Introduction to Information Retrieval and Boolean Query

Lecture 1

CS 510

Information Retrieval on the Internet

Information retrieval

"Information retrieval (IR) deals with the representation, storage, organization of, and access to information items."

- Baeza-Yates and Berthier Ribeiro-Neto in Modern Information Retrieval, p. 1

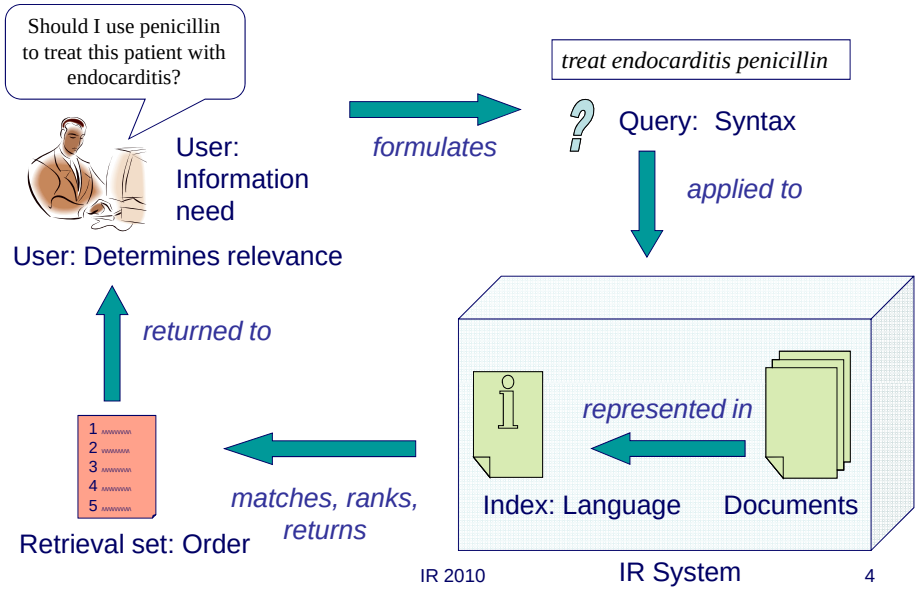
"*Information retrieval* is often regarded as being synonymous with *document retrieval* and, nowadays, with *text retrieval*, implying that the task of an IR system is to retrieve documents or texts with information content that is *relevant* to a user's information *need* ... the approaches that have been developed for this purpose are also applicable to a whole family of related information processing tasks that lie between, on the one hand, *data retrieval* and, on the other, *fact* or *knowledge retrieval*."

- Sparck Jones and Willett in Readings in Information Retrieval, p. 1

Information Retrieval

- History
  - Early work in IR in the 50's and 60's
  - Roots in library science much older, e.g.
    - Dewey Decimal system 1870s
    - Library of Congress Classification 1890s
- Important Components:
  - Indexing
  - Searching
  - User-system interaction

