vs clearheaded
  - mother-in-law

Identifying Tokens

- Multiple words as single token?
  - San Francisco
  - white space
  - New York University vs York University
- Tokens that aren’t words…
  - info@uci.edu
  - http://www.ics.uci.edu/
  - 192.168.0.1

Identifying Tokens

- **Early** tokenization methodology:
  - A sequence of 3 or more alphanumeric characters
  - A space or special character indicates the end of the token
  - All characters were converted to lower case

Identifying Tokens

- Early tokenization methodology:
  - Example:

“Bigcorp's 2007 bi-annual report showed profits rose 10%.”

Identifying Tokens

- Early tokenization methodology:
  - Example:

“Bigcorp's 2007 bi-annual report showed profits rose 10%.”

↓

“bigcorp 2007 annual report showed profits rose”

- Issues:
  - Too generic and simple : *resulting in a lot of lost information*

Assignment 1 : Tokenizer from scratch

You should write the tokenizer in Python

(3.6+; but **preferably 3.6 because this is what you will have in the openlab.ics.uci.edu machines**).

This will help you with your next two assignments!

Assignment 1 : Tokenizer from scratch

Very important: **At certain points, the assignment may seem underspecified – this is by design.**

In those cases, make your own choices and assumptions and be prepared to defend them.

Assignment 1 : Tokenizer from scratch

Your program must run!

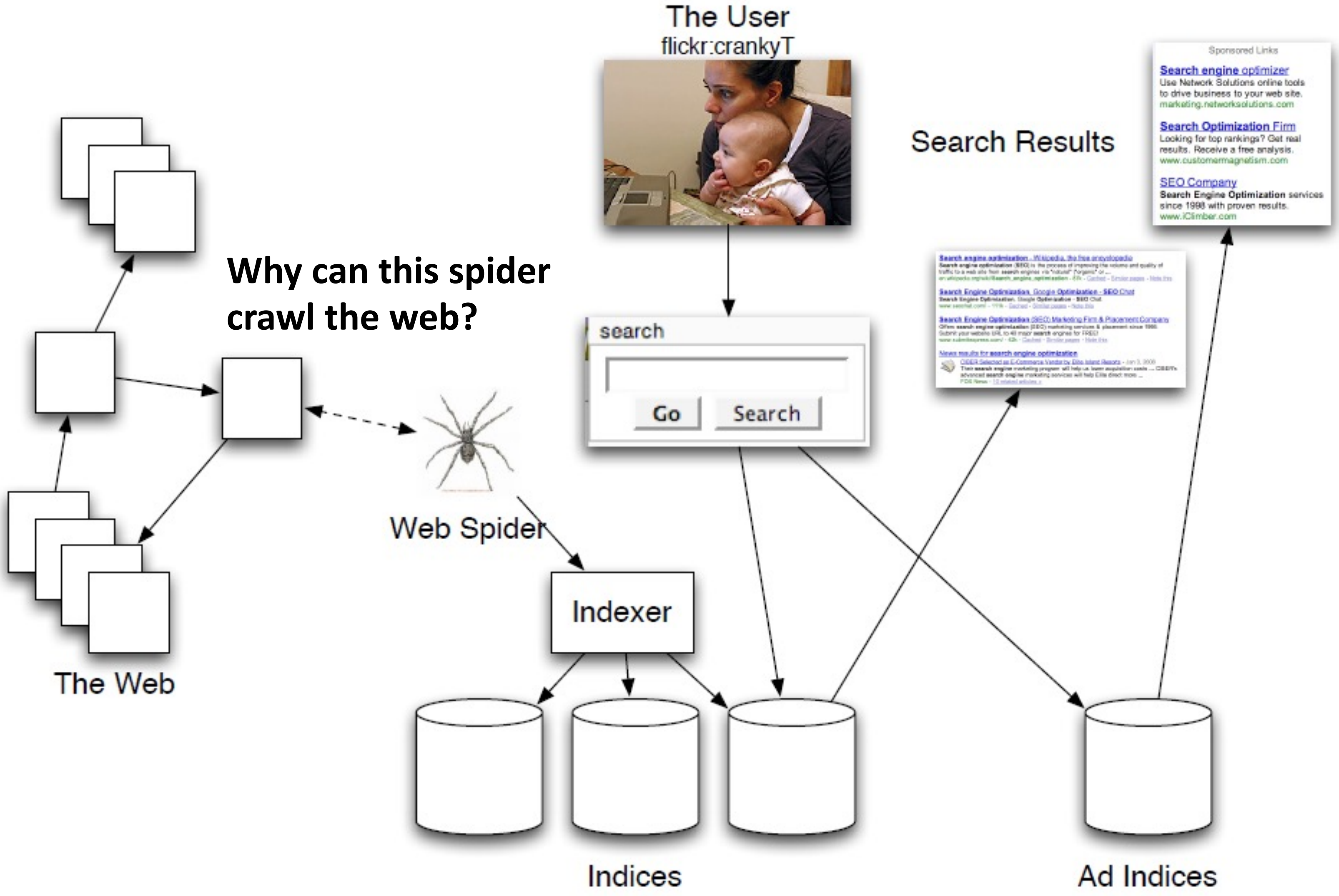
It must be executable from the command line.

