Queries

Lecture 4
CS 410/510
Information Retrieval on the Internet

Queries

- Query
  - An expression of the user's information need that is input to an IR system
- What's an information need?

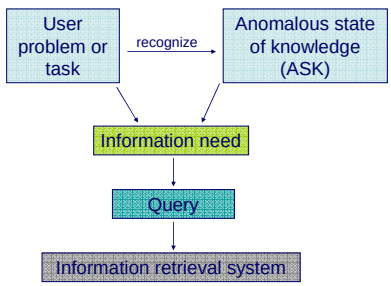
Information needs

- The ASK hypothesis

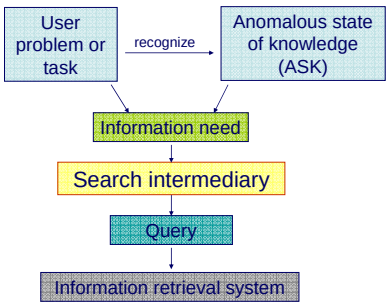
  "The ASK hypothesis is that an information need arises from a recognized anomaly in the user's state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is need to resolve that anomaly."

  – Belkin, Oddy, and Brooks. ASK for information retrieval: Part I. Background and theory. J of Documentation, 38(2), 1982.

One way to think about information needs and queries



Historical model



Context of information seeking

- Problem (user's perception of the problem)
  - Well-defined, poorly-defined?
- Intent
  - How the information will be used
  - Amount of time, effort willing to spend
  - Desired characteristics: completeness, precision, reliability, timeliness, language, source
- Internal knowledge state of user (wrt problem)
  - ASK
  - User cognitive characteristics – how knowledge is stored, organized, associated, retrieved, changed
- Public knowledge expectations, or estimate, by user
  - What knowledge is/is not available, how it is organized

- Saracevic et al. A Study of Information Seeking and Retrieving. I. Background and methodology. J Amer Soc for Inform Science, 39, 161-176, 1988.

Types of searches

- Traditional
  - Overview
  - Question answering/fact finding
  - Comprehensive research
- Web search*
  - Navigational
  - Informational
  - Transactional

* from Broder. A taxonomy of web search. SIGIR Forum, 2002.

Searches

- A search may consist of multiple queries
- Iterative refinement when initial results unsatisfactory
  - No/not enough hits
  - Too many hits
  - Unsatisfactory hits
- Some systems designed to enhance interactivity, progressive refinement

Queries

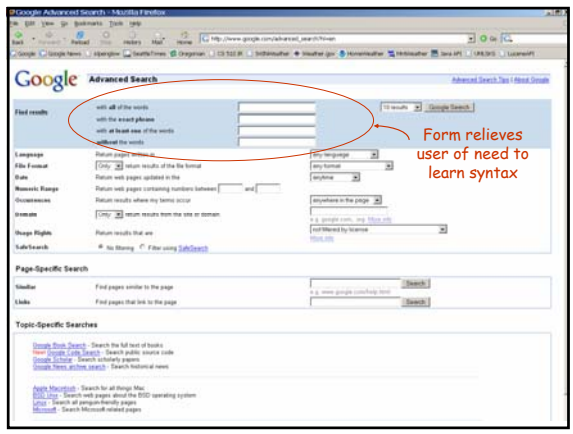
- Query
  - An expression of the user's information need that is input to an IR system
- Query language
  - The queries that are allowed as input to an IR system
    - Vocabulary
    - Syntax

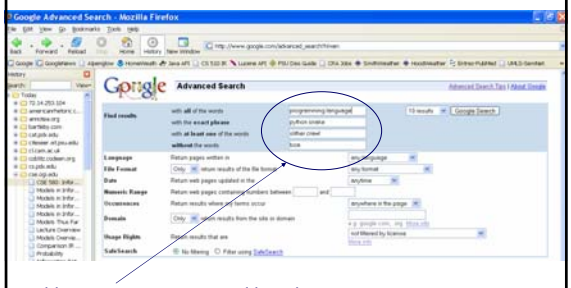
Query languages

- Vocabulary
  - Related to indexing language
  - Controlled vocabulary
  - Natural language
  - Mixed
- Syntax
  - Allows system to parse the query
  - Indicates how multiple terms are to be handled
  - Determines what types of queries are accepted
  - Understanding syntax can be key to getting good results

