Introduction to Information Retrieval

Lecture 1

CS 410/510

Information Retrieval on the Internet

Lecture Outline

- IR systems
  - Overview
  - IR systems vs. DBMS
  - Types, facets of interest
- User tasks
- Document representations
- Queries
- Retrieval
- Evaluation

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

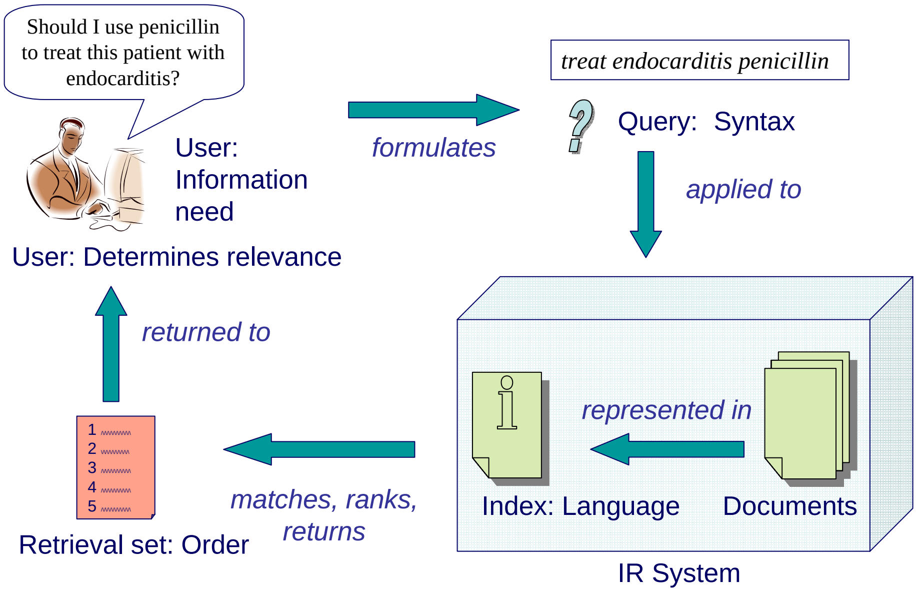
"*Information retrieval* is often regarded as being synonymous with *document retrieval* and, nowadays, with *text retrieval*, implying that the task of an IR system is to retrieve documents or texts with information content that is *relevant* to a user's information *need* ... the approaches that have 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



Should I use penicillin to treat this patient with endocarditis?

