

# Introduction to Information Retrieval

(CS429)

Nazli Goharian

nazli@ir.iit.edu

Slides are *mostly* based on Information Retrieval Algorithms and Heuristics, Grossman, Frieder

© Goharian, Grossman, Frieder, 2009

## Course Outline

- **Introduction**
- Retrieval Strategies
- Retrieval Utilities
- Evaluation
- Indexing
- Efficiency in indexing and query processing
- Integrating Structured Data and Text
- Distributed IR: Web & P2P
- Meta Search

© Goharian, Grossman, Frieder, 2009

## What is Information Retrieval?

- Salton (1968):

“Information retrieval is a field concerned with structure, analysis, organization, storage, searching, and retrieval of information”

© Goharian, Grossman, Frieder, 2009

## IR Tasks/Applications

- World Wide Web (web search) -- most common  
#pages indexed: ~50,000 (1994); 10s of billions (today)  
(ex: Google, Yahoo, Bing)
- Vertical/ Topical search (ex: MEDLINE, USPTO, LEXIS)
- Enterprise search (ex: Autonomy; Lucene – open source)
- Desktop search (ex: Microsoft Vista)
- Peer-to-peer search (ex: Limewire open source of Gnutella, KaZaA, eMule/eDonkey)

© Goharian, Grossman, Frieder, 2009

## IR Tasks/Applications (Cont'd)

- Informational (ad hoc)- Enterprise/desktop/Web
  - The real problem boils down to matching the language of the query to the language of the document.
- Navigational- Web
- Transactional- Web
- Question Answering
- Filtering/Routing
- Classification/Categorization

© Goharian, Grossman, Frieder, 2009

## Database vs. Information Retrieval

	Structured Data (Transactional)	Structured Data (Data Warehouse)	Text Data
Accuracy	100%	100%	~30-40%
Query Language	SQL	SQL, OLAP	Natural language
Volumes	10s TB	~500TB	~200TB (Web) 15-20%
Foundation	Algorithm	Algorithm	Heuristics
Validation	Objective	Objective	Subjective

© Goharian, Grossman, Frieder, 2009

## Definitions

- A *database* is a collection of documents.
- A *document* is a sequence of terms, expressing ideas about some topic in a natural language.
- A *term* is a semantic unit, a word, phrase, or potentially root of a word.
- A *query* is a request for documents pertaining to some topic.

© Goharian, Grossman, Frieder, 2009

## Hard Parts of IR

- Simply matching on words is a very brittle approach.
- One word can have a zillion different semantic meanings
  - Consider: Take
  - “take a place at the table”
  - “take money to the bank”
  - “take a picture”
  - “take a lot of time”
  - “take drugs”

© Goharian, Grossman, Frieder, 2009

## More Problems with IR

- You can't even tell what part of speech a word has:
  - “I saw her duck”
  - A query that searches for “pictures of a duck” will find documents that contains:  
“I saw her duck away from the ball falling from the sky”

© Goharian, Grossman, Frieder, 2009

## More Problems with IR

- Proper Nouns often use regular old nouns
- Consider a document with “a man named Abraham owned a Lincoln”
- A word matching query for “Abraham Lincoln” may well find the above document.

© Goharian, Grossman, Frieder, 2009

## What is Different about IR from the rest of Computer Science

- Most algorithms in computer science have a “right” answer:
  - Sort the following ten integers
  - Find the highest integer
- Now consider:
  - *Find the document most relevant to “hippos in the zoo”*

**Question:** How to measure the **relevance**?

© Goharian, Grossman, Frieder, 2009

## Relevance/Effectiveness

- An algorithm is deemed incorrect if it does not have a “right” answer.
- A heuristic tries to guess something close to the right answer. Heuristics are measured on “how close” they come to a right answer.
- IR techniques are essentially heuristics because we do not know the right answer.
- So we have to measure how *close* to the right answer we can come.

