Distributed Memory and Data Retrieval

Instructor: Dmitri A. Gusev

Fall 2007

CS 502: Computers and Communications Technology

Lecture 23, November 28, 2007

The Need Of Memory Hierarchy (recap)

- Massive data storage requirements (e.g. a satellite orbiting Earth generates 1 terabyte data per day)
- Difference in the cost and capacity of different types of memory
- Data location

Data Location Factors

- Cost
- Performance
- Reliability
- Application software

Directory

- A directory is a mechanism to locate a data object
  - Directory location: may be different from the data location
  - Directory server: a separate system, not a part of the application program
  - Directory structure: usually hierarchical
  - Directory search: finding out the data location and making use of this fact (e.g. designing an application)
  - Providing access control

Directory Systems

- Pure approaches
  - Master control directory: a single directory containing all the information
  - Directory server: a single computer providing all directory functions (contains the master directory)
  - Fully replicated directories: copies of the directory at different locations (concurrency control required)
  - Local directories: contain information on local data and usually stored at the same location
  - Point of control directories: located at the point where the access control is exercised (e.g. hard disk)
- Problems with scaling: *hierarchical system of directories* (different directories at different levels)
- Important issues
  - Directory backup
  - Coherency and consistency
  - Local caching

Directories and Access Control

- Points of access control
  - Application program (problems with using data provided by other applications)
  - DBMS (the most common approach)
  - Control at the physical location of data (e.g. library)
  - Function of the communication networks (e.g. preventing a user from accessing another node, telephone system)

Information retrieval

- Finding relevant data using irrelevant keys
- Example: database of photographic images sorted by number, date.
- DBMS: Well structured data according to the information content

Text document retrieval (not well structured data)

Document retrieval queries

- Natural Language Processing (NLP): requires structured data to match the translated query
- Keyword search: boolean, weighted
- Abstracting documents

Evaluating retrieval quality

- Precision: retrieved and relevant /retrieved
- Recall: retrieved and relevant /relevant

Access methods

- Full text scanning: hardware support
- Keyword or phrase indexing: storage overhead (30%-70%), index updating
- Document signature: a fixed length bit string for each document (precision and recall easily computed)

Document classification and clustering

- Semantic (NLP, domain knowledge), statistical (keyword frequencies), mixed
- Similarity measures: attribute distance, information compression metrics
- The Machine Learning approach

Two extreme search paradigms

Searching a RDBMS

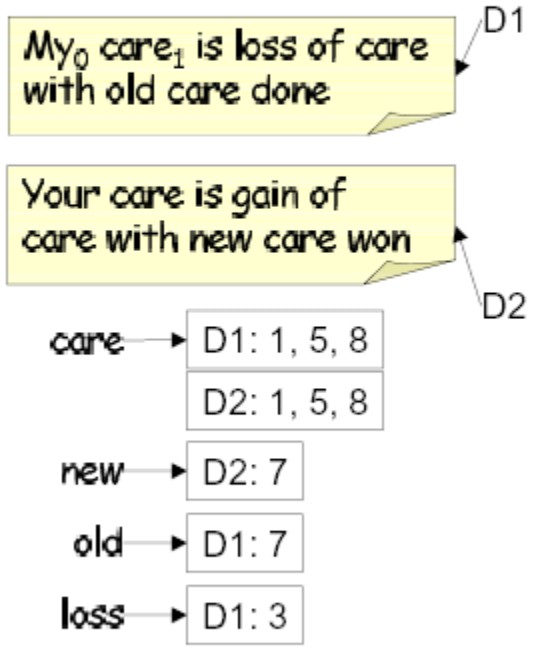
- Complex data model: tables, rows, columns, data types
- Expressive, powerful query language
- Need to know schema to query
- Answer = unordered set of rows
- Ranking: afterthought

Information Retrieval

- Collection = set of documents, document = sequence of terms
- Terms and phrases present or absent
- No (nontrivial) schema to learn
- Answer = sequence of documents
- Ranking: central to IR

