Introduction to Digital Libraries

hussein suleman
uct cs honours 2003

Quick Intro. to Information Retrieval

Introduction

- Information retrieval is the process of locating the most relevant information to satisfy a specific information need.
- Traditionally, librarians created databases based on keywords to locate information.
- The most common modern application is search engines.
- Historically, the technology has been developed from the mid-50's onwards, with a lot of fundamental research conducted pre-Internet!

Terminology

- Term
  - Individual word, or possibly phrase, from a document.
- Document
  - Set of terms, usually identified by a document identifier (e.g., filename).
- Query
  - Set of terms (and other semantics) that are a machine representation of the user's needs.
- Relevance
  - Whether or not a given document matches a given query.

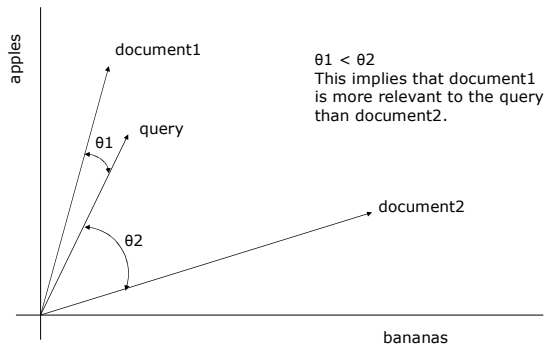
More Terminology

- Indexing
  - Creating indices of all the documents/data to enable faster searching.
- Searching
  - Retrieving all the possibly relevant results for a given query.
- Ranked retrieval
  - Retrieval of a set of matching documents in decreasing order of estimated relevance to the query.

Models for IR

- Boolean model
  - Queries are specified as boolean expressions and only documents matching those criteria are returned.
    - e.g., digital AND libraries
- Vector model
  - Both queries and documents are specified as lists of terms and mapped into an n-dimensional space (where n is the number of possible terms). The relevance then depends on the angle between the vectors.

Vector Model in 2-D



Naïve Vector Implementation

apples	Doc1: 15 Doc2: 5	19
bananas	Doc1: 4 Doc2: 20	24

- An inverted file for a term contains a list of document identifiers that correspond to that term.
- When a query is matched against an inverted file, the document weights are used to calculate the similarity measure (inner product or angle).

tf.idf

- Term frequency (tf)
  - The number of occurrences of a term in a document – terms which occur more often in a document have higher tf.
- Document frequency (df)
  - The number of documents a term occurs in – popular terms have a higher df.
- In general, terms with high "tf" and low "df" are good at describing a document and discriminating it from other documents – hence tf.idf (term frequency * inverse document frequency).

Implementation of Inverted Files

- Each term corresponds to a list of weighted document identifiers.
  - Each term can be a separate file, sorted by weight.
  - Terms, documents identifiers and weights can be stored in an indexed database.
- Search engine indices can easily take 2-6 times as much space as the original data.
  - The MG system (part of Greenstone) uses index compression and claims 1/3 as much space as the original data.

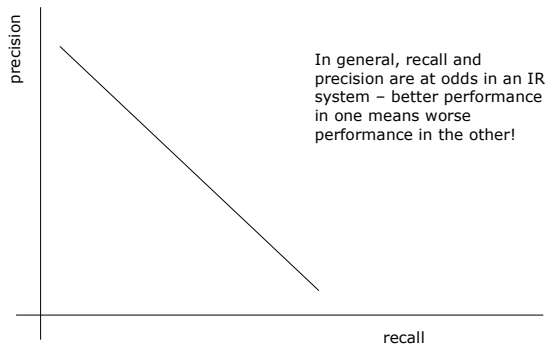
Clustering

- In term-document space, documents that are similar will have vectors that are close together.
- Even if a specific term of a query does not match a specific document, the clustering effect will compensate.
- Centroids of the clusters can be used as cluster summaries.

Recall and Precision

- Recall
  - The number of relevant results returned.
  - Recall = number retrieved and relevant / total number relevant
- Precision
  - The number of returned results that are relevant.
  - Precision = number retrieved and relevant / total number retrieved
- Relevance is determined by an "expert" in recall/precision experiments. High recall and high precision are desirable.

Typical Recall-Precision Graph



In general, recall and precision are at odds in an IR system – better performance in one means worse performance in the other!

Filtering and Ranking

- Filtering
  - Removal of non-relevant results.
  - Filtering restricts the number of results to those that are probably relevant.
- Ranking
  - Ordering of results according to calculated probability of relevance.
  - Ranking puts the most probably relevant results at the "top of the list".

Extended Boolean Models

- Any modern search engine that returns no results for a very long query probably uses some form of boolean model!
  - Altavista, Google, etc.
  - Vector models are not as efficient as boolean models.
- Some extended boolean models filter on the basis of boolean matching and rank on the basis of term weights (tf.idf).

Term Preprocessing

- Case Folding
  - Changing all terms to a standard case.
- Stemming
  - Changing all term forms to canonical versions.
    - e.g., studying, studies and study map to "study".
- Stopping
  - Stopwords are common words that do not help in discriminating in terms of relevance.

PageRank

- PageRank (popularised by Google) determines the rank of a document based on the number of documents that point to it, implying that it is an "authority" on a topic.
- In a highly connected network of documents with lots of links, this works well. In a diverse collection of separate documents, this will not work.
- Google uses other techniques as well!

Thesauri

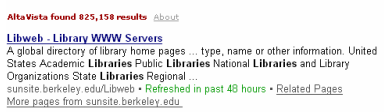
- A thesaurus is a collection of words and their synonyms.
  - e.g., According to Merriam-Webster, the synonyms for "library" are "archive" and "athenaeum".
- An IR system can include all synonyms of a word to increase recall, but at a lower precision.
- Thesauri can also be used for cross-language retrieval.

Metadata vs. Full-text

- Text documents can be indexed by their contents or by their metadata.
- Metadata indexing is faster and uses less storage.
- Metadata can be obtained more easily (e.g., using OAI-PMH) while full text is often restricted.
- Full-text indexing does not rely on good quality metadata and can find very specific pieces of information.

Relevance Feedback

- After obtaining results, a user can specify that a given document is relevant or non-relevant.
- Terms that describe a (non-)relevant document can then be used to refine the query – an automatic summary of a document is usually better at describing the content than a user.


AltaVista found 825,158 results About

Libweb - Library WWW Servers
A global directory of library home pages ... type, name or other information. United States Academic **Libraries** Public **Libraries** National **Libraries** and Library Organizations State **Libraries** Regional ...
sunsite.berkeley.edu/Libweb • Refreshed in past 48 hours • Related Pages
More pages from sunsite.berkeley.edu

Inference Engines

- Machine learning can be used to digest a document collection and perform query matching.
  - Connectionist models (e.g., neural networks)
  - Decision trees (e.g., C5)
- Combined with traditional statistical approaches, this can result in increased recall/precision.

Web Crawlers

- Web crawlers are often bundled with search engines to obtain data from the WWW.
- Crawlers follow each link (respecting robots.txt exclusions) in a hypertext document, obtaining an ever-expanding collection of data for indexing/querying.
- WWW search engines operate as follows:

Implications for Information Systems

- Free-text search should use an IR system – not a database and not keywords!
- Indexing and searching are two separate operations and require intermediate storage (for inverted files).
- Search engines can be obtained as components.
  - e.g., Lucene, Swish-E