
CS100: Introduction to Computer Science

Lecture 22: Information Retrieval, Question
Answering & Web Search

Review: Database and Data mining

- 1. What is a database?
 - 2. What operations can be performed on a database?
 - 3. Database management systems
 - 4. Database applications
 - 5. What is data mining?
 - 6. What are the basic tasks in data mining?
-

More Computer Applications

- Information Retrieval
 - Question Answering
 - Web search engines
-

Information Retrieval

- ***Information Retrieval (IR)***: retrieving desired information from documents.
 - Text retrieval
 - Match a query against free-text documents
 - Retrieve relevant documents
 - Retrieve relevant passages
 - Retrieve relevant sentences
 - Music retrieval
 - How to match a query to music?
 - Image retrieval
 - Retrieve images from a large database of digital images
 - Content based image retrieval – based on visual similarity
 - Traditional methods – based on keywords, captions, descriptions of images
-

Information Retrieval

- **Query:**

- *Short key-word based query (2-3 words)*
- *Long descriptive query (one or more sentence)*

- **Examples**

- <title> Unexplained Highway Accidents
- <desc> Description:

Identify documents that discuss fatal highway accidents where the cause of the accident cannot be determined.

Information Retrieval

- **Query:**
 - **Relevance**
 - **True relevance:** *A relevant document meets user's information need*
 - **Relevance score:** *A numeric score assigned to a search result, representing how well the result "match" the query.*
 - **Similarity:** measure of how close a query is to a document.
-

Information Retrieval (cont'd)

- ***Similarity Measures:***
 - *Dice, jaccarb, cosine, overlap*
 - *The similarity is being evaluated between two vectors*
 - Documents which are “close enough” to a query are retrieved.
 - Many similarity measures are used in data mining to mine text/web data.
-

Similarity Measures

Dice: $sim(t_i, t_j) = \frac{2\sum_{h=1}^k t_{ih}t_{jh}}{\sum_{h=1}^k t_{ih}^2 + \sum_{h=1}^k t_{jh}^2}$

Jaccard: $sim(t_i, t_j) = \frac{\sum_{h=1}^k t_{ih}t_{jh}}{\sum_{h=1}^k t_{ih}^2 + \sum_{h=1}^k t_{jh}^2 - \sum_{h=1}^k t_{ih}t_{jh}}$

Cosine: $sim(t_i, t_j) = \frac{\sum_{h=1}^k t_{ih}t_{jh}}{\sqrt{\sum_{h=1}^k t_{ih}^2 \sum_{h=1}^k t_{jh}^2}}$

Overlap: $sim(t_i, t_j) = \frac{\sum_{h=1}^k t_{ih}t_{jh}}{\min(\sum_{h=1}^k t_{ih}^2, \sum_{h=1}^k t_{jh}^2)}$

Similarity Measures

- Determine similarity between two objects.
- Similarity characteristics:

- $\forall t_i \in D, sim(t_i, t_i) = 1$
- $\forall t_i, t_j \in D, sim(t_i, t_j) = 0$ if t_i and t_j are not alike at all.
- $\forall t_i, t_j, t_k \in D, sim(t_i, t_j) < sim(t_i, t_k)$ if t_i is more like t_k than it is like t_j .

- Alternatively, distance measure measure how unlike or dissimilar objects are.

Similarity Measures — origin in measuring similarity between sets based on the intersection of the two sets

- Dice's coefficient
 - Relates the overlap to the average size of the two sets
 - Jaccard's coefficient
 - Relates the overlap to the size of the union
 - Cosine' coefficient
 - Relates the overlap to the geometric average of the two sets
 - Overlap
 - Determines to which degree the two sets overlap
-

Information Retrieval (cont'd)

- **Term Frequency**

- ***simply the number of times a given term appears in a document***
- *The importance of the term in the document.*

- **Inverse Document Frequency**

- a measure of the general importance of the term
- it is the logarithm of the number of all documents divided by the number of documents containing the term.