User: Information need

formulates

treat endocarditis penicillin

Query: Syntax

applied to

IR System

represented in

Index: Language

Documents

matches, ranks, returns

Retrieval set: Order

returned to

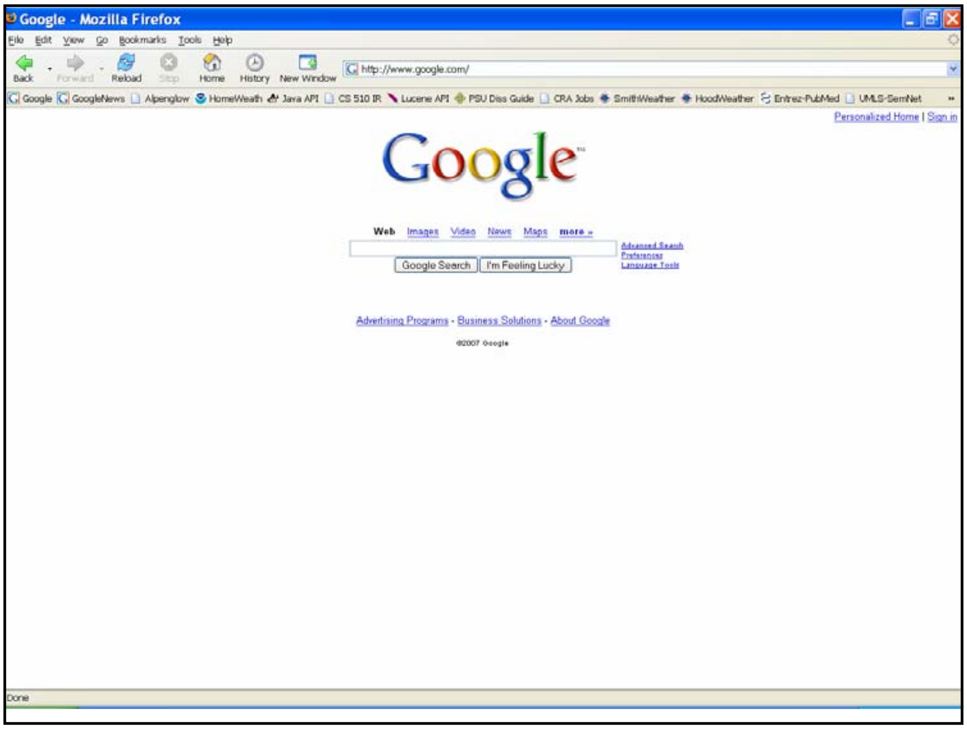
User: Determines relevance

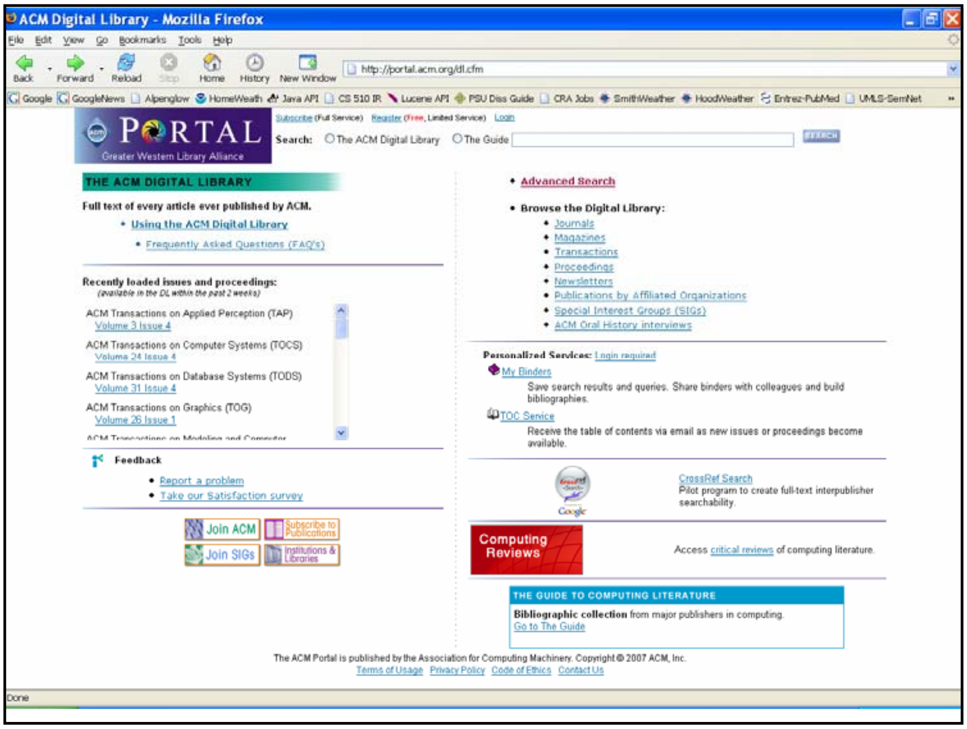
IR systems vs. DBMS

	DBMS	IR System
Target		
Queries / Query language		
Matching		
Results		

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
Advanced Search
Preferences
Language Tools
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
SEARCH
Done

Subscribe (Full Service) Register (Free, Limited Service) Login

Search: ○ The ACM Digital Library ○ The Guide

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

ACM Transactions on Modeling and Computer

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.

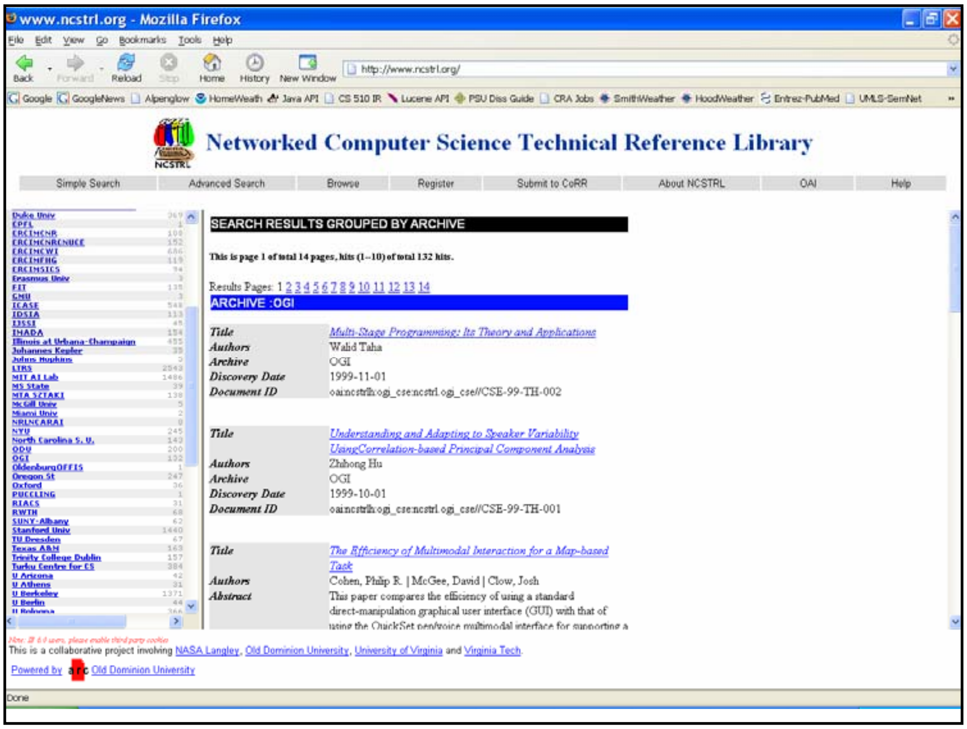
TOC Service
Receive the table of contents via email as new issues or proceedings become available.