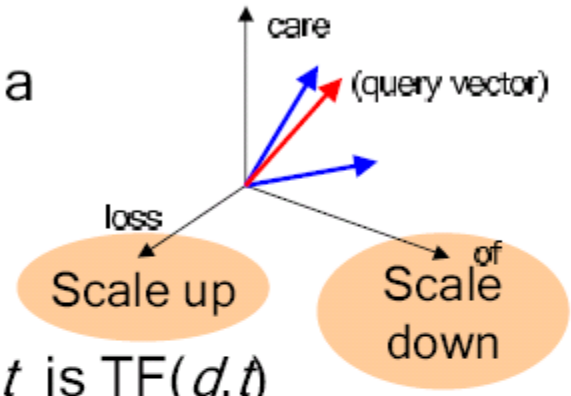
Text indexing basics

- "Inverted index" maps from term to document IDs
- Term offset info enables phrase and proximity ("near") searches
- Document boundary and limitations of "near" queries
- Can extend inverted index to map terms to
  - Table names, column names
  - Primary keys, RIDs
  - XML DOM node IDs



Information retrieval basics

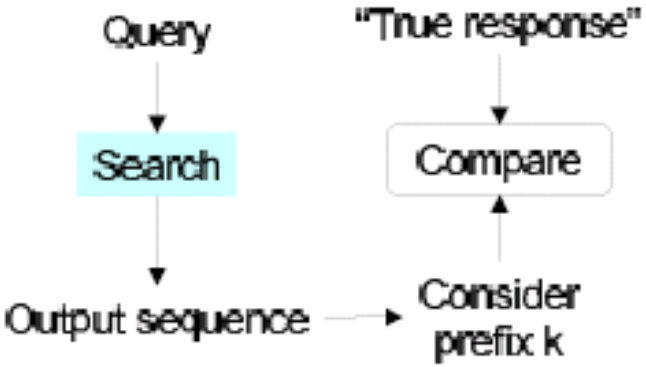
- Stopwords and stemming
- Each term t in lexicon gets a dimension in vector space
- Documents and the query are vectors in term space

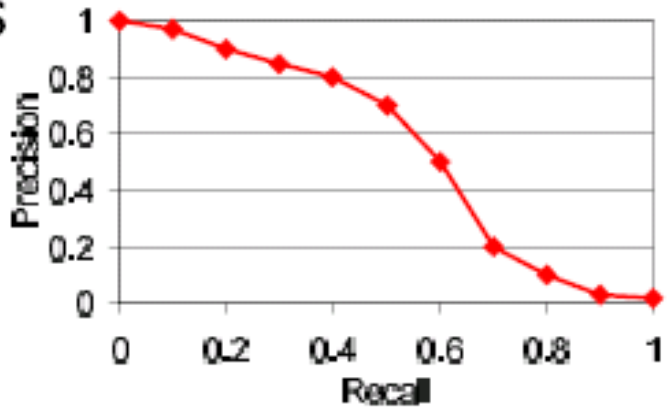


- Component of d along axis t is TF(d,t)
  - Absolute term count or scaled by max term count
- Downplay frequent terms: IDF(t) = log(1+|D|/|D_t|)
  - Better model: document vector d has component TF(d,t) IDF(t) for term t
- Query is like another "document"; documents ranked by cosine similarity with query

Relevance ranking

- Recall = coverage
  - What fraction of relevant documents were reported
- Precision = accuracy
  - What fraction of reported documents were relevant
- Trade-off
- 'Query' generalizes to 'topic'





Counting the cost

- Different types of classification errors often incur different costs.
- Example: predict cancer. Compare the cost of predicting "no" when the actual classification is "yes" and predicting "yes" when the actual classification is "no". Obviously the first error is much more costly.
- Confusion matrix:

Actual \ Predicted	yes	no
yes	True Positive (TP)	False Negative (FN)
no	False Positive (FP)	True Negative (TN)

- Total error = (FP+FN)/(TP+FP+TN+FN)
- Recall - precision (information retrieval):
  - Precision (retrieved relevant / total retrieved) = TP / (TP+FP)
  - Recall (retrieved relevant / total relevant) = TP / (TP + FN)