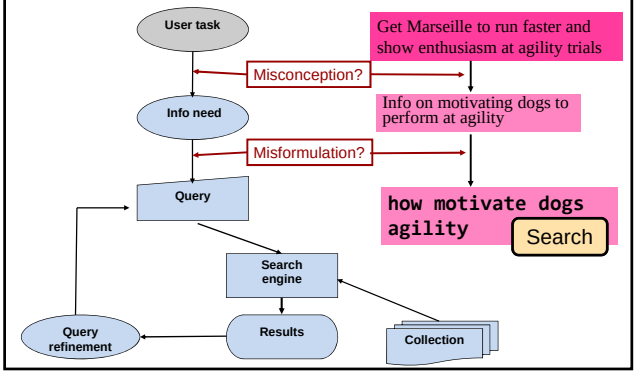
Information Retrieval
INFO 4300 / CS 4300

- Instructor: Claire Cardie
  - Professor in CS and IS (and CogSci)
- Three TAs at last count
  - Liz Murnane
  - Jon Park
  - Chenhao Tan
- One dog
  - Marseille (mahr-say)

Last class

- Classic search model
- Definitions of IR
- IR applications
- Cornell connections(!!)

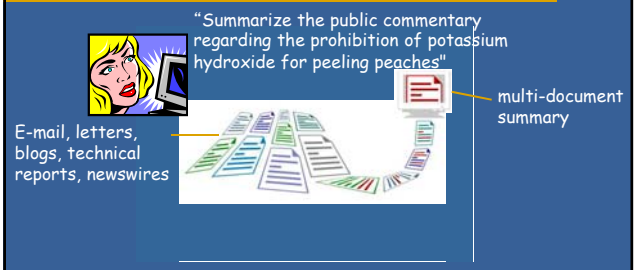
The Classic Search Model



Croft, Metzler & Strohman (2010)

- *"Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."* (Salton, 1968)
- General definition that can be applied to many types of information and search applications

IR applications: E-Rulemaking



Many Cornell Connections

- Gerard Salton
  - Father of IR
  - Co-founded our CS department

- Amit Singhal
  - PhD student of Salton's
  - Head of "search" at Google
  - Totally rewrote the search code at Google in 2001

Course Goals

- To help you to understand **search engines**, evaluate and compare them, and modify them for specific applications
- Provide broad coverage of the important issues in **information retrieval and search engines**
  - includes underlying (mathematical) models and current research directions

Topics for Today

- Big issues in IR: revisited
- Search engine architecture
  - Issues for each component

Big Issues in IR

- Relevance
  - A *relevant document* contains the info that a person was looking for when he/she submitted the query.
- Sounds simple.
  - Vocabulary mismatch
  - Topical relevance vs. user relevance

Addressing relevance

- ***Retrieval models*** define a view of relevance
  - Formal representation of the process of matching a query to a document
  - The basis of *ranking algorithms* used in search engines
- Need to account for user relevance
- Model the statistical properties of language (e.g. word counts) rather than linguistic properties (e.g. adjective/noun counts) --- since 1950s
  - This view of text wasn't popular in NLP until the 1990s.

Big Issues in IR

- Evaluation
  - Long tradition (since 1960s) of using empirical procedures and evaluation measures to compare system output with user expectations
    - » Precision
    - » Recall --- problem for web search?
  - Often use *test collections:* documents, typical queries, and *relevance judgments*
    - » Most commonly used are TREC collections
  - Clickthrough data

Big Issues in IR

- Users and their information needs
  - Search evaluation is necessarily user-centered
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as *query suggestion*, *query expansion*, *relevance feedback* improve ranking

IR and Search Engines

- A **search engine** is the practical application of IR techniques to large-scale text collections
  - Web search engines are best-known examples
- Big issues include main IR issues but also some others
  - Performance
  - Dynamic data
  - Scalability
  - Adaptability
  - Specific problems (e.g. spam)

In-Class Exercise

- Name some web services or sites that appear to use search (not including web search engines)

In-Class Exercise

- Precision/Recall