You should get the file names from command line arguments.

A bit further on the web

Information Retrieval

Web Search Engine



Sponsored Links

Search engine optimizer
Use Network Solutions online tools to drive business to your web site.
marketing.networksolutions.com

Search Optimization Firm
Looking for top rankings? Get real results. Receive a free analysis.
www.customermagnetism.com

SEO Company
Search Engine Optimization services since 1998 with proven results.
www.iClimber.com

Search engine optimization - Wikipedia, the free encyclopedia
Search engine optimization (SEO) is the process of improving the volume and quality of traffic to a web site from search engines via "natural" ("organic" or ...
en.wikipedia.org/wiki/Search_engine_optimization - 89k - Cached - Similar pages - Note this

Search Engine Optimization, Google Optimization - SEO Chat
Search Engine Optimization, Google Optimization - SEO Chat
www.seochat.com/ - 111k - Cached - Similar pages - Note this

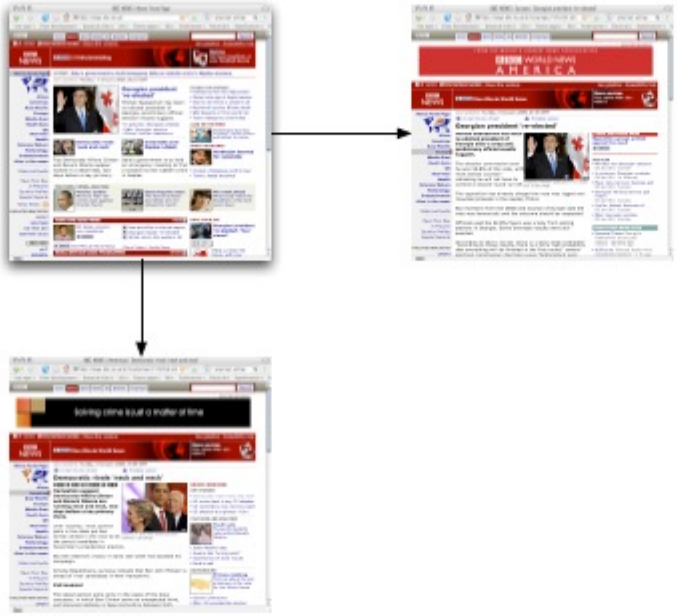
Search Engine Optimization (SEO) Marketing Firm & Placement Company
Offers search engine optimization (SEO) marketing services & placement since 1998. Submit your website URL to 40 major search engines for FREE!
www.submitexpress.com/ - 42k - Cached - Similar pages - Note this

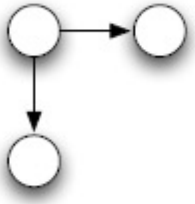
News results for search engine optimization
OBEER Selected as E-Commerce Vendor by Elite Island Resorts - Jan 3, 2008
Their search engine marketing program will help us lower acquisition costs ... OBEER's advanced search engine marketing services will help Elite direct more ...
PDB News - 12 related articles »

Why can this spider crawl the web?