- **tf-idf weight**

- ***evaluate how important a word is to a document in a collection***
 - Many different ways to calculate the weight
 - *High tf-idf weight reached by High tf and low idf, filter out common terms*
-

Information Retrieval-metrics

- Metrics:

- **Precision** = $\frac{|\text{Relevant and Retrieved}|}{|\text{Retrieved}|}$

- *The proportion of retrieved and relevant documents to all the documents retrieved:*

- **Recall** = $\frac{|\text{Relevant and Retrieved}|}{|\text{Relevant}|}$

- The proportion of relevant documents that are retrieved, out of all relevant documents available
-

Information Retrieval-metrics

■ **F-measure**

- The weighted harmonic mean of precision and recall, the traditional F-measure or balanced F-score is:
- $F = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

■ **Average Precision for a query**

- Calculate by averaging precision as recall increases.
- Widely used in IR.

■ **Mean Average Precision**

- The average of many queries' average precision values
-

Example: Evaluate an IR system

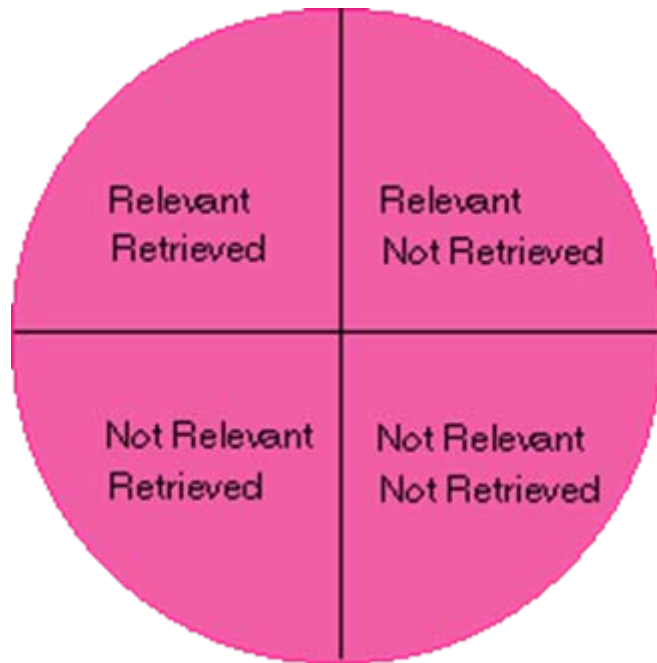
- Query 1: (Total number of relevant documents is 5)
 - Ranked list: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
 - Query 2: (Total number of relevant documents is 5)
 - Ranked list: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
 - **How to calculate** Recall, Precision, Average Precision, Mean Average Precision
-

Precision and Recall Applied to Classification

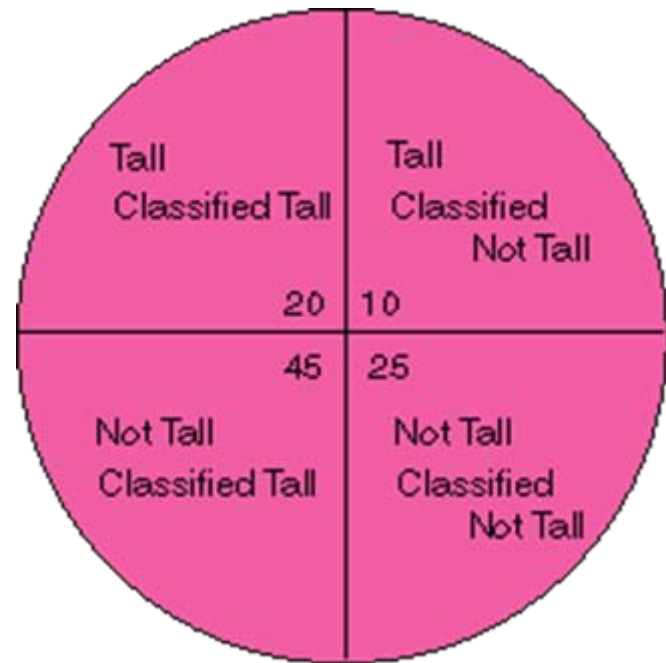
■ Example:

