Introduction to Information Retrieval

(COSC 488)
Fall 2013

Nazli Goharian
Information Retrieval Lab
Department of Computer Science
Georgetown University
nazli@ir.cs.georgetown.edu

Course Outline

- Introduction
- Retrieval Strategies (Models)
- Retrieval Utilities
- Evaluation
- Indexing
- Efficiency in indexing and query processing
- Integrating Structured Data and Text
- Distributed IR: Web
- Text Classification
- Recommender systems

What is Information Retrieval?

- Salton (1968):

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

Early Developments in IR

- Motivating factors: libraries, library science
- 50’s: Hans Luhn, Eugene Garfield, Philip Bagley, Calvin Moores
- 1962: First book on IR: Joseph Becker, Robert Hayes
- 60’s: Gerald Salton, Karen Spark Jones,..introduced concepts leading to today’s ranking in IR
- 1968: IR book by Gerard Salton
- 1978: First IR conference

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)

IR Tasks/Applications (Cont'd)

- Informational (ad hoc)- Enterprise/desktop/Web
- Navigational- Web
- Transactional- Web
- Question Answering
- Filtering/Routing
- Classification/Categorization

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

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.

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”

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”

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.

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?

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.

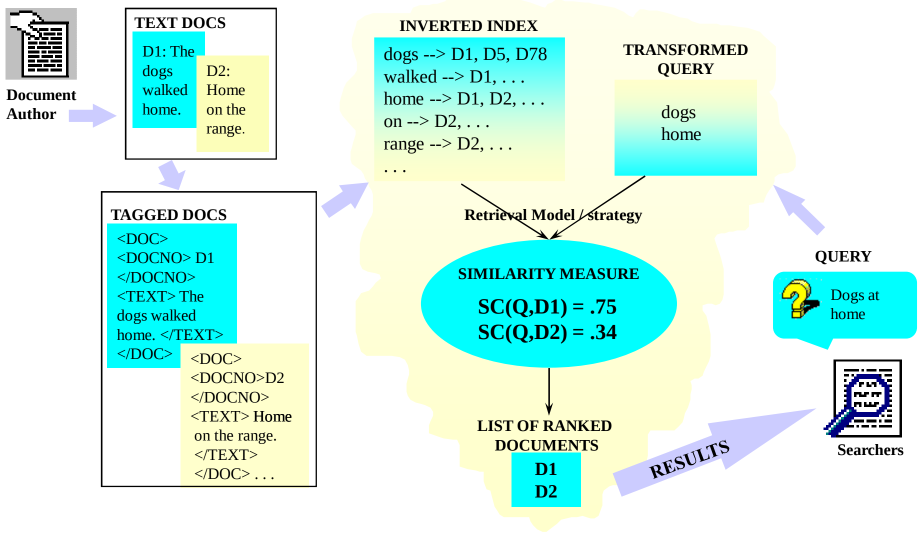
Retrieval Model/Strategy

- 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*)

Models/Strategies

- Manual
  - Boolean
- Automatic
  - Probabilistic
    - OKAPI BM25, Robertson/Sparck Jones
    - Kwok
  - Language Models
  - Vector Space Model
  - Inference Networks
  - Latent Semantic Indexing (LSI)
- Adaptive Models
  - Genetic Algorithms
  - Neural Networks

Measuring Relevance



Document Author

TEXT DOCS

D1: The dogs walked home.

D2: Home on the range.

TAGGED DOCS

<DOC>
<DOCNO> D1
</DOCNO>
<TEXT> The
dogs walked
home. </TEXT>
</DOC>

<DOC>
<DOCNO>D2
</DOCNO>
<TEXT> Home
on the range.
</TEXT>
</DOC> . . .

INVERTED INDEX

dogs --> D1, D5, D78
walked --> D1, . . .
home --> D1, D2, . . .
on --> D2, . . .
range --> D2, . . .
. . .

TRANSFORMED QUERY

dogs
home

Retrieval Model / strategy

SIMILARITY MEASURE

SC(Q,D1) = .75
SC(Q,D2) = .34

LIST OF RANKED DOCUMENTS

D1
D2

QUERY

Dogs at home

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
- …..

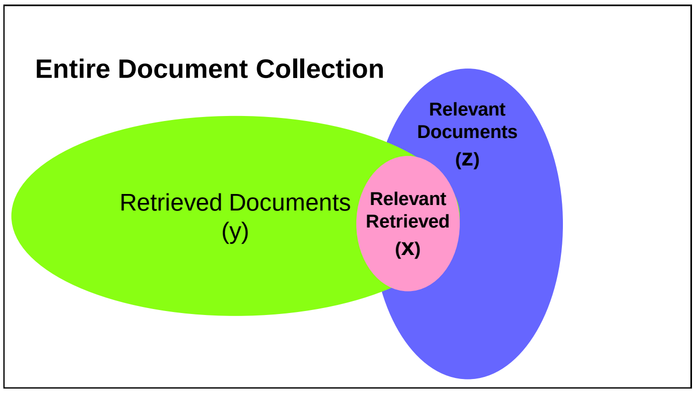
Evaluating Engine's Effectiveness

- *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!*

Precision / Recall

Precision = x / y

Recall = x / z



Existing Testbeds

- Cranfield (1970): A small (megabytes) domain specific testbed with fixed documents and queries, along with an exhaustive set of relevance judgment