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


www.ncstrl.org - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home History New Window
http://www.ncstrl.org/
Google GoogleNews Alpenglow HomeWeath Java API CS 510 IR Lucene API PSU Diss Guide CRA Jobs SmithWeather HoodWeather Entrez-PubMed UMLS-SemNet
NCSTRL
Done

Networked Computer Science Technical Reference Library

Simple Search | Advanced Search | Browse | Register | Submit to CoRR | About NCSTRL | OAI | Help

Duke Univ	369
EPFL	1
ERCIMCNR	109
ERCIMCNRCNUCE	152
ERCIMCWI	686
ERCIMFHG	119
ERCIMSICS	94
Erasmus Univ	3
FIT	135
GMU	3
ICASE	548
IDSIA	113
IISSI	45
IHADA	154
Illinois at Urbana-Champaign	455
Johannes Kepler	35
Johns Hopkins	5
LTRS	2543
MIT AI Lab	1486
MS State	39
MTA SZTAKI	138
McGill Univ	5
Miami Univ	2
NRLNCARAI	0
NYU	245
North Carolina S. U.	143
ODU	200
OGI	132
OldenburgOFFIS	1
Oregon St	247
Oxford	36
PUCCLING	1
RIACS	31
RWTH	68
SUNY-Albany	62
Stanford Univ	1440
TU Dresden	67
Texas A&M	163
Trinity College Dublin	157
Turku Centre for CS	384
U Arizona	42
U Athens	31
U Berkeley	1371
U Berlin	44
U Bologna	366

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

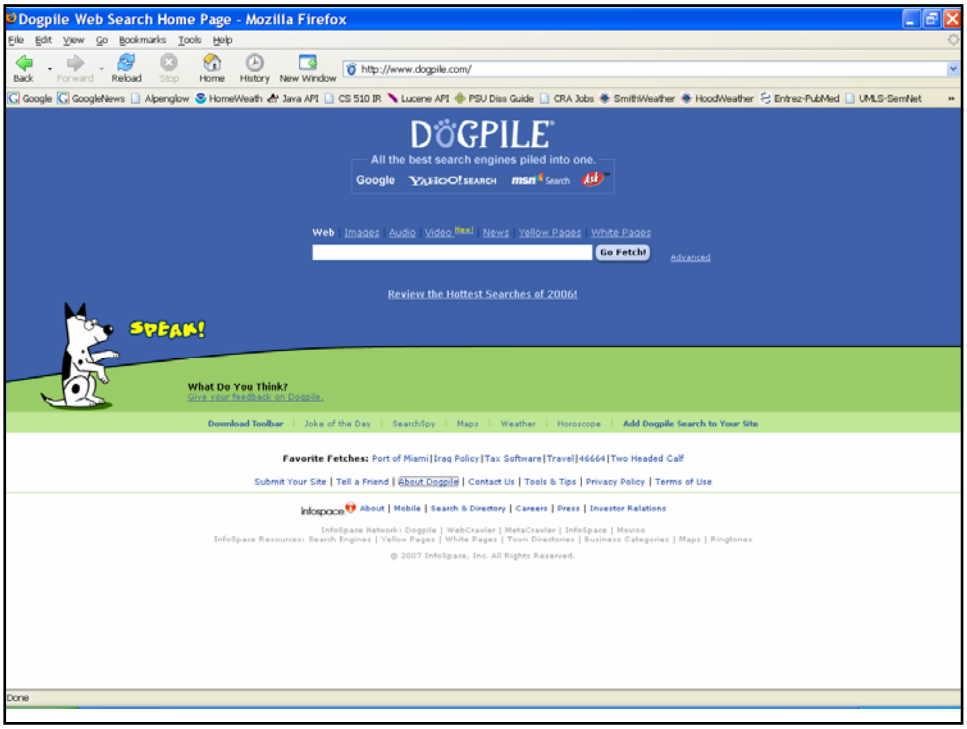
Title	*Multi-Stage Programming: Its Theory and Applications*
Authors	Walid Taha
Archive	OGI
Discovery Date	1999-11-01
Document ID	oai:ncstrlh:ogi_cse:ncstrl.ogi_cse//CSE-99-TH-002

Title	*Understanding and Adapting to Speaker Variability UsingCorrelation-based Principal Component Analysis*
Authors	Zhihong Hu
Archive	OGI
Discovery Date	1999-10-01
Document ID	oai:ncstrlh: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 using the QuickSet pen/voice multimodal interface for supporting a

Note: IE 6.0 users, please enable third party cookies
This is a collaborative project involving NASA Langley, Old Dominion University, University of Virginia and Virginia Tech.