Query languages

- Implementation of query syntax
  - Embedded in interface – user
    - fills out a form
    - doesn't need to learn syntax
    - avoids usage errors for complicated inflexible syntax
  - Symbolic operators
    - Default operator for a list of words
    - e.g. "AND" "OR" "NOT" "+" "-"





This query gets converted into the query:

programming language slither OR crawl "python snake" -boa

Keyword-based queries

- Single word
- Context
  - Phrase
  - Proximity
- Boolean
- Natural language

Single word queries

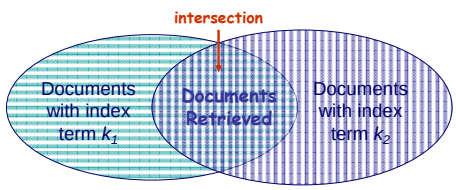
- Need to define a word
  - Word characters, separators, handle special characters and punctuation

Context queries

- Phrases
  - Usually specified with quotation marks
  - May preferentially rank phrases that match a query containing multiple single word terms (not designated as a phrase query)
  - Strictness of match may vary (e.g. require stop words to be present)
- Proximity
  - Similar to phrase but with allowed degree of "slop"
  - May or may not require same ordering of terms

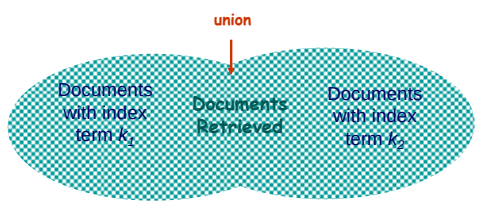
Boolean queries

- Terms (words or phrases) connected with Boolean operators



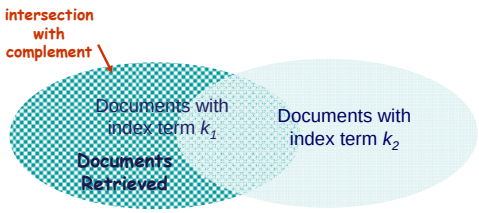
Query: k_1 **AND** k_2

Boolean queries



Query: k_1 **OR** k_2

Boolean queries



NOTE: Baeza-Yates calls this query "BUT"; can think of it as "AND NOT"

Query: k_1 **NOT** k_2

Natural language queries

- User expresses query in "natural language" – can be in the form of a question, sentence, fragment, or list of terms (words or phrases)
  - Usually treated as a bag of words
  - Some systems do some language processing to "interpret" questions
- Some systems are almost a mix of natural language and Boolean
  - Allow specification of words that *must* or *must not* be present in otherwise natural language interface

Pattern matching

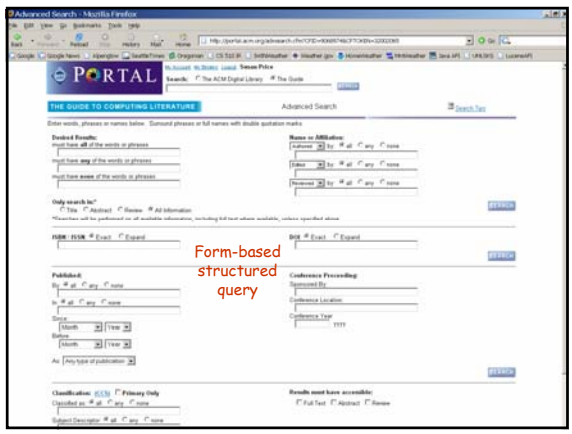
- String of characters (a "word")
- Prefix, suffix
  - e.g. *system-* to find systems, systemic, systematic
  - e.g. *-ology* to find geology, biology, anthropology
- Substring
  - e.g. *verb* matches verbal, nonverbal, verbose
- Ranges
  - Determined by lexicographic sort (words)
  - Some systems also handle ranges of numbers, dates

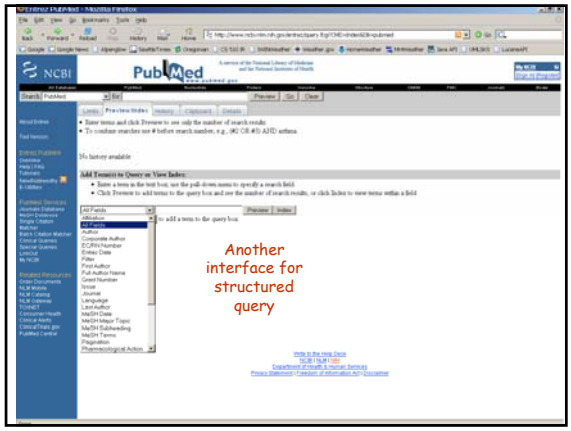
Pattern matching

- Match within a given error threshold
  - e.g. Levenshtein edit distance
- Regular expressions
  - Much more powerful (and faster to type) than a large set of "OR"s
  - Can be very helpful, e.g. trying to match gene names
  - Some systems allow only a subset of regular expressions
- How to rank regular expression matches?

