# Information Retrieval

Anne Denton

Department of Computer Science
North Dakota State University

# Outline

# Table of Contents

# Structured, semi-structured, and unstructured data

1. Relational databases intended for structured data
   - Modeling, normalization, join queries depend on rigorous structure
2. Unstructured data
   - Text, images, video
3. Semi-structured data
   - Markup languages allow adding some structure to unstructured information
   - XML and document database

# History of information retrieval

- Need for retrieval from large text repositories almost as old as databases
- Information retrieval concepts proposed in a 1968 book by Gerald Salton
- Topic of Library and Information Science
- Spawned off web search
  - Internet data is effectively unstructured data
  - Link structure offers some additional information that is used in web search
  - Differences in size of data caused development of new techniques

# Characteristics of information retrieval systems

- Allow free-form search requests or keyword search request
- Even inexperienced users develop their own queries
- Approximate measures used in evaluating matches
- Search results in list of pointers to documents
- Results typically ranked for relevance
- Rich data operations

  Statistical approaches give weights to word occurrences, e.g. vector space and probabilistic models

  Semantic approaches use sentence structure and lexical information

# Table of Contents

# Information retrieval pipeline

- Steps in preparing data
  - Preprocessing
  - Modeling
  - Indexing
- Steps in searching
  - Query formation
  - Query processing
  - Search mechanism
- Other components
  - Relevance feedback can improve system
  - Metadata integration from external ontologies

# Preprocessing

**Stopword removal:** Words like “the", “is", “at" don’t contribute to searches and are removed (at least in conventional statistical models)

**Stemming:** Different grammatical forms of a word represented by one term

**Data cleaning:** Removing hyphens, digits, etc.

## Question 1 (multiple answers can be correct)

In information retrieval, a stop word is

1. a word that signifies the end of a sentence
2. a word that is so frequent that it does not say much about the content of the document
3. a word that is typically at the end of a combination of frequent words

# Data modeling

- Conventional statistical models ignore grammatical structure or word sequences

  Bag-of-words model: Treat text as a bag of words
- Even with the bag-of-words approximation there are many possible models

  Boolean model: Only considers existence of term in documents

  Vector space model: Considers frequency of terms in document vs. corpus

  Probabilistic model: Computes likelihood that document is in relevant set

# Indexing

- Construction of an inverted index (Boolean model)
- Word occurrences extracted from document

| | $W_1$ | $W_2$ | $W_3$ | |
|---|---|---|---|---|
| $D_1$ | 1 | 0 | 1 | ... |
| $D_2$ | 1 | 1 | 0 | ... |
| $D_3$ | 0 | 0 | 1 | ... |

- Inverted index

| | $D_1$ | $D_2$ | $D_3$ | |
|---|---|---|---|---|
| $W_1$ | 1 | 1 | 0 | ... |
| $W_2$ | 0 | 1 | 0 | ... |
| $W_3$ | 1 | 0 | 1 | ... |

- Note that formally this is the transpose of the above matrix not the inverse

## Question 2 (multiple answers can be correct)

An inverted index in information retrieval is an index that

1. links a document to all the words it contains
2. links a word to all the documents that contain it
3. is based on the matrix inverse of the word frequency matrix

# Ontologies

- Terms in a text can mean a variety of things
  - Cat can stand for Computer Aided Tomography or for a small furry animal
- The concept of markup allows creating schemas with a defined meaning and providing a level of structure
  - Tools for semi-structured data, e.g. XML, JSON, YAML
- Subject domains may create “ontologies", i.e. defined vocabularies, typically hierarchical, e.g., Gene Ontology
  - `https://en.wikipedia.org/wiki/Gene_ontology`
- Ontologies often get represented through markup language schemas, including for the Gene Ontology
- Another example of an ontology / markup language / data format, is the spatial data format kml

  `https://developers.google.com/kml`

# Table of Contents

# Similarity between documents

- Document similarity typically evaluated using a “Vector-space Model"
- Considers term frequencies, not only Boolean or binary (i.e., 0/1) existence of a term
- Term frequencies are normalized based on the overall frequency of the term
- The cosine of the angle between the vectors for two documents is a measure of its similarity
  - Since vectors are the same length, only the angle between them matters
  - In linear algebra it is shown that the dot product between vectors ($x_0 * y_0 + x_1 * y_1 + \ldots$) is the cosine of the angle between them