Typical information retrieval (IR) system

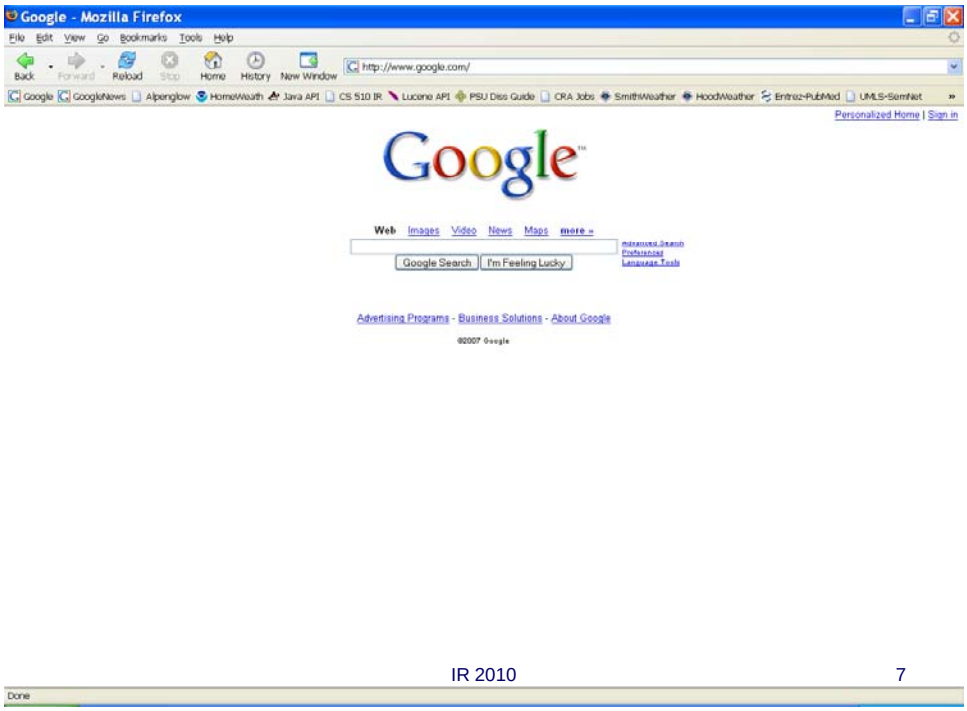


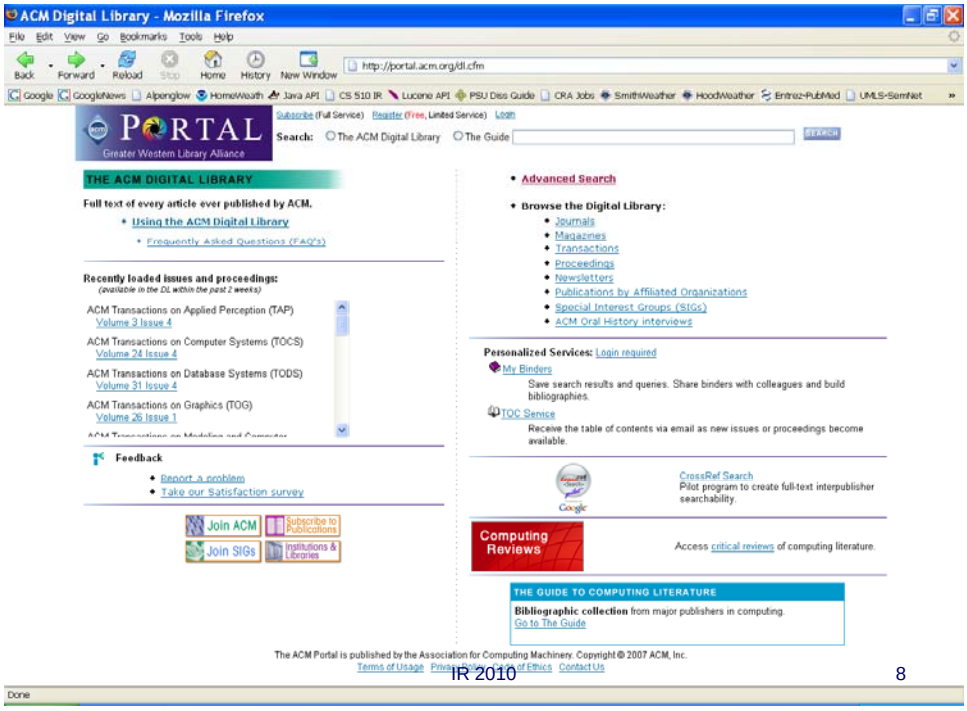
IR systems vs. DBMS

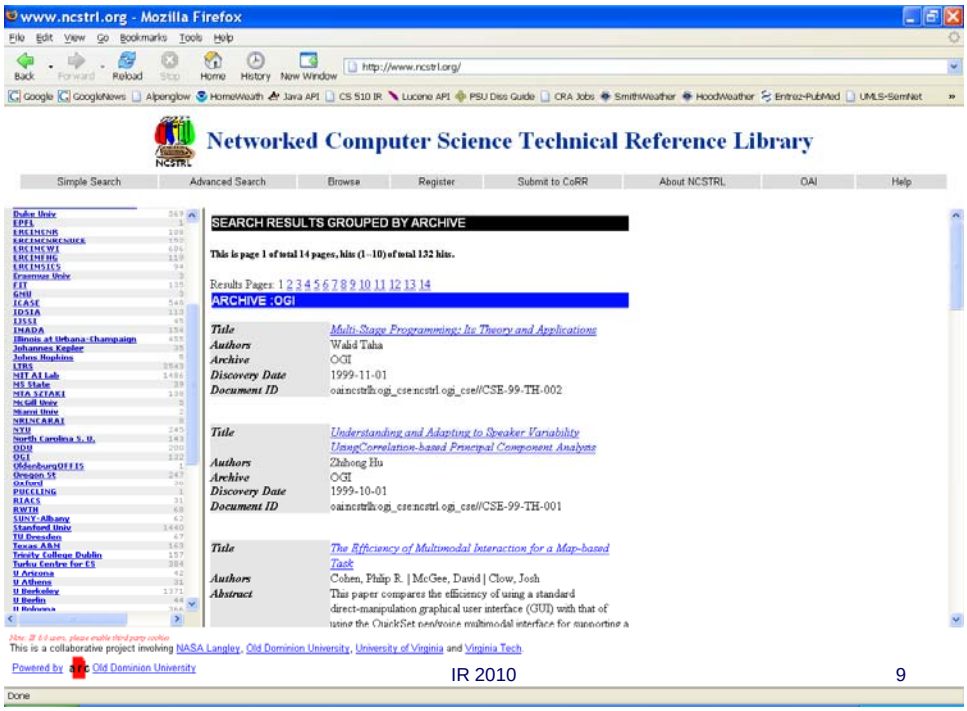
	DBMS	IR System
Target	Data (structured)	Information (text, multimedia)
Query/ Query language	Precise specification SQL	Imprecise specification Controlled vocabulary?
Matching	Exact match	“Similarity”
Results	Set of records	Set or ranked list of documents