The Web Graph

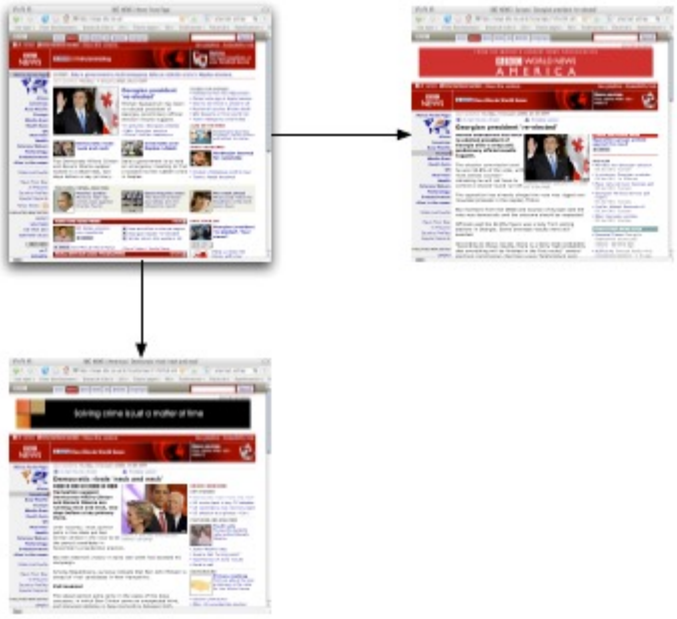
- The Web is a graph
    - Pages are nodes
    - Hyperlinks are directed edges

