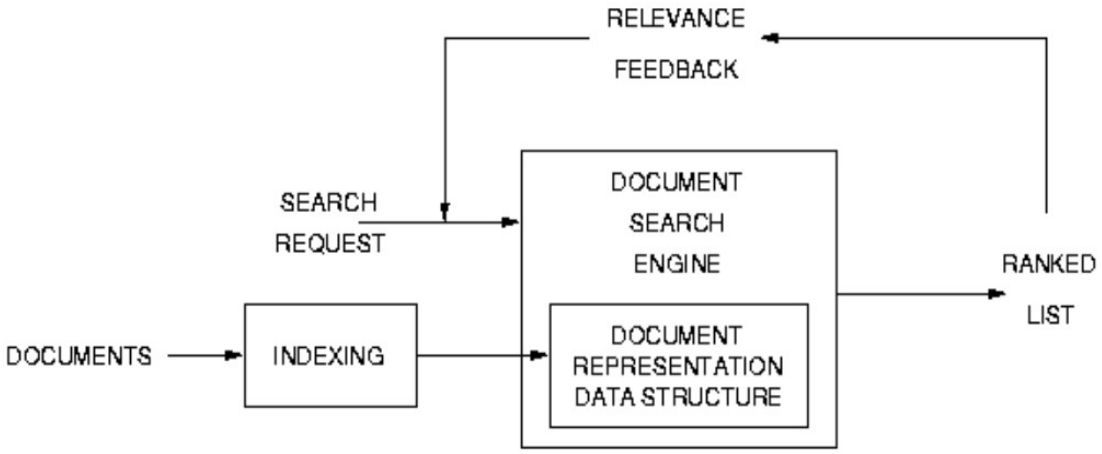
Search Tech Notes

IR Systems

Definition:

- The purpose of an IR system is to satisfy a users information need.
- Seeks to locate documents relevant to this information need

Components of an text-retrieval system:



- Document collection:
  - Gather documents from sources, need to be pre-procesed, stop word removal, stemming, markup removal etc..
- Document indexing
  - convert into fast searchable format
- Search request
  - use similar text preprocessing as for documents
- Document searching
  - calculate set of potentaily relevant documents in a ranked order of relevance.
- Relevance feedback
  - modify search and re run

Conflation:

- Designed to bring toghter words that are related to each other in some way.
- Two main classes:
  - Stemming algorithms
  - String-simularity measures
- Need for conflation
  - Incorrect spelling
  - Alternitive spelling
  - Multi-word concepts
  - Transliteration
  - Affixes

Stemming Algorithms:

- matching the ending of a word against a suffix dictonary
- removing any suffix identified
- Checking whether context-sensitive rules apply

Two approaches to comparison:

- Longest match: a word such as ALARMINGLY is stemmed to ALARM if both -INGLY and -LY are in the dictionary
- Iterative Search: ALARMINGLY is first stemmed to ALARM by removal of -LY and then to ALARM by removal of -ING

Context-sensitive rules list exception to stem removal rules

Recoding can be used to cover changes in spelling that occur from removal of suffixes

Stemming problems:

Obvious stemming error

Word	Stem
MEDICAL	MED
MEDIA	MED
MEDIAN	MED

Over- and under- stemming

Word	Stem
GASES	GAS (correct)
GAS	GA (over)
GASEOUS	GASE (under)

Language dependant issues:

Splitting compounds or **decompounding** in productive languages such as German is important words are joined toghter to create new meanings these compounds can be rare or unique in documents and requests are unlikely to match. Usefull search terms must be extracted by **decompounding** the word (splitting).

Its is important to **segment** sentances in **agglutinating** languages as these do not have spaces in between words and as such would need to segment them to generate search terms.