# Towards quality measures

- Used in much of statistics and data science
- For Boolean quantities, such as relevant vs. irrelevant, four outcomes are possible

  True positive (TP) Returned result is in relevant set
  False positive (FP) Returned result is not in relevant set
  False negative (FN) Relevant result is not returned
  True negative (TN) Irrelevant result is not returned
- Can be combined into many quality measures
- For ranked results cutoff can be selected arbitrarily
  - Each cutoff corresponds to different TP, FP, FN, and TN designations

## Question 3

Consider a search for “CAT scans" as a medical procedure. If a result were returned that relates to browsing through listings of house cats would be considered a

1. True positive
2. False positive
3. False negative
4. True negative

## Question 4

Consider a search for “CAT scans" as a medical procedure. If a highly relevant result about the medical procedure failed to show up in the result it would be considered a

1. True positive
2. False positive
3. False negative
4. True negative

# Quality measures

- Precision
  - Fraction of returned documents that is relevant
  - Precision=TP/(TP+FP)
- Recall (also called sensitivity)
  - Fraction of relevant documents that are returned
  - Recall=TP/(TP+FN)
- These quality measures apply to any predicted true vs. actually true comparisons in data science
- Other data science applications also consider specificity (=TN/(TN+FP)), but in IR it is usually too close to 1
- Measures for ranked results
  - There are also measures to combine all cutoffs, in particular the “Area under the Receiver Operating Characteristic" or AUC

## Question 5

Assume that out of 10 documents that are returned, 4 are relevant to your query. Also assume that 50 documents in the corpus in total are considered relevant, and that there are 1000 documents in the corpus. How many false positive results are there?

1. 4
2. 996
3. 46
4. 6

## Question 6

Assume that out of 10 documents that are returned, 4 are relevant to your query. Also assume that 50 documents in the corpus in total are considered relevant, and that there are 1000 documents in the corpus. How many false negative results are there?

1. 4
2. 996
3. 46
4. 6

## Question 7

Assume that out of 10 documents that are returned, 4 are relevant to your query. Also assume that 50 documents in the corpus in total are considered relevant, and that there are 1000 documents in the corpus. What is the “precision" of the result?

1. 0.4
2. 0.004
3. 0.05
4. 0.08

## Question 8

Assume that out of 10 documents that are returned, 4 are relevant to your query. Also assume that 50 documents in the corpus in total are considered relevant, and that there are 1000 documents in the corpus. What is the “recall" of the result?

1. 0.4
2. 0.004
3. 0.05
4. 0.08

# Table of Contents

# Differences between web search and IR systems

- Importance of link structure
  - Until link structure was considered, web search was largely useless
  - Ask an old person (like myself) about Altavista
  - Google PageRank is considered to be among the most important algorithms ever invented
- Size
  - Challenges of indexing the web have prompted new distributed processing paradigms
  - MapReduce framework of process, in which data are distributed and processing is pushed to the data, and newer frameworks, such as implemented in Apache Mahout

# PageRank

- Represents probability that a person who is randomly clicking on links will arrive at a particular page
- Related to number of ingoing links
- Also considers indirect references, i.e. ingoing links into pages with ingoing links
- Continues recursively
- `https://en.wikipedia.org/wiki/PageRank`

# Statistical vs. semantic approaches in web search

- Web search mostly relies on statistical measures
- Goes beyond bag-of-words model through word n-grams
  - Combinations of words are considered together
  - “cat scan" not the same as “scan cat"
  - Google maintains frequent word n-grams up to length 5 (large volume of data, 11 DVDs already years ago)
- Allowed treating problems “big data" problems that had been considered semantic, e.g. translation
  - Machine translation now extensively existing translations on the web

## Question 9 (multiple answers can be correct)

Distinguishing “the cat ate the mouse" from “the mouse ate the cat"

1. can be done using an analysis of grammar that has the typical English word order encoded
2. can be done using the bag-of-word model
3. can be done using a vector-space model for words that incorporates word frequencies
4. can conceivably be done using word n-grams, provided at least one relevant subset of words occurs frequently enough in the respective order

## Question 10

The early success of Google search, in comparison with earlier search engines like Altavista, was primarily due to the use of

1. grammar analysis
2. fast processing frameworks like MapReduce or Apache Mahout
3. vector-space word models instead of bag-of-word models
4. n-gram word models that encode the combination of terms
5. the link structure of hypertext links