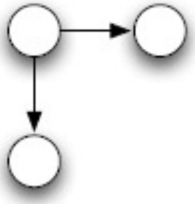




The Web Graph

- The Web is a graph
  - Pages are nodes
  - Hyperlinks are directed edges



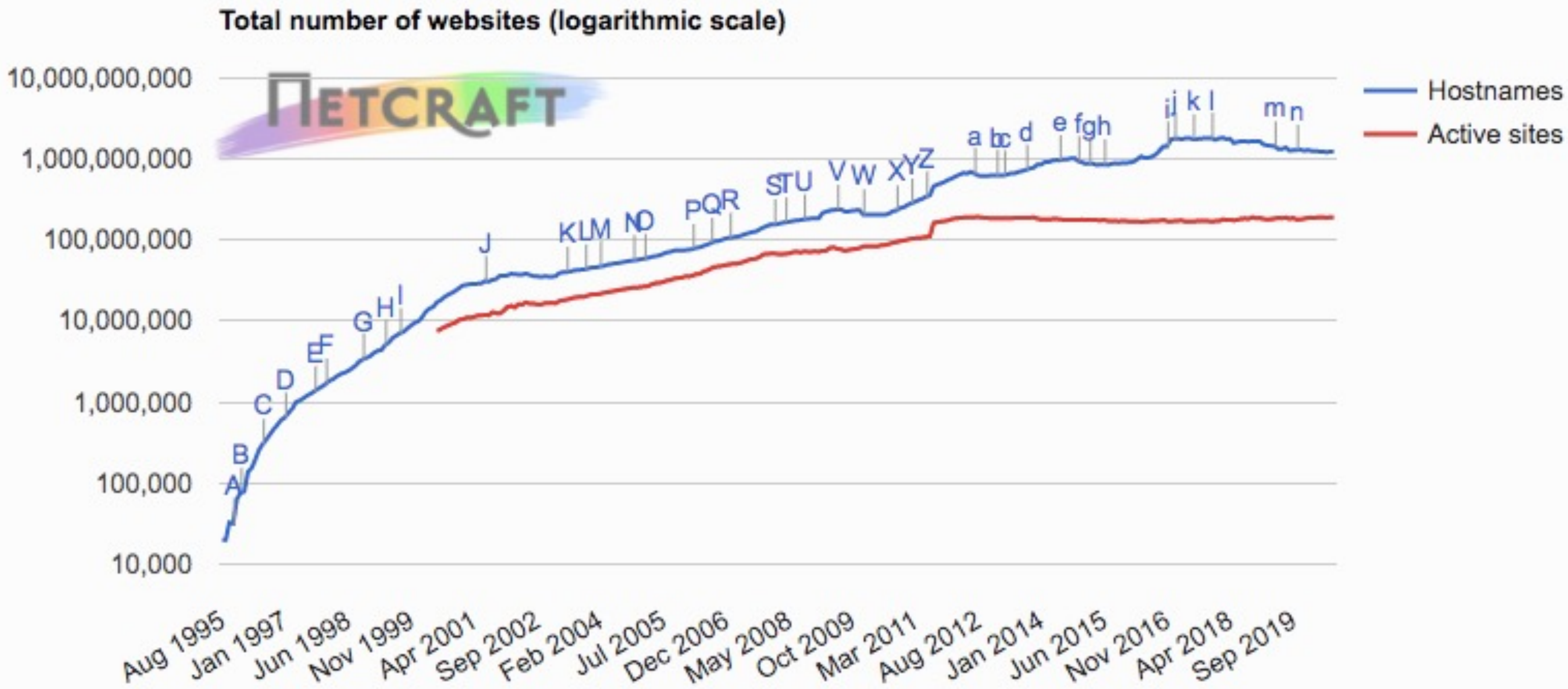


What are the characteristics of this graph?

Characteristics

- Significant duplication
  - 30%-40% in some studies [Brod97, Shiv99]
  - 25%-30% Google's number from 2013 [https://www.youtube.com/watch?v=mQZY7EmjbMA]
  - www.copyscape.com
- High linkage
  - More than 8 links per page
- Very large
  - *How large?*

Web Characteristics: Size



https://news.netcraft.com/archives/2020/09/23/september-2020-web-server-survey.html

Characteristics

- Significant duplication
  - 30%-40% in some studies [Brod97, Shiv99]
  - 25%-30% Google's number from 2013 [https://www.youtube.com/watch?v=mQZY7EmjbMA]
  - www.copyscape.com
- High linkage
  - More than 8 links per page
- Very large
  - *Hosts* O(10^9)
  - *Pages O(???)*

Characteristics

- Significant duplication
  - 30%-40% in some studies [Brod97, Shiv99]
  - 25%-30% Google's number from 2013 [https://www.youtube.com/watch?v=mQZY7EmjbMA]
  - www.copyscape.com
- High linkage
  - More than 8 links per page
- Very large
  - *Hosts* O(10^9)
  - *Pages O(???)*
- Spam, misleading and false information
  - O($10^{8\text{-}9?}$) of pages of it

Characteristics

- High rate of change
  - – [Cho00] 720k pages from 270 popular sites sample daily for 5 months in 1999
    - **?? % changed weekly, ?? % daily**

How much do you think?

Characteristics

- High rate of change
  - [Cho00] 720k pages from 270 popular sites sample daily for 5 months in 1999
    - **40% changed weekly, 23% daily**

Characteristics

- High rate of change
  - [Cho00] 720k pages from 270 popular sites sample daily for 5 months in 1999
    - **40% changed weekly, 23% daily**
  - [Fett02] 151M pages checked over a few months
    - **Significant changes: 7% weekly**
    - **Some changes: 25% weekly**

Characteristics

- High rate of change
  - [Cho00] 720k pages from 270 popular sites sample daily for 5 months in 1999
    - **40% changed weekly, 23% daily**
  - [Fett02] 151M pages checked over a few months
    - **Significant changes: 7% weekly**
    - **Some changes: 25% weekly**
  - [Ntul04] 154 **large** sites recrawled from scratch weekly
    - **8% had new pages every week**
    - **8% die**
    - **5% new content**
    - **25% new links**

The Web: Evolution

- 1st phase: static content (documents)
- 2nd phase: dynamic content (applications)
- 3rd phase: user-generated content (social)
- 4th phase: mobile
- 5th phase: ??? AR? VR? Brain-interfaces?