- Suppose 100 college students are to be classified based on height. In actuality, there are 30 tall students and 70 who are not tall. A classification techniques classifies 65 students as tall and 35 as not tall.
 - Calculate the precision and recall
-

IR Query Result Measures and Classification



IR



Classification

More About IR – Other models

■ Boolean Model

- Exact match
- Queries are logical expressions with document features as operands
- Boolean operators: AND, OR, AND-NOT

■ Statistical Language Model

- Each document has a language model M_D : a probability distribution over terms
 - Documents are ranked by $P(Q|M_D)$, probability of observing Q during random sampling from the language model of document D .
-

Cross-Lingual Retrieval

- Accepting query in one language (English), retrieving documents in another language (French).
 - Typical approach is to translate query and then use monolingual search engines
 - Language resources
 - Bilingual dictionary, parallel collection, MT systems.
 - Major issue: Translation ambiguity
 - Multiple translations for each word
 - Translation probabilities required for some approaches
-

Question Answering

- Question in QA vs. Query in IR
 - Exact answers vs. relevant documents for results
 - A typical QA system
 - Question analysis
 - Search engine (a revised IR system)
 - Answer extraction
-

Question Answering

- Require more complex natural language processing techniques
 - Named entity Extraction, Parser, Part-of-speech tagger
 - A wide range of question:
 - Factoid: (person, location, date, organization, money amount, etc.)
 - List, why, how, definition
 - Closed-domain QA & Open-domain QA
 - Domain specific knowledge, accept a limited type of questions
 - General knowledge, user can basically ask any type of question
-

Parser

- Find the rules of a grammar that are used to construct a sentence
- “John loves Mary”
- Given a grammar
 - Sentence \rightarrow noun_phrase, verb_phrase
 - Verb_phrase \rightarrow verb, noun_phrase
 - Noun_phrase \rightarrow det, noun
 - Noun_phrase \rightarrow p_name
 - Verb \rightarrow [love]
 - P_name \rightarrow [Mary]
 - P_name 00 \rightarrow [john]

Example questions

- Who is the president of United States?
 - How many members are in the U.S. congress?
 - When was Abraham **Lincoln** born?
 - Name 22 cities that have a subway system.
-

Example

- Question: When was Abraham **Lincoln** born?
 - **Abraham Lincoln** was **born Sunday, February 12, 1809**, in a log cabin near Hodgenville, Kentucky. He was the son of Thomas and Nancy Hanks Lincoln, and he was named for his paternal grandfather. Thomas Lincoln was a carpenter and farmer. Both of Abraham's parents were members of a Baptist congregation which had separated from another church due to opposition to slavery.
-

More Example

- Who is the president of the United States?
 - Chinese President Hu Jintao's visit to the United States is fruitful and of milestone importance to the bilateral relations, Chinese Foreign Minister Li Zhaoxing said on April 22.
-

Web Search Engine

- An important application of IR
 - Search in the large amount of data on the web
 - Pages with heterogeneous data and extensive hyperlinks
 - Multi-language
 - Various sources
 - Different formats
 - ...
-

Web Search Engines - problems

- Abundance
 - Too much data, a query only retrieves a small subset of it
 - Limited coverage
 - Search a result from a subset of the web.
 - Only periodically update the index
 - Limited query
 - Key-word based searching, works for short queries
 - Limited customization
 - Query results determined by the query itself.
 - IR systems consider the background and knowledge of the users as well.
-

Announcements:

- Lab 6 this week (optional)
 - Next lecture: Review session, guideline to the final exam.
-