Powered by arc Old Dominion University


Dogpile Web Search Home Page - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home History New Window
http://www.dogpile.com/
Google GoogleNews Alpenglow HomeWeath Java API CS 510 IR Lucene API PSU Diss Guide CRA Jobs SmithWeather HoodWeather Entrez-PubMed UMLS-SemNet
DOGPILE
SPEAK!
Done

All the best search engines piled into one.

Google YAHOO! SEARCH msn Search Ask

Web | Images | Audio | Video New! | News | Yellow Pages | White Pages

Go Fetch! Advanced

Review the Hottest Searches of 2006!

What Do You Think?
Give your feedback on Dogpile.

Download Toolbar | Joke of the Day | SearchSpy | Maps | Weather | Horoscope | Add Dogpile Search to Your Site

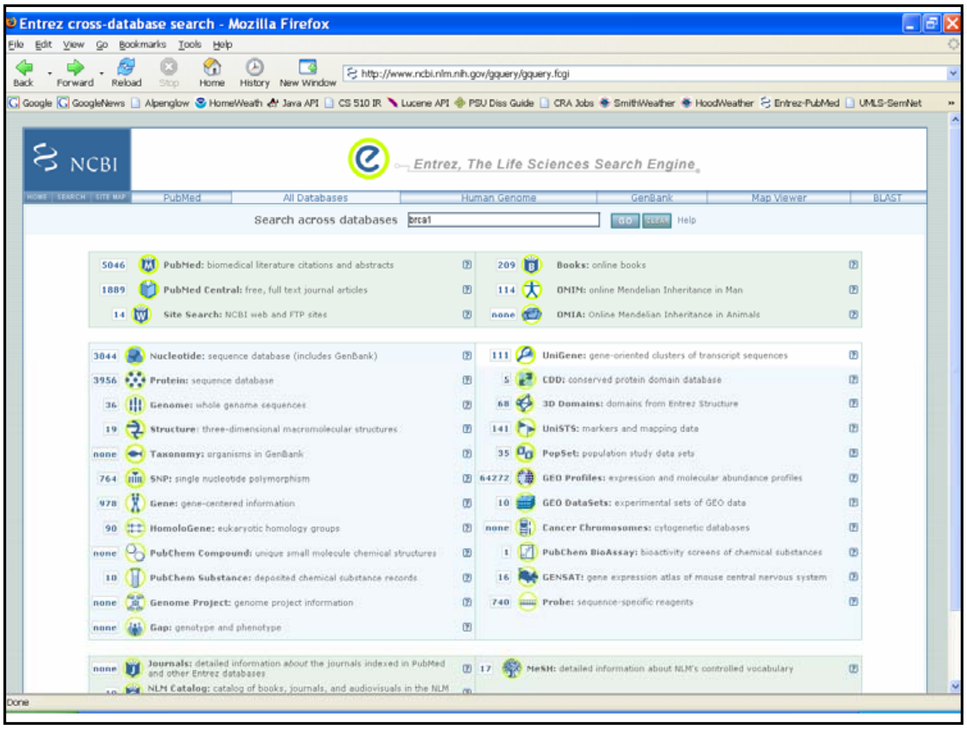
Favorite Fetches: Port of Miami | Iraq Policy | Tax Software | Travel | 46664 | Two Headed Calf

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

InfoSpace Network: Dogpile | WebCrawler | MetaCrawler | InfoSpace | Moviso
InfoSpace Resources: Search Engines | Yellow Pages | White Pages | Town Directories | Business Categories | Maps | Ringtones

© 2007 InfoSpace, Inc. All Rights Reserved.