The Web: Evolution

- 1st phase: static content (documents)
- 2nd phase: dynamic content (applications)

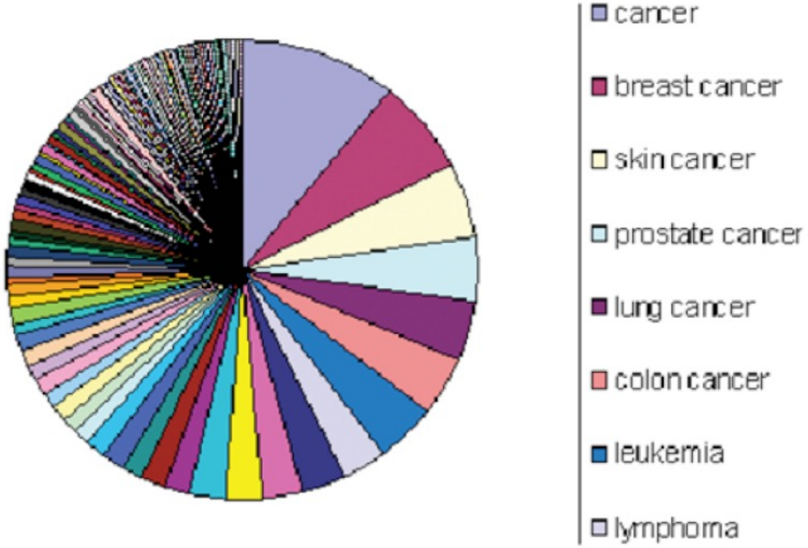
One need has been kept invariant: to find information

Present-day solution : IR Systems for SEARCH and RANKING

- 4th phase: mobile
- 5th phase: ??? AR? VR? Brain-interfaces?

The Web: Web search characteristics

- Few popular broad queries
- Many rare specific queries
  - E.g. distribution of all cancer related searches:



