CIS 560 – Database System Concepts

Document Retrieval and Inverted Indexes (Section 14.1.8)

Credits for slides: Hofmann, Mihalcea, Mobasher, Mooney, Schutze.

Information Retrieval

- The processing, indexing and retrieval of textual documents.
- Concerned firstly with retrieving *relevant* documents to a query.
- Concerned secondly with retrieving from *large* sets of documents *efficiently*.

Key Terms Used in IR

- Query: a representation of what the user is looking for - can be a list of words or a phrase.
- Document: an information entity that the user wants to retrieve
- Collection or corpus: a set of documents
- Index: a representation of information that makes querying easier
- Term: word or concept that appears in a document or a query

Typical IR Task

Given:

- A corpus of textual natural-language documents
- A user query in the form of a textual string

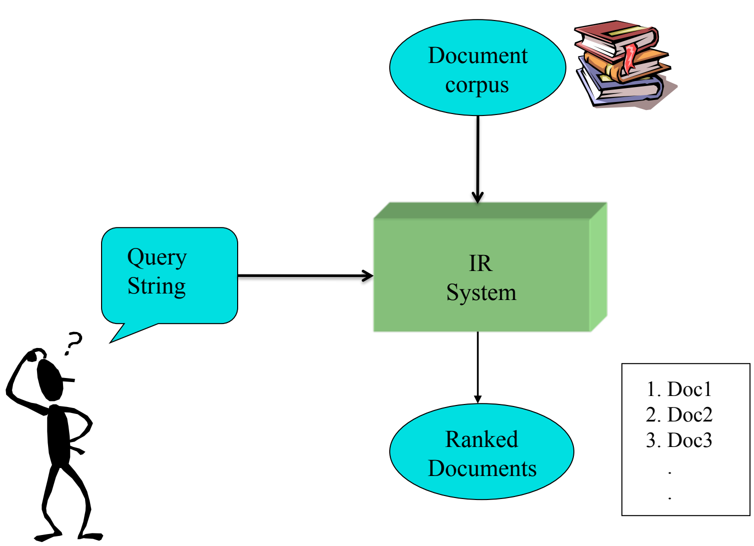
Find:

- A ranked set of documents that are relevant to the query

Relevance

- Relevance is a subjective judgment and may include:
  - Being on the proper subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the goals of the user and his/her intended use of the information (*information need*)
- Main relevance criterion: an IR system should fulfill user's information need

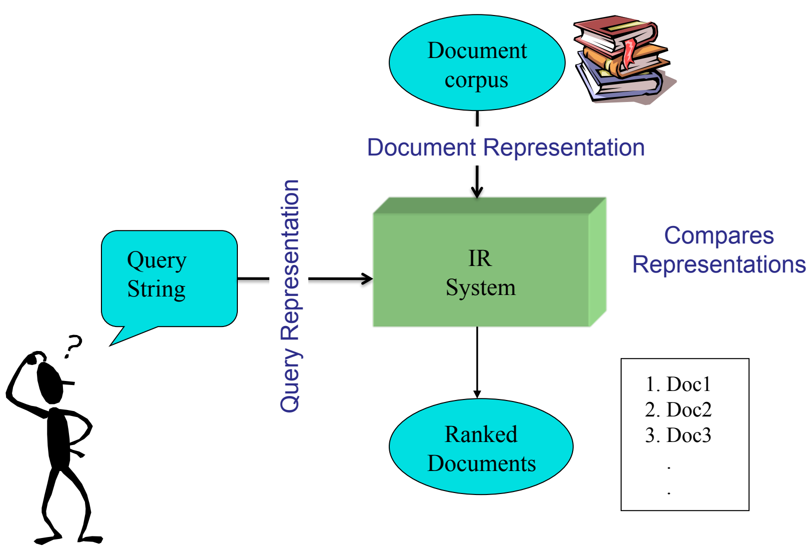
Typical IR System Architecture



Retrieval Models

- A retrieval model specifies the details of:
  - Document representation
  - Query representation
  - How do we compare representations - retrieval function?
- Determines a notion of relevance.
- Notion of relevance can be binary or continuous (i.e. *ranked retrieval*).

Typical IR System Architecture



Classes of Retrieval Models

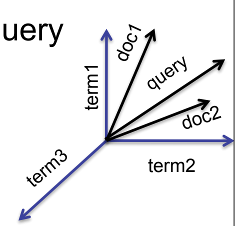
- Boolean models (set theoretic)
  - Extended Boolean
- Vector space models (algebraic)
  - Generalized VS
  - Latent Semantic Indexing
- Probabilistic models
  - Inference Networks
  - Belief Networks

Exact match

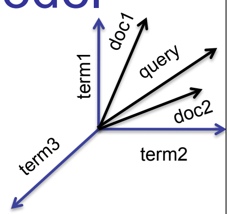
Ranking - “Best” match

Vector Space Model

- Key idea: Everything (documents, queries, terms) is a vector in a high-dimensional space.
- Geometry of space induces a similarity measure between documents
- Rank documents based on their similarity with query
- History:
  - Invented by Gerald Salton (1960/70)
  - Lucene (popular open source engine written in Java)
  - Most Web search engines are similar