Entrez cross-database search - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home History New Window
http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
Google GoogleNews Alpenglow HomeWeath Java API CS 510 IR Lucene API PSU Diss Guide CRA Jobs SmithWeather HoodWeather Entrez-PubMed UMLS-SemNet
NCBI
HOME SEARCH SITE MAP
Done

Entrez, The Life Sciences Search Engine

PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

Search across databases brca1 GO CLEAR Help

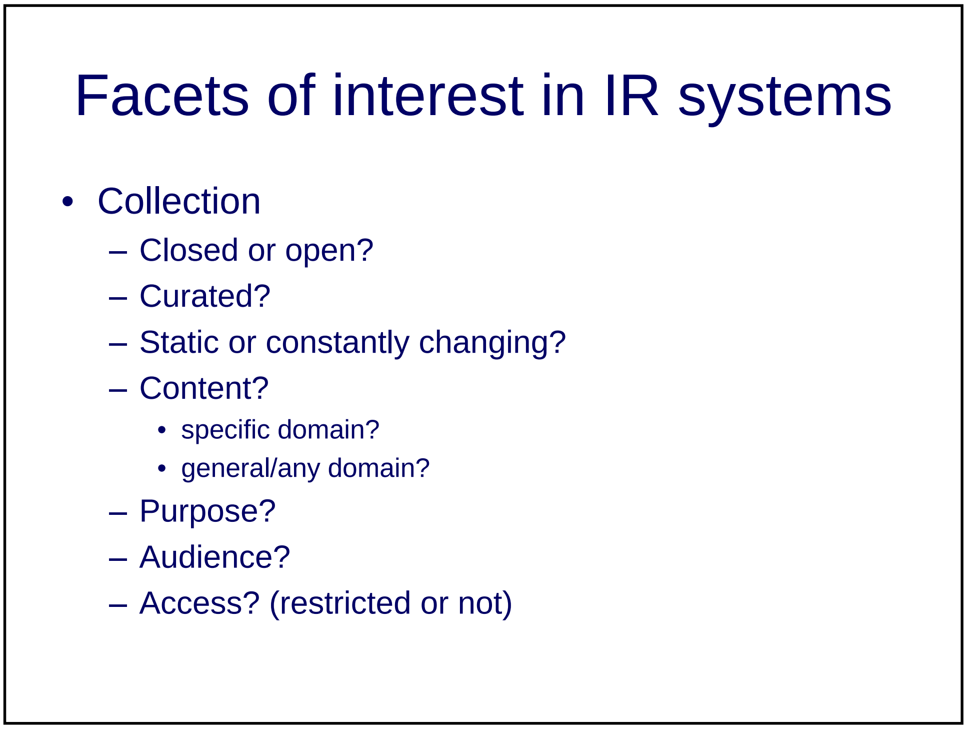
5046	**PubMed:** biomedical literature citations and abstracts	209	**Books:** online books
1889	**PubMed Central:** free, full text journal articles	114	**OMIM:** online Mendelian Inheritance in Man
14	**Site Search:** NCBI web and FTP sites	none	**OMIA:** Online Mendelian Inheritance in Animals
3844	**Nucleotide:** sequence database (includes GenBank)	111	**UniGene:** gene-oriented clusters of transcript sequences
3956	**Protein:** sequence database	5	**CDD:** conserved protein domain database
36	**Genome:** whole genome sequences	68	**3D Domains:** domains from Entrez Structure
19	**Structure:** three-dimensional macromolecular structures	141	**UniSTS:** markers and mapping data
none	**Taxonomy:** organisms in GenBank	35	**PopSet:** population study data sets
764	**SNP:** single nucleotide polymorphism	64272	**GEO Profiles:** expression and molecular abundance profiles
978	**Gene:** gene-centered information	10	**GEO DataSets:** experimental sets of GEO data
90	**HomoloGene:** eukaryotic homology groups	none	**Cancer Chromosomes:** cytogenetic databases
none	**PubChem Compound:** unique small molecule chemical structures	1	**PubChem BioAssay:** bioactivity screens of chemical substances
10	**PubChem Substance:** deposited chemical substance records	16	**GENSAT:** gene expression atlas of mouse central nervous system
none	**Genome Project:** genome project information	740	**Probe:** sequence-specific reagents
none	**Gap:** genotype and phenotype		
none	**Journals:** detailed information about the journals indexed in PubMed and other Entrez databases	17	**MeSH:** detailed information about NLM's controlled vocabulary
	NLM Catalog: catalog of books, journals, and audiovisuals in the NLM		

Facets of interest in IR systems

- Site
  - – single controlled website (e.g. intranet)
  - – distributed across multiple sites