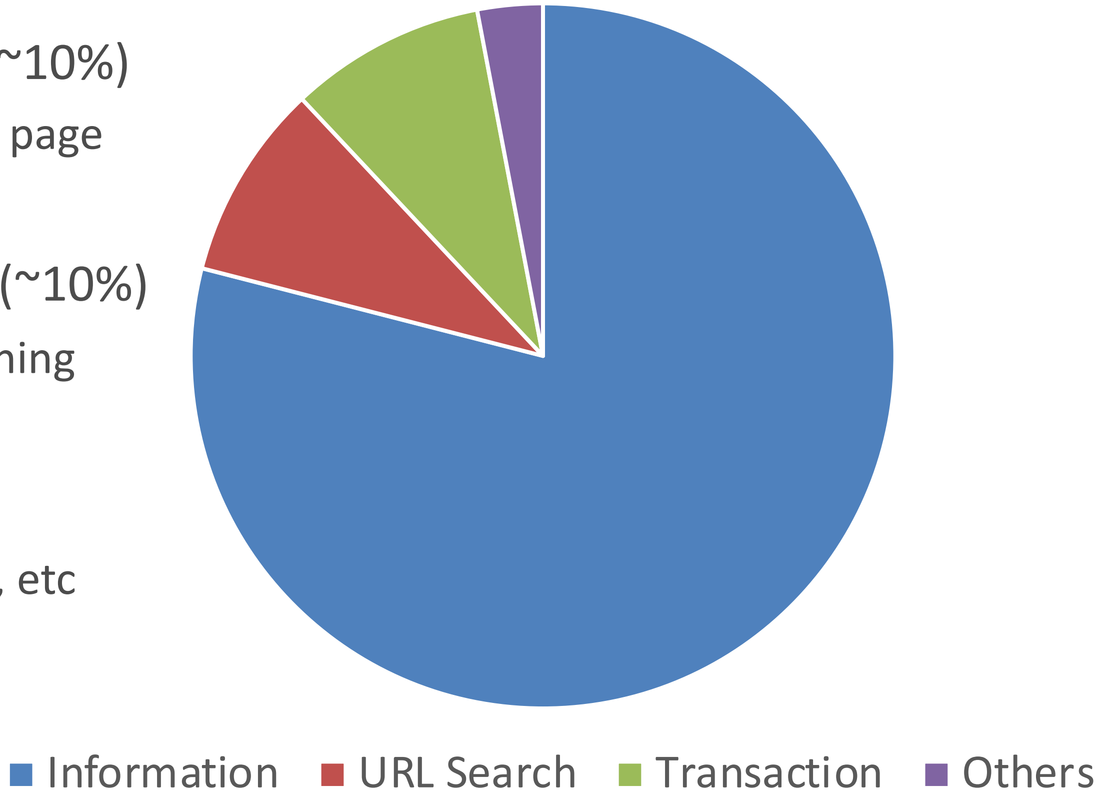
Web Users: Popular Searches

- Term Interest can be explored at google trends
  - https://www.google.com/trends/

- A search engine uses trend information to
  - Auto-complete:
    - Enhance user experience;
    - Correct spelling errors;
  - Optimize search results
    - Trending “near you” can help in relevance scoring;
  - Smart-cache search results
    - If some search is trending, keep results in cache!

Web Users: Motivations

- [Jansen et al. 2008]
  - Informational needs (~80%)
    - Want to learn about something
  - Navigational needs (~10%)
    - Want to go to that page
  - Transactional needs (~10%)
    - Want to do something
  - Miscellaneous
    - Exploration, social, etc



Web Users: Query formation

- Most queries are **ill-defined** queries

Web Users: Query formation

- Most queries are **ill-defined** queries
  - Do you have any idea of the size of queries?

Web Users: Query formation

- Most queries are **ill-defined** queries
  - Short (80% < 3 words)
  - Imprecise terms
  - No logical operators

Web Users: Query formation

- Most queries are **ill-defined** queries
  - Short (80% < 3 words)
  - Imprecise terms
  - No logical operators
  - Low effort: spelling mistakes