General types of IR systems

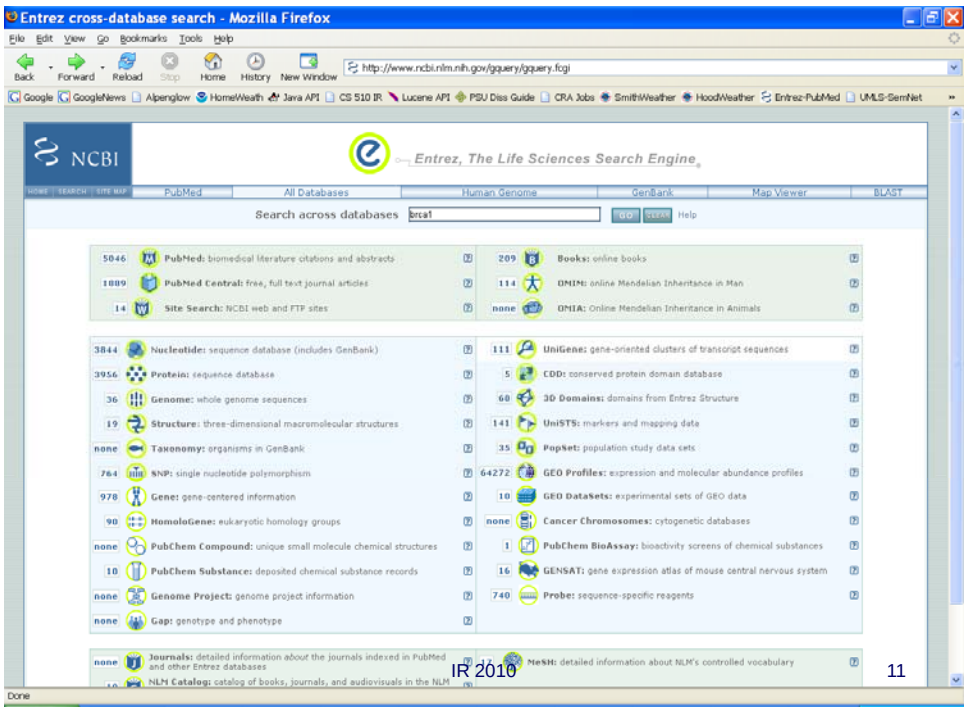
- Web
- Full text documents
- Bibliographic
- Distributed variations
  - Metasearch
  - Virtual document collections


Google - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home History New Window
http://www.google.com/
Google GoogleNews Alpenglow HomeWeath Java API CS 510 IR Lucene API PSU Diss Guide CRA Jobs SmithWeather HoodWeather Entrez-PubMed UMLS-SemNet
Personalized Home | Sign in
Google
Web Images Video News Maps more »
Google Search I'm Feeling Lucky
Advertising Programs - Business Solutions - About Google
©2007 Google
Done