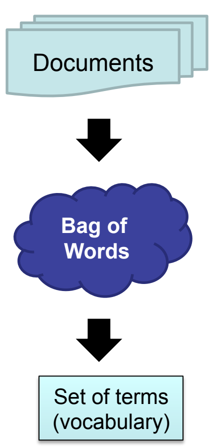


Issues for Vector Space Model



- How to determine important words in a document?
  - How to select basis vectors (dimensions)
- How to convert objects into vectors?
  - Documents, queries, terms
- Assumption - not all terms are equally useful for representing the document contents, less frequent terms allow identifying a narrower set of documents
  - The *importance* of the index terms is represented by weights associated to them.
  - How to determine the degree of importance of a term within a document and within the entire collection?
- How to compare objects in the vector space?
  - How to determine the degree of similarity between a document and the query?
- In the case of the web, what is a collection and what are the effects of links, formatting information, etc.?

Preprocessing



tokenization, case folding, stopword removal, stemming

~~syntax~~, ~~semantics~~, ~~word knowledge~~, etc.

The Vector-Space Model

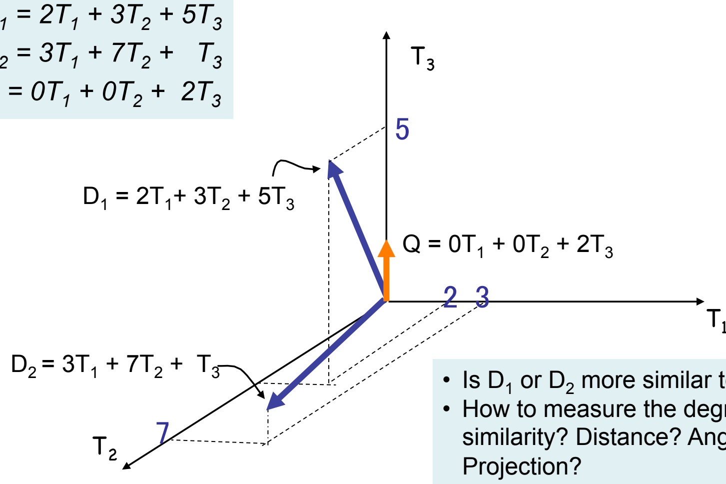
- Assume *t* distinct terms remain after preprocessing; call them index terms or the vocabulary.
- These "orthogonal" terms form a basis of a vector space.

  Dimension = *t* = |vocabulary|
- Each term, *i*, in a document or query, *j*, is given a real-valued weight, w_{ij}.
- Both documents and queries are expressed as *t*-dimensional vectors:

  $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$

Example Graphical Represenation

$D_1 = 2T_1 + 3T_2 + 5T_3$
$D_2 = 3T_1 + 7T_2 + T_3$
$Q = 0T_1 + 0T_2 + 2T_3$



- Is D_1 or D_2 more similar to Q?
- How to measure the degree of similarity? Distance? Angle? Projection?

Document Collection

- A collection of *n* documents can be represented in the vector space model by a term-document matrix.
- An entry in the matrix corresponds to the "weight" of a term in the document; zero means the term has no significance in the document or it simply doesn't exist in the document.

$$
\begin{array}{c}
\\ D_1 \\ D_2 \\ : \\ : \\ D_n
\end{array}
\begin{pmatrix}
T_1 & T_2 & \ldots. & T_t \\
w_{11} & w_{21} & \ldots & w_{t1} \\
w_{12} & w_{22} & \ldots & w_{t2} \\
: & : & & : \\
: & : & & : \\
w_{1n} & w_{2n} & \ldots & w_{tn}
\end{pmatrix}
$$

Term Weights: Term Frequency

- More frequent terms in a document are more important, i.e. more indicative of the topic.

  f_{ij} = frequency of term i in document j

- May want to normalize *term frequency* (*tf*)
  - e.g. by dividing by the frequency of the most common term in the document:

    $tf_{ij} = f_{ij} / max_i\{f_{ij}\}$

Term Weights: Inverse Document Frequency

- Terms that appear in many *different* documents are *less* indicative of overall topic.

  df_i = document frequency of term i

  = number of documents containing term i

  idf_i = inverse document frequency of term i,

  $= \log_2 (N/ df_i)$

  (N: total number of documents)
- An indication of a term's *discrimination* power.
- Log used to dampen the effect relative to *tf*.

TF-IDF Weighting

- A typical combined term importance indicator is *tf-idf weighting*:

  $$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N/ df_i)$$
- A term occurring frequently in the document but rarely in the rest of the collection is given high weight.
- Many other ways of determining term weights have been proposed.
- Experimentally, *tf-idf* has been found to work well.

Computing TF-IDF -- An Example

Given a document containing terms with given frequencies:

A(3), B(2), C(1)

Assume collection contains 10,000 documents and document frequencies of these terms are:

A(50), B(1300), C(250)

Compute tf, idf, tf-idf?