Web Users: Characteristics

- Wide variance in
  - Needs
  - Expectations
  - Knowledge

Web Users: Characteristics

- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth

Web Users: Characteristics

- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth

Your search engine will be better if you take these points seriously

Web Users: Characteristics

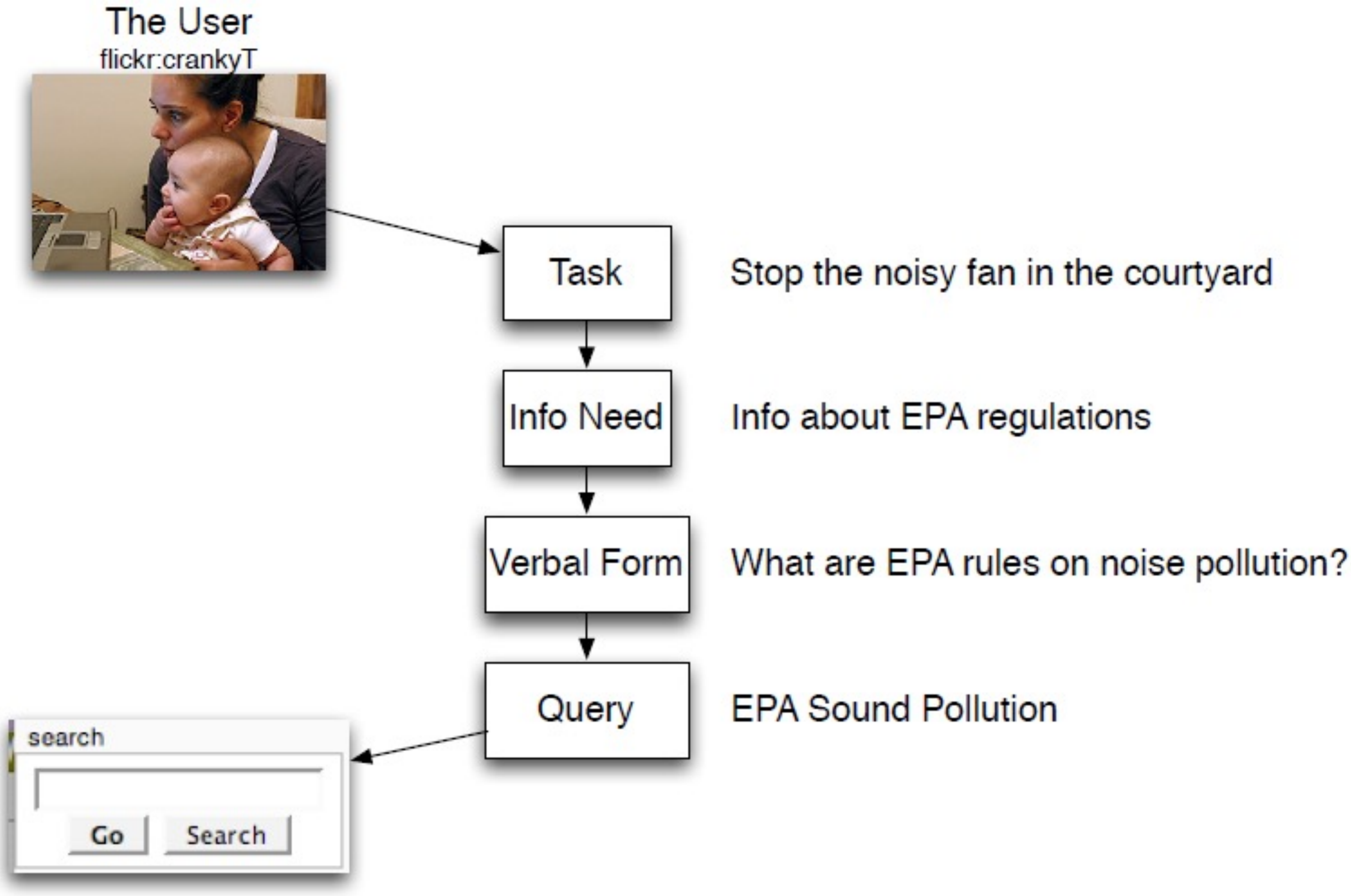
- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth
- Behavior
  - 85% look over one result screen only **(!!!)**

Web Users: Characteristics

- Wide variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth

- Behavior
  - 85% look over one result screen only **(!!!)**
  - 78% of queries are not modified
  - Follow links ("The scent of information")

Information need pipeline



The User
flickr:crankyT

Task	Stop the noisy fan in the courtyard
Info Need	Info about EPA regulations
Verbal Form	What are EPA rules on noise pollution?
Query	EPA Sound Pollution

Answering the need behind the query

- Query is often imprecise indicator of what the user wants
- What can we do to improve this?

Answering the need behind the query

- Query is often imprecise indicator of what the user wants

- What can we do to improve this?
  - Design retrieval models and mathematical definitions for the Relevance Score that are generic, albeit not ideal for all possible “specific topics”.
  - User context: who, where, what
  - Guess the type of information (image, map, math, etc.)
  - Correct and/or expand queries

Final announcements

- Remember to book the office hours if you need/want to!
  - I can try to find alternative slots if it is necessary!