ACM Digital Library - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home History New Window
http://portal.acm.org/dl.cfm
Google GoogleNews Alpenglow HomeWeath Java API CS 510 IR Lucene API PSU Diss Guide CRA Jobs SmithWeather HoodWeather Entrez-PubMed UMLS-SemNet
PORTAL
Greater Western Library Alliance
Subscribe (Full Service) Register (Free, Limited Service) Login
Search: The ACM Digital Library The Guide
SEARCH
THE ACM DIGITAL LIBRARY
Full text of every article ever published by ACM.
Using the ACM Digital Library
Frequently Asked Questions (FAQ's)
Recently loaded issues and proceedings:
(available in the DL within the past 2 weeks)
ACM Transactions on Applied Perception (TAP)
Volume 3 Issue 4
ACM Transactions on Computer Systems (TOCS)
Volume 24 Issue 4
ACM Transactions on Database Systems (TODS)
Volume 31 Issue 4
ACM Transactions on Graphics (TOG)
Volume 26 Issue 1
Feedback
Report a problem
Take our Satisfaction survey
Join ACM
Subscribe to Publications
Join SIGs
Institutions & Libraries
Advanced Search
Browse the Digital Library:
Journals
Magazines
Transactions
Proceedings
Newsletters
Publications by Affiliated Organizations
Special Interest Groups (SIGs)
ACM Oral History interviews
Personalized Services: Login required
My Binders
Save search results and queries. Share binders with colleagues and build bibliographies.
TOC Service
Receive the table of contents via email as new issues or proceedings become available.
CrossRef Search
Pilot program to create full-text interpublisher searchability.
Computing Reviews
Access critical reviews of computing literature.
THE GUIDE TO COMPUTING LITERATURE
Bibliographic collection from major publishers in computing.
Go to The Guide
The ACM Portal is published by the Association for Computing Machinery. Copyright © 2007 ACM, Inc.
Terms of Usage Privacy Policy Code of Ethics Contact Us
Done


www.ncstrl.org - Mozilla Firefox
http://www.ncstrl.org/
Networked Computer Science Technical Reference Library
Simple Search
Advanced Search
Browse
Register
Submit to CoRR
About NCSTRL
OAI
Help
SEARCH RESULTS GROUPED BY ARCHIVE
This is page 1 of total 14 pages, hits (1–10) of total 132 hits.
Results Pages: 1 2 3 4 5 6 7 8 9 10 11 12 13 14
ARCHIVE :OGI
Title
Multi-Stage Programming: Its Theory and Applications
Authors
Walid Taha
Archive
OGI
Discovery Date
1999-11-01
Document ID
oaincstrlk:ogi_cse:ncstrl.ogi_cse//CSE-99-TH-002
Title
Understanding and Adapting to Speaker Variability Using Correlation-based Principal Component Analysis
Authors
Zhihong Hu
Archive
OGI
Discovery Date
1999-10-01
Document ID
oaincstrlk:ogi_cse:ncstrl.ogi_cse//CSE-99-TH-001
Title
The Efficiency of Multimodal Interaction for a Map-based Task
Authors
Cohen, Philip R. | McGee, David | Clow, Josh
Abstract
This paper compares the efficiency of using a standard direct-manipulation graphical user interface (GUI) with that of
This is a collaborative project involving NASA Langley, Old Dominion University, University of Virginia and Virginia Tech.
Powered by Old Dominion University
Done


Dogpile Web Search Home Page - Mozilla Firefox
http://www.dogpile.com/
DOGPILE
All the best search engines piled into one.
Google
YAHOO! SEARCH
msn Search
Web
Images
Audio
Video
News
Yellow Pages
White Pages
Go Fetch!
Advanced
Review the Hottest Searches of 2006!
SPEAK!
What Do You Think?
Download Toolbar
Joke of the Day
SearchSpy
Maps
Weather
Horoscope
Add Dogpile Search to Your Site
Favorite Fetches: Port of Miami | Iraq Policy | Tax Software | Travel | 46664 | Two Headed Calf
Submit Your Site | Tell a Friend | About Dogpile | Contact Us | Tools & Tips | Privacy Policy | Terms of Use
infospace
About | Mobile | Search & Directory | Careers | Press | Investor Relations
© 2007 InfoSpace, Inc. All Rights Reserved.
Done



Facets of IR systems: Scope

- Web
- Limited in Focus, Kind or Location
  - PubMed
  - Patents, News Articles
  - Company documents
- Local
  - Files
  - Email

Facets of IR systems: Documents

- Format
  - HTML
  - PDF
  - Word processed
  - Scanned OCR
- Type
  - Text
  - Multimedia
  - Semistructured (XML)
- Dynamic

Facets of IR systems: Engine

- Supported tasks: search, browse
- Basic model: Boolean vs. ranked results
- Indexing language: controlled keywords vs. natural language
- Indexing target: bibliographic data vs. full text
- Search syntax; available operations