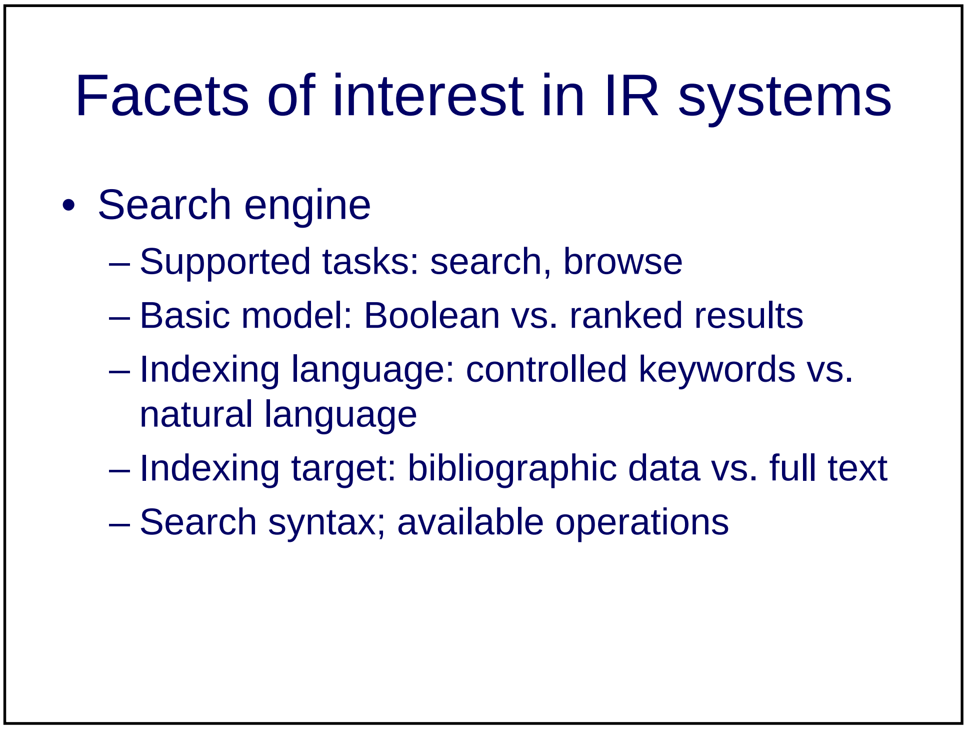
Facets of interest in IR systems

- Documents
  - – Hyperlinked?
  - – Format?
    - HTML
    - PDF
    - Word processed
    - Scanned OCR?
  - – Type?
    - Text?
    - Multimedia?
    - Semistructured (XML)?
  - – Static or dynamic?



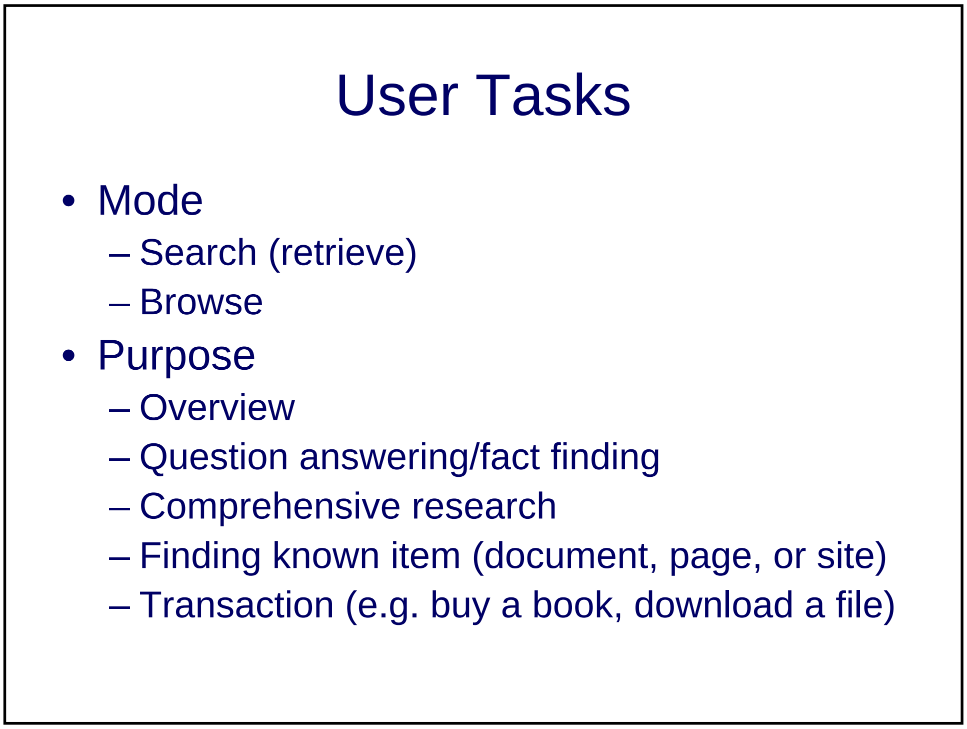
Facets of interest in IR systems

- Collection
  - Closed or open?
  - Curated?
  - Static or constantly changing?
  - Content?
    - specific domain?
    - general/any domain?
  - Purpose?
  - Audience?
  - Access? (restricted or not)