© Goharian, Grossman, Frieder, 2009

## Retrieval Model/Strategy

- An IR *model* or *strategy* is a technique by which a *relevance measure* is obtained between a query and a document.
- Depending on the model various query and *document statistics* are used.
- Additional factors maybe used such as:
  - Link analysis (*page popularity, anchor text*)
  - User Log data (*ex. Clickthrough data, dwell time*)

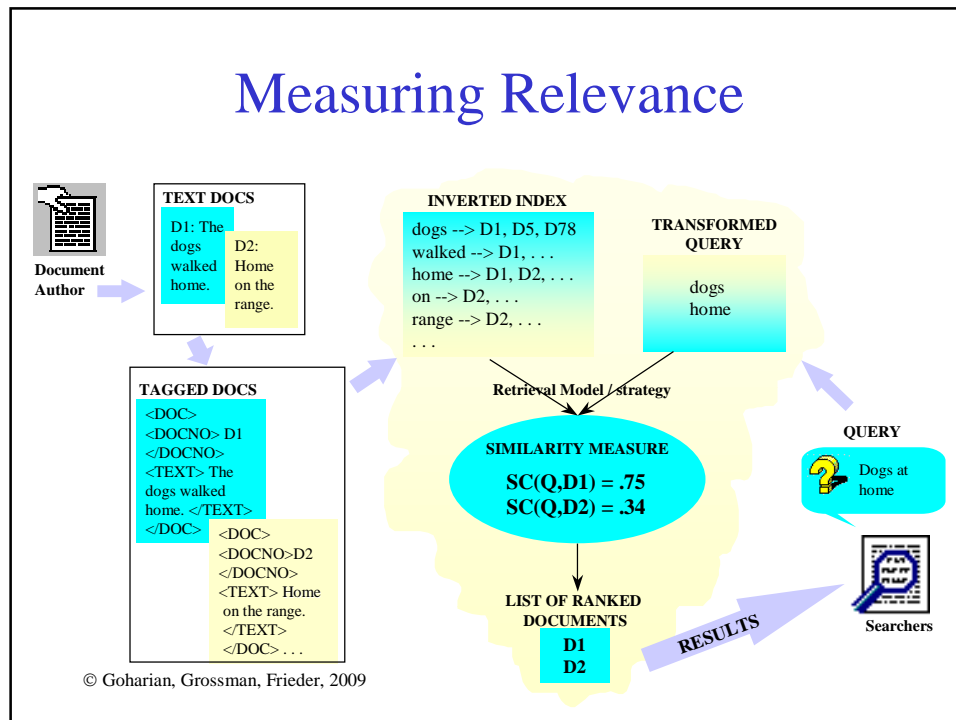
© Goharian, Grossman, Frieder, 2009

## Models/Strategies

- Manual
  - Boolean
- Automatic
  - Probabilistic
    - OKAPI, Robertson/Sparck Jones
    - Kwok
  - Language Models
  - Vector Space Model
  - Inference Networks
  - Latent Semantic Indexing (LSI)
- Adaptive Models
  - Genetic Algorithms
  - Neural Networks

© Goharian, Grossman, Frieder, 2009

# Measuring Relevance



## Model/Strategy vs. Utility

- An IR *model* is a technique by which a relevance assessment (*relevance ranking*) is obtained between a query and a document.
- An IR *utility* is a technique that may be used to improve the assessment (*effectiveness*) given by a model.

## Utilities

- Parsing
- Stemming
- N-grams
- Thesauri
- Relevance Feedback
- Clustering
- Passage-based retrieval
- Semantic Networks

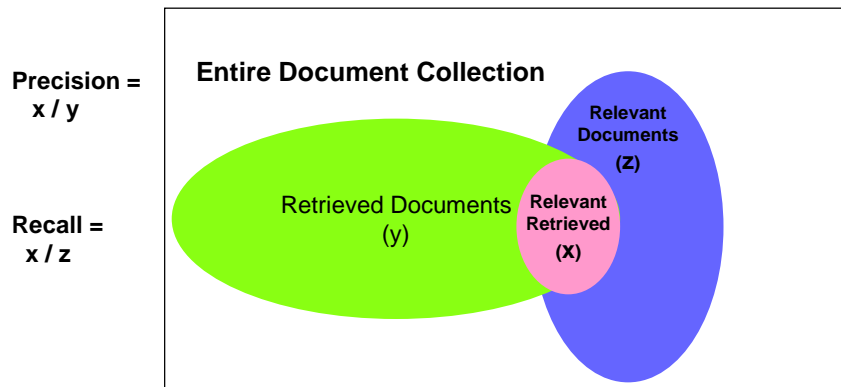
© Goharian, Grossman, Frieder, 2009

## Evaluating Engine's Effectiveness

- *Recall* is the fraction of relevant documents retrieved from the set of total relevant documents collection-wide. *In Web search Recall measure is not possible.*
- *Precision* is the fraction of relevant documents retrieved from the total number retrieved.
- Variations of these measures exist – *will be discussed later!*