Indexing:

- convert into fast searchable format

Inverted file index

- File structure to represent indexed files

Example:

After stop word removal the following words remain:

Word	Word No
computer	T1
database	T2
information	T3
management	T4
retrieval	T5
Systems	T6

Representing in their original form:

Document	Contents
D1	T3,T5,T6
D2	T2,T4,T6
D3	T1,T3,T5,T6

inverted file:

Term	Document
T1	D3
T2	D2
T3	D1,D2
T4	D2
T5	D1,D3
T6	D1,D2,D3

Hashing Algorithm

Used to efficiently compute term locations

Must fulfil the following conditions:

- it must be repeatable
- ideally have an even distribution
- minimise synonyms.

Retriaval models

- Boolean – Probabilistic model
  - attempts to compute the probability that a document is relevant to a query
  - ranked in decreasing order of probability of relevance
- Best match – vector space model
  - Documents and queries are represented as vectors in a t dimensional space. Where t is the number of unique index terms in the collection.
  - Similarity is calculated between document and query as the cosine of the angle between the two vectors
  - documents are ranked in decreasing order of similarity

Term Weighting

- Collection Frequency
  - Terms that occur in only a few documents are often more valuable than ones which occur in many documents
  - Also known as Inverse document frequency(IDF)
  - $$cfw(i) = \log \frac{N}{n(i)}$$
  - n(i) = the number of documents term t(i) occurs in,
  - N = the total number of documents in the collection archive
- Term Frequency
  - The more often a term occurs in a document, the more likely it is to be important
  - tf(i,j) = the number of occurrences of term t(i) in document d(j)
- Document Length
  - Document relevance is independent of document length
  - dl(j) = the total number of term occurences in document d(j)
  - $$ndl(j) = \frac{dl(j)}{average\ dl\ for\ all\ documents}$$

OKAPI BM25

$$cw(i,j) = cfw(i) \times \frac{tf(i,j) \times (k_1 + 1)}{k_1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

- cw(i,j) indicates the combined weighting scheme

- k and b are experimentally determined constants that control the offect of tf(i,j) and the degree of length normalisation respectively

Relevance feedback

Relevance feedback uses relevance information from an initial retriaval run for the current query to:

- add significant terms from relevant documents to the current query
- modify the weights of terms to improve the ranks of documents which are likely to be relevant.

Relevance information can come from three possible sources:

- Explicit feedback: The user can mark documents from the current run as relevant or non-relevant
- Implicit feedback: The system can assume that returned documents clicked on by the user are relevant.
- Blind or pesudo feedback: The system can simply assume the top ranked documents are all relevant.

In vector space model:

Query expansion can be seen as adjusting term weights in the query vector.

In probabilistic model:

Term re-weighting and query expansion are treated separately.

Computing relevance weights seeks to answer the question:
“How much evidence does the presence of this term provide for the relevance of this document?”

Selecting terms to add to a query should answer:
“how much will adding this term to the request benefit the overall performance of the search formulation?”

Evaluation of IR systems

1. does it work correctly?
2. Is it useful?
3. Is it better that other systems which attempt to do the same thing?

Evaluation of 2 and 3 must adopt a planned strategy. Should rely on evidence from individual users with individual requests.

- Qualitative evaluation: From user feedback
- Quantitative evaluation: From user tests and accuracy of document retrieval.

Recall: $\frac{\text{No of relevant docs retrieved}}{\text{total relevant docs in collection}}$

- Proportion of the overall relevant documents retrieved

Precision $\frac{\text{No of relevant docs retrived}}{\text{Total docs retrived}}$

- Proportion of retrieved documents are relevant.

Summarization

Definitions:

Summary: a condensed derivative of a source text

Selection: forming a summary focused on a subset of the topical content of the source document in detail.

Generalization: Forming a summary which overviews the entire topical contents of the source document.

Factors to take into account to form effective snippet summaries are:

- The form of the source text
- The purpose of the summary
  - What is the function of the summary
  - is the summary for a narrow targeted or more general audience
  - what level of subject knowledge should be assumed.

Enterprise search

Refers to the application of search technologies to information within an organisation.

Intended for use within an organisation by employees or other authorises seeking information held internally in a variety of formats and potentially at a number of locations.

Constaints:

- Considerations of security
- inablity to index specialised content(images, flash files)
- Difficulty intergrating structured and non-structured content
- Cost, time, difficulty.