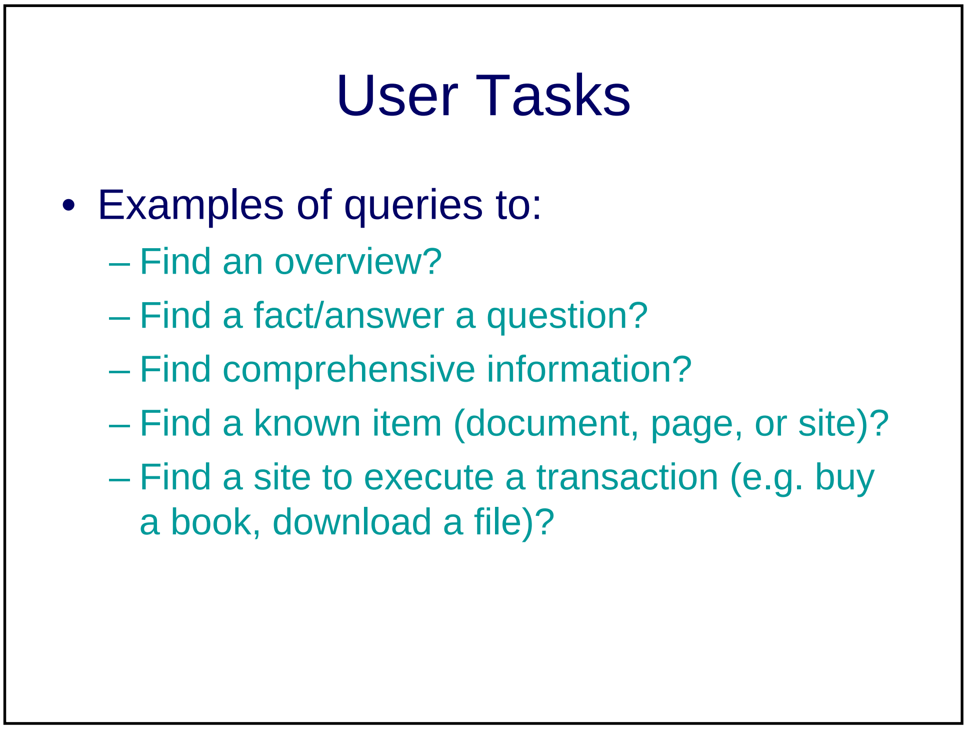
Facets of interest in IR systems

- Search 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



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)



User Tasks

- Examples of queries to:
  - Find an overview?
  - Find a fact/answer a question?
  - Find comprehensive information?
  - Find a known item (document, page, or site)?
  - Find a site to execute a 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)
  - Other (e.g., try to anticipate what questions the document could answer)