$w_{ij} = tf_{ij}\ idf_i = (f_{ij} / max_i\{f_{ij}\}) * \log_2 (N/ df_i)$

Query Vector

- Query vector is typically treated as a document and also tf-idf weighted.
- Alternative is for the user to supply weights for the given query terms.
  - Weighted query terms:
    Q = < database 0.5; text 0.8; information 0.2 >
  - Unweighted query terms:
    Q = < database; text; information >

Similarity Measure

- A similarity measure is a function that computes the *degree of similarity* between two vectors.
- Using a similarity measure between the query and each document:
  - It is possible to rank the retrieved documents in the order of presumed relevance.
  - It is possible to enforce a certain threshold so that the size of the retrieved set can be controlled.

Vector Space Similarity: Common Measures

Sim(X,Y)	Binary Term Vectors	Weighted Term Vectors
Inner product	$\lvert X \cap Y \rvert$	$\sum x_i \cdot y_i$
Dice coefficient	$\frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum x_i \cdot y_i}{\sum x_i^2 + \sum y_i^2}$
Cosine coefficient	$\frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert}\sqrt{\lvert Y \rvert}}$	$\frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}}$
Jaccard coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum x_i \cdot y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i \cdot y_i}$

Inner Product

- Similarity between vectors for the document $\boldsymbol{d}_i$ and query $\boldsymbol{q}$ can be computed as the vector inner product (a.k.a. dot product):

$$\text{sim}(\boldsymbol{d}_j,\boldsymbol{q}) = \boldsymbol{d}_j \bullet \boldsymbol{q} = \sum_{i=1}^{t} w_{ij} w_{iq}$$

where w_{ij} is the weight of term i in document j and w_{iq} is the weight of term i in the query

- For binary vectors, the inner product is the number of matched query terms in the document (size of intersection).
- For weighted term vectors, it is the sum of the products of the weights of the matched terms.

Properties of Inner Product

- The inner product is (usually) unbounded.
- Favors long documents with a large number of unique terms.
- Measures how many terms matched but not how many terms are *not* matched.

Inner Product -- Examples

Binary:

	retrieval	database	architecture	computer	text	management	information
D =	1,	1,	1,	0,	1,	1,	0
Q =	1,	0 ,	1,	0,	0,	1,	1

Size of vector = size of vocabulary = 7
0 means corresponding term not found in document or query

sim(D, Q) = ?

Weighted:

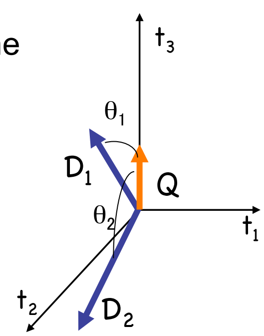
$D_1 = 2T_1 + 3T_2 + 5T_3$ $D_2 = 3T_1 + 7T_2 + 1T_3$
$Q = 0T_1 + 0T_2 + 2T_3$

sim(D_1 , Q) = ?
sim(D_2 , Q) = ?

Cosine Similarity Measure

- Cosine similarity measures the cosine of the angle between two vectors.
- Inner product normalized by the vector lengths.

$$\text{CosSim}(\mathbf{d}_j, \mathbf{q}) = \frac{\vec{d_j} \cdot \vec{q}}{\left|\vec{d_j}\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^{2} \cdot \sum_{i=1}^{t} w_{iq}^{2}}}$$



$D_1 = 2T_1 + 3T_2 + 5T_3$ CosSim(D_1 , Q) = ?
$D_2 = 3T_1 + 7T_2 + 1T_3$ CosSim(D_2 , Q) = ?
$Q = 0T_1 + 0T_2 + 2T_3$

Vector Space Summary

- Very simple
  - Map everything to a vector
  - Compare using angle between vectors
- Challenge is mostly finding good weighting scheme
  - Variants on tf·idf are most common
- Another challenge is comparison function
  - Cosine comparison is most common
  - Generic inner product (without unit vectors) also occurs
- Considers both local (*tf*) and global (*idf*) word occurrence frequencies.
- Provides partial matching and ranked results.
- Tends to work quite well in practice despite obvious weaknesses.

Problems with Vector Space Model

- Missing semantic information (e.g. word sense).
- Missing syntactic information (e.g. phrase structure, word order, proximity information).
- Assumption of term independence (e.g. ignores synonymy).
- Lacks the control of a Boolean model (e.g., *requiring* a term to appear in a document).
  - Given a two-term query “A B”, may prefer a document containing A frequently but not B, over a document that contains both A and B, but both less frequently
- Implementation?

Naïve Implementation

Convert all documents in collection D to tf-idf weighted vectors, $\boldsymbol{d}_j$, for keyword vocabulary V.

Convert query to a tf-idf-weighted vector $\boldsymbol{q}$.

For each $\boldsymbol{d}_j$ in D do

Compute score $s_j = \text{cosSim}(\boldsymbol{d}_j, \boldsymbol{q})$

Sort documents by decreasing score.

Present top ranked documents to the user.

Time complexity?

Practical Implementation

- Based on the observation that documents containing none of the query keywords do not affect the final ranking
- Try to identify only those documents that contain at least one query keyword
- Actual implementation of an inverted index