© Goharian, Grossman, Frieder, 2009

## Precision / Recall



© Goharian, Grossman, Frieder, 2009

## TREC

- Text Retrieval Conference- sponsored by NIST
- A benchmark for evaluating IR systems.
  - Standard document sets (2GB SGML, 10 GB Web HTML, 436 GB planning to increase to 1 TB, and more.....)
  - Relevance assignments for 50 queries each year
  - Ad-hoc: evaluation using new queries
  - Routing: evaluation using new documents
  - Other tracks: CLIR, Multimedia, Question Answering, Biomedical Search, etc.
  - Check out: <http://trec.nist.gov/>

© Goharian, Grossman, Frieder, 2009

## User Interaction

- User query box and various functionalities
- Query suggestion
- Query expansion
- Providing snippets of documents to users
- Providing the ranked retrieved list
- Highlighting important terms
- Displaying the translated results
- .....

© Goharian, Grossman, Frieder, 2009

## Efficiency

- How **fast** index is built
- How fast each query is answered (*query response time*)
- How many queries are answered within a unit of time (*query throughput*)
- How collection size and number of users are handled (*scalability*)

© Goharian, Grossman, Frieder, 2009

## Efficiency

- Indexing
- Compression
- Index Pruning (Top Doc)
- Efficient Query Processing
- Duplicate Document Detection
- .....

© Goharian, Grossman, Frieder, 2009

## IR Engine Main Components

- IR engine has two main components
  - Indexing: to index documents
    - Most IR systems use a structure called an *inverted index* to index documents.
  - Query Processing: to accept and process queries.

© Goharian, Grossman, Frieder, 2009

## IR Engine Other Components

- Main components: Index builder & Query Processor
- Other components:
  - Crawler
  - Document conversion
  - Document data store
  - Tokenizer
  - Information extractor
  - Index distributor
  - Query broker
  - Logging

© Goharian, Grossman, Frieder, 2009

## Search Engine Requirements

- Scalability
  - Must handle large document collections
- Index Efficiency
  - Must build indexes in a reasonable amount of time
- Query Efficiency
  - Queries must run fast
- Query Effectiveness
  - Result set must be relevant

© Goharian, Grossman, Frieder, 2009

## Important IR References (Latest Research Papers on IR & DB)

### Journals

- ACM Transactions on Information Systems
- ACM Transactions on Database Systems
- J. of the American Society of Information Sci. & Tech. (JASIST)
- Information Retrieval Journal

### Conferences

- ACM Special Interest Group on Information Retrieval (SIGIR)
- ACM Conf. on Information and Knowledge Management (CIKM)
- European Conference on Information Retrieval (ECIR)
- Web Search and Data Mining Conference (WSDM)
- World Wide Web Conference (WWW)
- ACM Symposium on Applied Computing (SAC)
- ACM Special Interest Group on Management of Data (SIGMOD)
- Conference on Very Large Databases (VLDB)
- TREC (Text Retrieval Evaluation Conference)

© Goharian, Grossman, Frieder, 2009

## Information Retrieval Books

- G. Salton, Automatic Text Processing. Addison-Wesley, 1989.
- K. Sparck Jones & P. Willett, Readings in Information Retrieval. Morgan Kaufmann, 1997.
- I. Witten, A. Moffat, & T. Bell, Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann, Second Edition, 1999.
- R. Baeza-Yates & B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley 1999.
- D. Grossman & O. Frieder, Information Retrieval Algorithms and Heuristics. Springer 2004.
- C. Manning, P. Raghavan & H. Schütze, Introduction to Information Retrieval. Cambridge University Press. 2008.

© Goharian, Grossman, Frieder, 2009