Structured queries

- Query for matches in a particular field
  - Title
  - Author
  - Subject
  - Publisher
- Query applied to metadata and/or actual text
- Some systems that are indexed with controlled vocabulary will only search metadata
- Many sites offer simple search as default, fielded search as "advanced search"

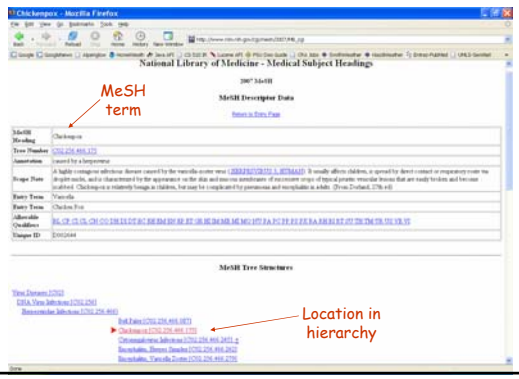




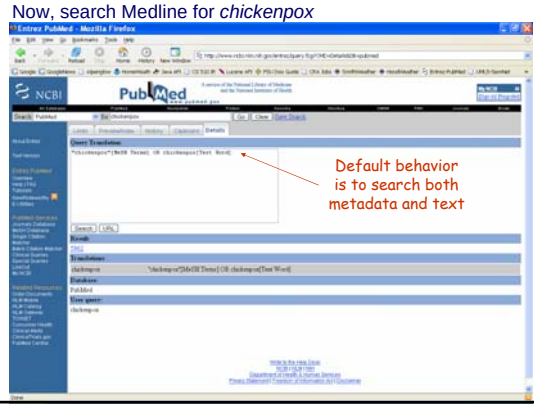
Controlled vocabulary query

- Suppose we want to get information about *chicken pox* from Medline
- Medline
  - A bibliographic database of journal articles from the National Library of Medicine
  - Indexed (manually) using Medical Subject Headings (MeSH) vocabulary
  - Indexed (automatically) using words in title and abstract
- PubMed
  - Free publicly available web interface to Medline
  - Also access to other services and databases

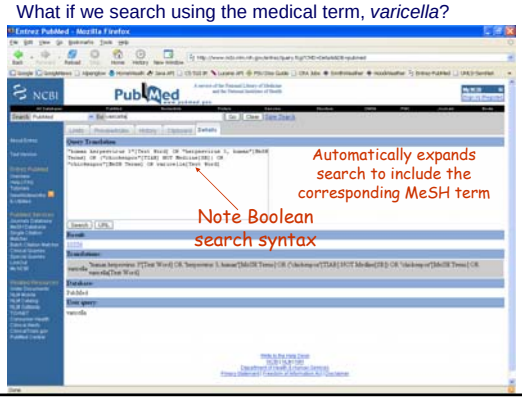
First, let's search the MeSH database for *chicken pox*



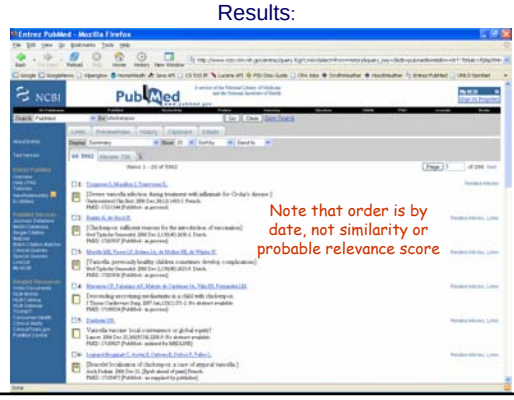
Now, search Medline for *chickenpox*



What if we search using the medical term, *varicella*?