Types of users:

- class 1: Members of an organisation who may be familiar with the information or documents they are searching for.
- Class 2: looking for information held within the organisation, but they do no know where it may be found.
- Class 3: Third parties

Access control:

“early binding security” - Access control attributes are stored when the document is indexed

“late binding security” = each entry in the results list is checked at display time.

Document Clustering

Documents can be clustered bassed simply on metadata fields or based on their content.

Users can browse clusters of documents rather then looking one by one.

Faceted Search

Technique of accessing information by filtering items based on facets of the information. Each facet typically corresponds to the possible values of a property common to all objects.

Recommender Systems

Seek to predict items that will be of interest to a user.

Make their decision based on previous feedback or ratings from the user to whom the recommendation are made and/or other users.

Two categories:

- content-based filtering (CDF)
  - analysis the contents of a set of items rated by the user and use the contents of the items as well as the users ratings of the items
  - information is used to create a user profile that can be usd to recommend items
- collaborative filtering (CF)
  - use ratings from multiple previous users to make recommendations
  - provide recommendations based on ratings of items provided by other users who share common interests to the current user.
  - Ratings information is gatered in two ways:
    - explicitly: user provides rating
    - implicitly: ratings inferred from user behaviour with the item.
  - Two approaches:
    - memeory-based algorithms
      - compute simularity between each user and current user and select closest neighbors to current user
    - model-based algorithms:
      - construct a model to represent the behaviour of the users to predict
      - ratings

Problems:

- Sparsity of ratings
- Cold start problem: Two types: User side and item-side problems
  - User-side problem relates to new users who have so far provided little rating information
  - Item-side problem where new items have not been rated often enough
- Spam attacks

Multimedia search

research Seeks to:

- better understand user needs and their cognitive abilities and preferences for media interaction.
- Develop technologies to create effective multimedia search systems

Automatic speech recognition (ASR):

Seeks to generate imperfect index information for documents

Challanges:

- Speech variability
  - Each utterances of a word is unique and vary in many ways
  - speech patterns of different people vary.
- speaker variability
  - accents
- acoustuc ambiguity
  - to, too, two
    - homophones, sound the same
  - bee and pea,
    - very small acoustic distinction
- continuous speech
- context-dependency
  - all words can be broken down into small set of constituent sounds

two fundamental components:

- Acoustic models – statistical models of all the possible speech sounds of the language that is being spoken
- a language model- a statistical model of the expected word sequences of the language

Content-based retrieval of visual media

- Successful content-based retrieval systems are generally task or domain specific
- Automatic understanding tool have been to date impossible to develop
- Systems are thus typically highly interactive involving users developing queries through iterative search to steer the system towards relative items (human in the loop)

The semantic gap refers to the gap between the contents of a multimedia data stream and its meaning as interpreted by humans.

Retrieval of Image data

- level 1 :image primitives
  - based on extracted features e.g. colour, texture, shape or spacial placement
- level 2:iconography
  - describes a picture's actual contents derived attributes such as the presence of specific objects e.g. chairs around a table, or names specific individual
- level 3: iconology
  - describes a picture's deeper, artistic significance inferred abstract attributes which do not correspond directly to content in the image but to some inferred attribute e.g. if we have footballers, a football and a goal post we have "A football match"

A **shot boundary** is crossed when a new camera is used or a recording instance ends and a new one begins(shot cuts).
A simple process of shot boundary detection is by examining every frame and looking for shot cuts which can be based on a change in colour, texture or intensity/ brightness of the adjacent frame.
Problems typically encountered are:

- Dropped shot boundaries:
  - fade-in and fade-out
  - dissolving
  - morphing
  - wipes
  - etc
- False shot boundaries:
  - zooming and panning
  - tilting
  - booming and tracking
  - events in the content – camera flashes

Four methods of selecting key frames from a shot can be:

- take the First or last frame
- Choosing the middle frame
- Choosing the frame with the most average colour histogram from the shot
- A virtual centroid – a frame that does not actually exists, but contains the average of all colour data in the shot.