- TREC (Text Retrieval Conference- sponsored by NIST; starting 1992): Various data sets for different tasks.
  – Most use 25-50 queries (topics)
  – Collections size (2GB, 10GB, half a TByte (GOV2), …….and 25 TB ClueWeb09)
  – No exhaustive relevance judgment

Existing Testbeds (Cont’d)

- GOV2 (Terabyte):
  – 25 million pages of web; 100-10,000 queries; 426 GB
- Genomics:
  – 162,259 documents from the 49 journals; 12.3 GB
- ClueWeb09 (25 TB):
  – Residing at Carnegie Mellon University, 1 billion web pages (ten languages). TREC Category A: entire; TREC Category B: 50,000,000 English pages)

- Text Classification datasets:
  – Reuters-21578 (newswires)
  – Reuters RCV1 (806,791 docs),
  – 20 Newsgroups (20,000 docs; 1000 doc per 20 categories)
  – Others: WebKB (8,282), OHSUMED(54,710), GENOMICS (4.5 million),….

TREC

- Text Retrieval Conference- sponsored by NIST
- Various benchmarks for evaluating IR systems.
- Sample tasks:
  - 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/

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
- ……

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*)

Efficiency

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

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.

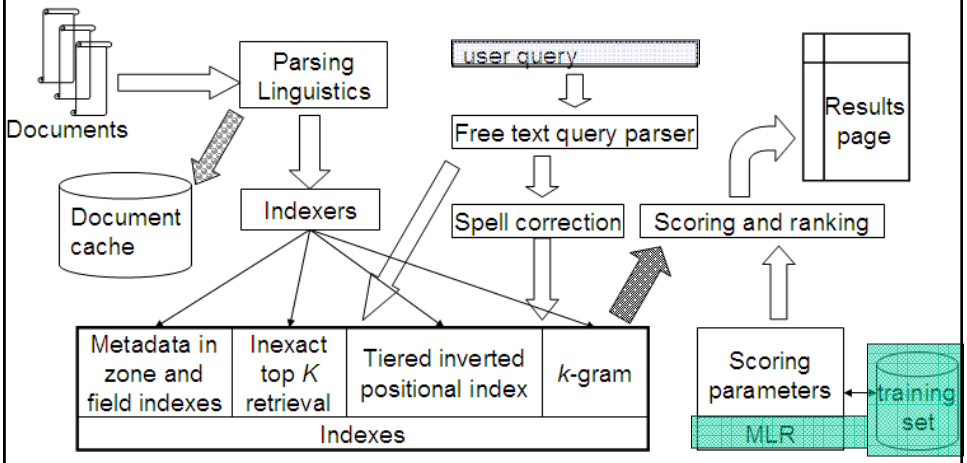
IR Engine Other Components

- Main components: Index builder & Query Processor
- Other components:
  - Crawler   (full vs. vertical)
  - Document conversion
  - Document data store
  - Tokenizer
  - Information extractor
  - Index distributor
  - Query broker
  - Logging

Sec. 7.2.4

Putting it all together (borrowed from:

©D. Manning, P. Raghavan, H. Schutze, Introduction to Information retrieval, p 135, Cambridge University Press., 2008.



Documents

Parsing Linguistics

Document cache

Indexers

user query

Free text query parser

Spell correction

Scoring and ranking

Results page

Metadata in zone and field indexes

Inexact top *K* retrieval

Tiered inverted positional index

k-gram

Indexes

Scoring parameters

MLR

training set

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

Important IR References
(Latest Research Papers on IR)

Journals

- ACM Transactions on Information Systems (TOIS)
- Journal of the American Society of Information Science & Technology (JASIST)
- Information Retrieval Journal
- ACM Transactions on Web
- Information Processing and Management (IP&M)
- IEEE Transactions on Knowledge and Data Engineering (TKDE)

Conferences

- ACM Special Interest Group on Information Retrieval (SIGIR)
- ACM Conf. on Information and Knowledge Management (CIKM)
- World Wide Web Conference (WWW)
- Web Search and Data Mining Conference (WSDM)
- European Conference on Information Retrieval (ECIR)
- ACM Symposium on Applied Computing (SAC)
- Joint ACM-IEEE Conference on Digital Libraries (JCDL)
- European Conference on Digital Libraries (ECDL)

Retrieval Evaluation Conferences

- Text REtrieval Conference (TREC)
- INitiative for the Evaluation of XML Retrieval (INEX)
- Cross Language Evaluation Forum (CLEF)

Information Retrieval Books

- G. Salton, Automatic Text Processing. Addison-Wesley, 1968, 2nd Edition, 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.
- D. Grossman & O. Frieder, Information Retrieval Algorithms and Heuristics, 1998, 2nd Edition, Springer, 2004.
- C. Manning, P. Raghavan & H. Schütze, Introduction to Information Retrieval. Cambridge University Press., 2008.
- B. Croft, D. Metzler, T. Strohman, Search Engines: Information Retrieval in Practice, The MIT Press, 2010
- S. Buttcher, C. Clarke, G. Cormack, Information Retrieval: Implementing and Evaluating search Engines, Addison Wesley, 2010
- R. Baeza-Yates & B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999, 2nd Edition, 2011