Fact of IR Systems: Content vs. Context

- Content only
  Boolean Query – presence or absence of term
- Extended content
  Meta-data, anchor text
- Collection statistics
  Word frequency: “snake” vs. “anaconda”
- Collection structure
  In-links

User Tasks

- Mode
  – Search (retrieve)
  – Browse
- Purpose
  – Overview
  – Question answering/fact finding
  – Comprehensive research
  – Finding known item (document, page, or site)
  – Transaction (e.g. buy a book, download a file)

Document Representation: Logical

- Goal:
  - Represent the content
  - Represent other aspects (sometimes)
- Methods:
  - Assign descriptors (usually selected from a predetermined list)
  - Extract features (usually words or phrases if text document)
  - Descriptors provided by others: taggin

Document Representation: Logical

- History
  - Early systems mostly bibliographic
    - Systems contained brief surrogates for documents
      - Title, author, abstract, location of full document
      - Remember card catalogs??
    - Provided reference to the full document
  - Limited by available storage and processing power
  - Affected the way document content was represented

Document Representation: Logical

- Methods:
  - Manual
    - usually assign terms from a controlled vocabulary
  - Automatic
    - usually extract terms from the document
- Considerations:
  - Size of the representation
  - Improve likelihood of appropriate matching to queries

Document Representation: Logical

Abstract: The objectives of this study were to determine if (1) children with migraine experience greater sleep disturbances than their siblings, (2) those with more severe migraine have greater levels of sleep disruption, and (3) these sleep disturbances lead to greater behavioral problems and more missed school. Children aged 6 to 18 years with a diagnosis of migraine for > 6 months, who had at least one sibling without migraine in the same age range, were identified through our neurology clinic database or at the time of the clinic visit. Parents completed the (1) demographic, general health, and migraine information questionnaire; (2) Child Sleep Habits Questionnaire; and (3) Behavior Assessment System for Children: Second Edition (BASC-2) Parent Rating Scales for each child. Cases with migraine had higher total sleep (P < .02), sleep delay (P < .03), and daytime sleepiness scores (P < .001) than controls. Cases with more severe migraines had higher total sleep (P < .01) and sleep duration scores (P < .03) than those with milder headaches. In cases, higher total sleep...

Assigned Indexing Terms:
Activities of Daily Living
Adolescent
Case-Control Studies
Child
Child Behavior
Disorders/*complications/psychology
Circadian Rhythm/physiology
Female
Humans
Male
Migraine
Disorders/*complications/psychology
Severity of Illness Index
Sleep
Disorders/*complications/psychology
Sleep Stages/physiology
Wakefulness/*physiology

Document Representation: Logical

- Techniques to improve representation
  - Remove stopwords
  - Stemming
  - Exploit document structure (title, subject:)
- Possible text processing steps
  - Determine *what* to index (frames? page title? metadata?)
  - Strip formatting (e.g. html tags, w.p. instr.)
  - Recognize structure
  - Recognize words (or phrases)

Document Representation: Physical

- Associate document identifier with various descriptors
- If descriptors are extracted terms, may also include
  - Frequency of terms
  - Position of terms
- Considerations
  - Efficient storage
  - Representations that can be searched to allow a fast response to user requests

Queries

- User has an information need
- Information need communicated to IR system as a request
- Request encoded in a query expression
- Query expression must be interpretable by the system (query language)
  - Encoded in the interface
  - Terms and operators
  - Pattern syntax

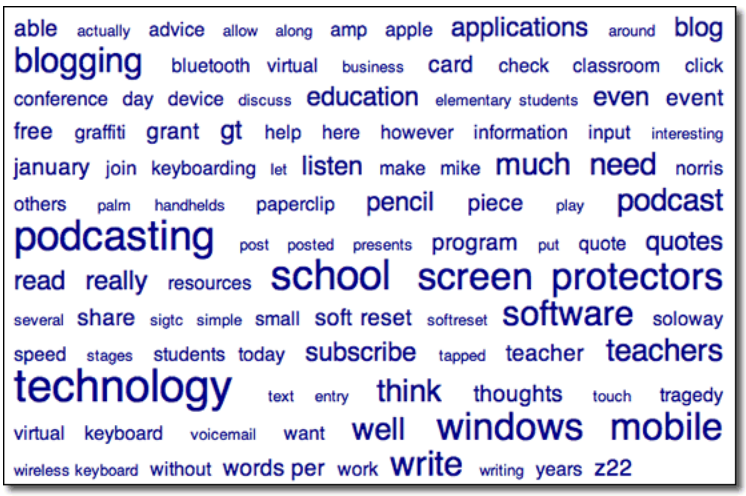
Retrieval

- Match information request representation (query) to document representations (stored in an index)
- Use an algorithm to compute the matching
  - Yes/no match (Boolean expressions)
  - Similarity match
- Order and Organize results
  - Ranked by similarity (or other)
  - Clustered
  - Organized by type (images, news articles)