Document Representation: Logical

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

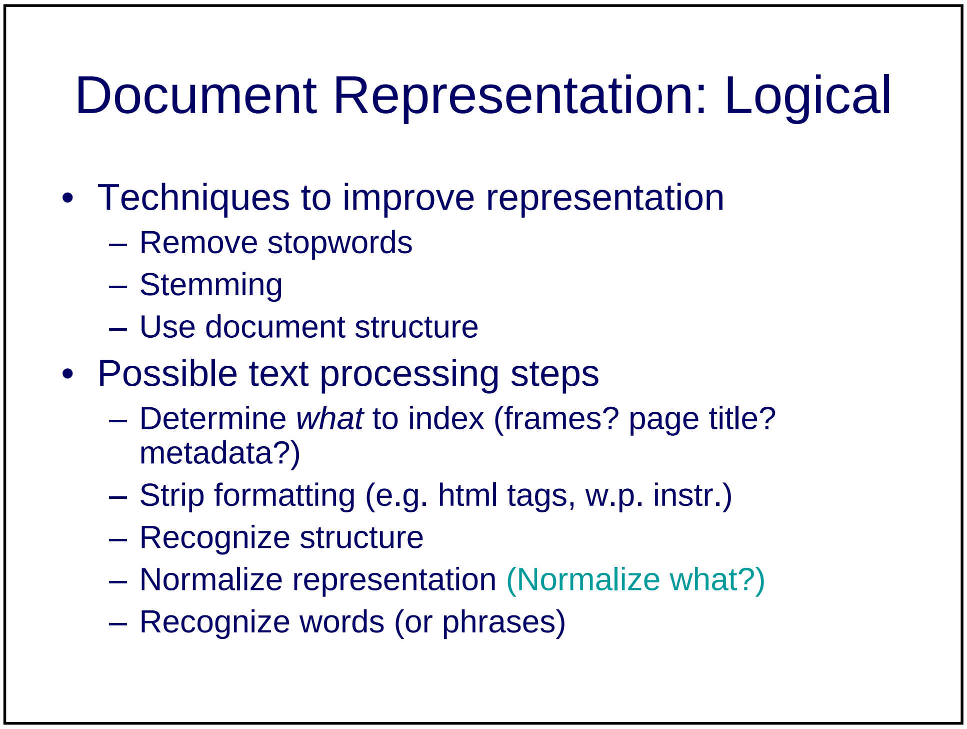
Document Representation: Logical

- 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

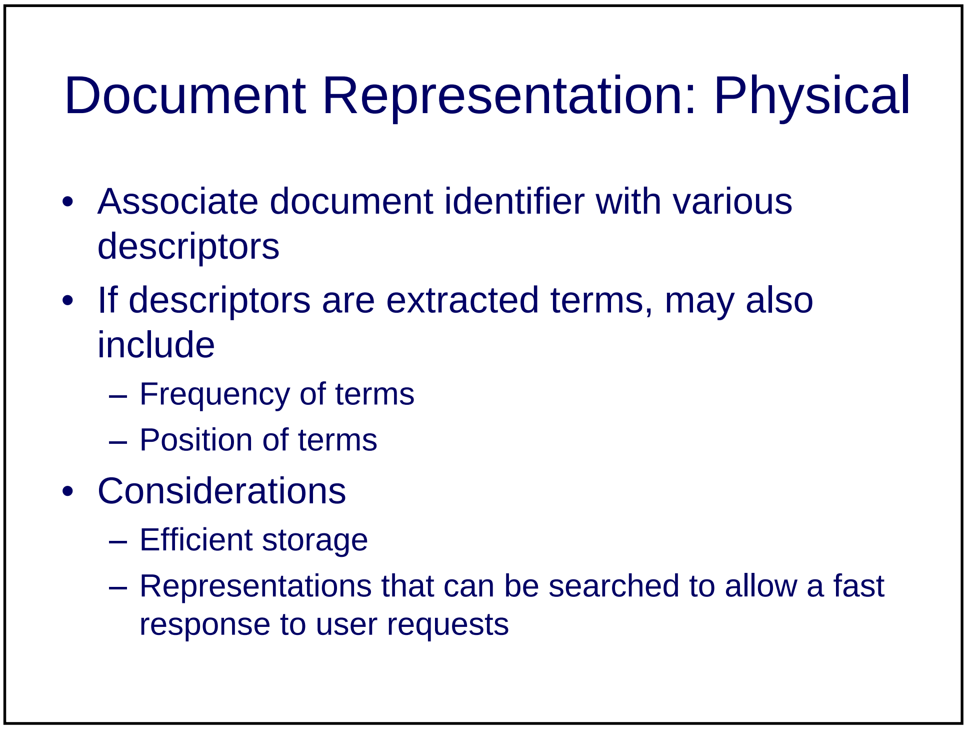
Abstract: The objectives of this study were to determine if (1) children with migraine experience greater sleep disturbances than their siblings, (2) those with more severe migraine have greater levels of sleep disruption, and (3) these sleep disturbances lead to greater behavioral problems and more 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
  - Use document structure
- Possible text processing steps
  - Determine *what* to index (frames? page title? metadata?)
  - Strip formatting (e.g. html tags, w.p. instr.)
  - Recognize structure
  - Normalize representation (Normalize what?)
  - 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
  - Query expressions that use symbols to express operators

Retrieval

- Match information request representation (query) to document representations (stored in an index)
- Use an algorithm to compute the matching
  - Exact match (Boolean expressions)
    - How should results be ordered?
  - Ranked list based on similarity computation (and possibly other factors)

IR as Classification

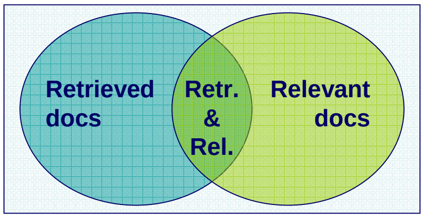
- Indexing techniques classify documents
  - Keywords, either assigned or extracted, define categories
- Information requests also identify one or more categories of interest
- Compare to bricks-and-mortar libraries?

Evaluation

- How can we evaluate performance of an IR system?
  - System perspective
  - User perspective
- Relevance
  - (How well) does a document satisfy the user’s need (or “anomalous state of knowledge”)?

Evaluation: Metrics

- There are many
- For now, two basics:



$$\text{Recall} = \frac{\text{\# documents retrieved and relevant}}{\text{\# documents relevant}}$$

$$\text{Precision} = \frac{\text{\# documents retrieved and relevant}}{\text{\# documents retrieved}}$$

Questions?

- Next: Overview of text processing