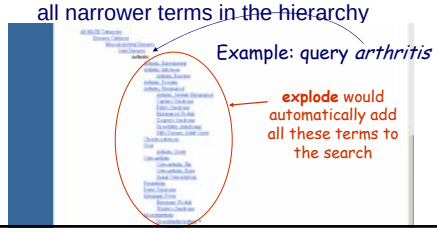


Results:



Explosions

- Some IR systems with controlled vocabularies provide an explode operator
  - Search is automatically expanded to include all narrower terms in the hierarchy



Ultraseek

- Commercial search engine, originally from Verity
  - Used for portals, intranets, etc.
  - Indexes a variety of document types
  - Has a configurable crawler
  - Allows considerable customization
  - Internals, including matching algorithms, are proprietary

Ultraseek search syntax

- Details in the online administrator guide
  - http://www.ultraseek.com/support/docs/pdf/UltraseekAdmin.pdf
- Automatic phrase recognition
  - Not intuitive behavior
  - Query *white house cat* is automatically interpreted as "white house cat"
  - To get "white house" cat, enter with quotes, or put comma at end of phrase *white house, cat*

Ultraseek search syntax

- Search terms are not weighted equally
  - Matches search terms in consecutive order
  - First term is primary, second term secondary
- Capitalization
  - If query all lower case, match is case-insensitive
  - If any upper case characters in query, match is case-sensitive
    - Requires exact match on *ClientServer*

Ultraseek search syntax

- Plus and minus
  - + before a term indicates the term **must** be in each hit
  - – before a term indicates to **prefer** hits without the term
- Query refinement
  - | Ultraseek finds documents with term(s) to the left of the bar, then searches within that set to find documents with the term(s) to the right

Ultraseek search syntax

- Filtering
  - || Ultraseek finds documents that match the query on the left side of the || then ranks that set based on the query to the right of the ||
  - Terms to the left only select the initial set; they do not affect ranking
- Wildcard expansion
  - * matches up to 40 arbitrary characters, so *cat** matches cats, catalogue, catatonic
  - ? matches a single character, so *c?t* matches cat, cot, cut

Ultraseek search syntax

- Fielded searches
  - Can search within a metadata field
  - title:brown finds documents with *brown* in the title
  - author:brown finds documents authored by *brown*
- URL searches
  - url:www.cs.pdx.edu would find the department home page and any pages below it in the directory

Ultraseek search syntax

- And there's more Ultraseek syntax ...
- Important points:
  - Failure to understand search syntax can lead to failed searches and unexpected results
  - Decisions made in designing query language for an IR system affect user experience and success
  - Trade off between flexibility and ease of use

How do real users query the Web?

- Method: analyses of Web search logs
- Query characteristics
  - In 9 studies of logs from 5 search engines collated by Jansen and Spink* :
    - 20–55% of queries consisted of a single word
    - 2–20% of queries used an advanced search operator
  - Studies reporting mean query length
    - Aol, December 2003: 2.2 words*
    - Excite, 2001: 2.6 words+

* Jansen and Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. IPM 42, 2006.

+ Spink et al. U.S. versus European web searching trends. SIGIR Forum, 36, Fall 2002

How do real users query the Web?

- Method: analyses of Web search logs*
- Most common query topics: AltaVista 2002 study
  - *People, places, or things* (49.3%)
  - *Commerce, travel, employment, or economy* (12.5%)
  - *Computers or Internet* (12.4%)

*Jansen and Spink. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. IPM 42, 2006.

What makes a good query?

Why do queries fail?

Failure analysis

- Lancaster, 1969, evaluation of MEDLARS
- Real search requests, submitted by user
  - Actual searches done by search analysts
  - Search requestors
    - Evaluated results for relevance
    - Reported relevant documents found in other ways
- Boolean query language, controlled vocabulary
  - Exhaustivity: coverage of various facets of information need (terms or groups of terms AND'd together)
  - Specificity: depth in hierarchy of controlled vocabulary

Failure analysis

- Recall failures (relevant articles not retrieved)
  - Did not cover all reasonable approaches to retrieval (33.6% of searches)
    - Omitted term(s), or combinations, that would have been successful
  - Search formulation too exhaustive (13.0%)
    - Requiring coverage of too many facets (AND'd)
  - Search formulation too specific (3.8%)

Failure analysis

- Precision failures (unwanted articles retrieved)
  - Search formulation not specific (31.3%)
  - Search formulation not exhaustive (22.3%)
  - Used inappropriate terms or combinations (11.2%)

Next: Introduction to Indexing