Unit of Retrieval

- Single page or file
- Site or cluster of documents
- Document fragment
- Summary or important terms (word cloud)
- Structured facts

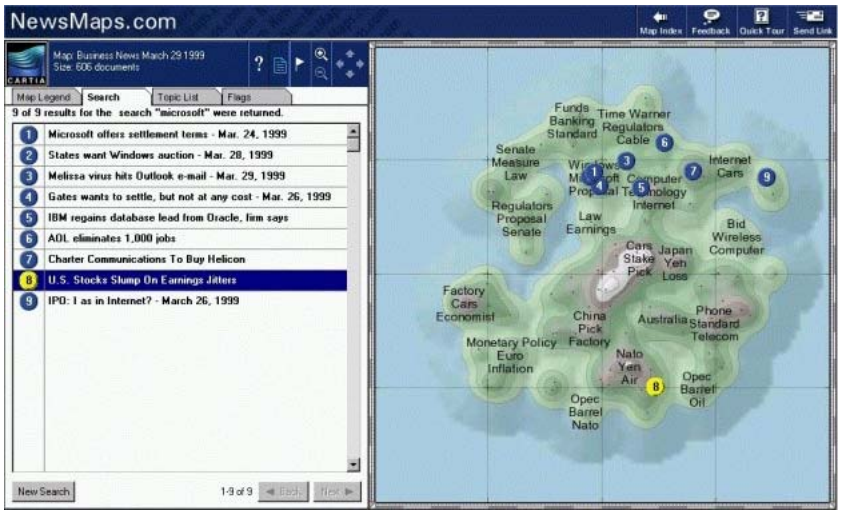
Word Cloud



Other IR Tasks

- Classification
- Clustering
- Synoptic view
- Question answering
- Segmentation
- Summarization
- Information extraction

Themescape



System Issues

- Index organization and construction
- Scaling
- Collection building
- “Publicizing”

Course Organization

http://web.cecs.pdx.edu/~maier/cs510iri/index.htm

Please print off your own lecture notes for subsequent classes

Boolean Search

Example of retrieval process

- Model: Document as set of words
- Query: Boolean combination of words
- Index: Inverted file
- Retrieval: Merging of lists

Boolean Search: Model

Document: Unit of retrieval

Term: Basic constituent of documents and queries

- Words: `snake, cobra, python`
- Stems: `mak-, determin-`
- Phrases: `"garter snake"`
- Special Chars: `1978, s70, help@ask.com`

Document is a set of terms

- `"to be or not to be"` → `{to,be,or,not}`
- Term is in a document or not (not count, not position)

Boolean Search: Query

Terms connected with `AND, OR, NOT`

```
cobra AND python
(cobra OR python) NOT language
```

Could be entered through a form

Get all documents that satisfy condition, none that don't

Boolean Search: Indexing

Inverted file: for each term, as list of documents (IDs) containing the term

- Usually sorted, elements called "postings"

```
cobra: 1,6,7,9,12
python: 1,5,8,9,11,12
language: 2,3,5,7,8,10,13,…
```

Terms organized in a dictionary, pointing to appropriate list

Boolean Search: Retrieval

Union (OR), Intersection(AND), Different (NOT) of lists

```
cobra AND python
   cobra: 1,6,7,9,12
   python: 1,5,8,9,11,12

(cobra OR python) NOT language
   t1 = cobra OR python
   cobra: 1,6,7,9,12
   python: 1,5,8,9,11,12

   t1 NOT language
   language: 2,3,5,7,8,10,13,…
```


Optimization

Look at sizes of lists to decide merge order

```
snake AND cobra AND python
11,345     280       156
```

Rewrite for smaller intermediate results

```
(snake OR language) AND python
 11,345    23,667        156
(snake AND python) OR
 (language AND python)
```

Extensions

Proximity of words

```
python NEAR language
python NEAR(3) language
python /P language
```

(same paragraph)

How can we process such queries?