# Information Retrieval

## Lecture 2: Models of Document Retrieval and Text Representation

Computer Science Tripos Part II


UNIVERSITY OF CAMBRIDGE

Simone Teufel

Natural Language and Information Processing (NLIP) Group

sht25@cl.cam.ac.uk


(Lecture Notes after Stephen Clark)

## Document Retrieval



- Retrieve documents with content that is relevant to a user's information need
- Document set is fixed
  (size can vary from 10s of documents to billions)
- Information need is not fixed (ad-hoc retrieval)

Goal:

- Documents relevant to query should be returned
- documents not relevant to query should not be returned

## Models of Retrieval



- A model is an abstraction of a process
- A retrieval model can be a description of the human process or the computational process of retrieval
  - the process by which information needs are articulated and refined
  - the process of choosing documents for retrieval
- Here we focus on the description of the computational process

## Models of the Retrieval Process



- Boolean Model
  - simple, but common in commercial systems
- Vector Space Model
  - popular in research; becoming common in commercial systems
- Probabilistic Model
- Statistical language models
- Bayesian inference networks

Assumption shared by all models: both document content and query can be represented by a bag of terms

## Some Terminology



- Document: an item which may satisfy the user's information need
  - task specific: web pages, news reports, emails, . . .
- Query: representation of user's information need
  - initial query formulated by user . . .
  - transformed into final query used for search
- Term: any word or phrase that can serve as a link to a document

## Boolean Model



- Boolean model is simple, but is popular in commercial systems (e.g. bibliographic databases)
- Queries consist of terms connected by:
  AND ($\wedge$), OR ($\vee$), NOT ($\neg$)
- Key assumptions (and weaknesses)
  - Terms are either present or absent in a document
    (frequency not taken into account)
  - Terms are all equally informative when determining relevance
  - A document is either relevant or not relevant
    (no partial matches)

## Example Boolean Retrieval



- User need: I'm interested in learning about vitamins other than vitamin e that are anitoxidants
- User query:
  vitamin AND antioxidant AND NOT vitamin e
- Suppose there's a document which discusses the antioxidant properties of vitamin e and vitamin c
  - does the user want it?
  - does the user get it?

## Advantages and Disadvantages of the Boolean Model



Advantages:

- Simple framework; semantics of a Boolean query is well-defined
  - can be implemented efficiently
  - works well with well-informed user

Disadvantages:

- Complex queries often difficult to formulate
- difficult to control volume of output
- no ranking facility
  - may require trained intermediary
  - may require many reformulations of query

## Vector Space Model



- Documents, Queries, Terms are all vectors in some high-dimensional vector space
- Key questions:
  - What are the basis vectors (dimensions)?
  - What is magnitude along a dimension?
  - How can objects in the space be compared?

## Basis Vectors

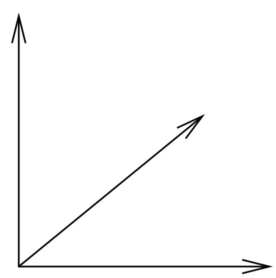


- A Vector Space is defined by a linearly independent set of basis vectors
  - A set of vectors $\{\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_n\}$ is linearly independent if the only solution to the equation $\lambda_1\overline{v}_1 + \lambda_2\overline{v}_2 + \ldots + \lambda_n\overline{v}_n = 0$ is $\lambda_i = 0$ for all $i$
  - Each vector in the set cannot be expressed as a linear combination of the remaining vectors
- Any vector in the space can be expressed as a linear combination of the basis vectors
  - basis vectors determine what objects can be described in the space

## Orthogonal Basis Vectors



- If $\overline{v} \cdot \overline{w} = 0$ then $\overline{v}$ and $\overline{w}$ are orthogonal
  - $\overline{v} \cdot \overline{w} = \Sigma_i \, v_i.w_i$
  - $\overline{v} \cdot \overline{w} = |\overline{v}||\overline{w}| \cos\theta$
- If a set of vectors is pairwise orthogonal then it is linearly independent



Basis vectors for 3 dimensions

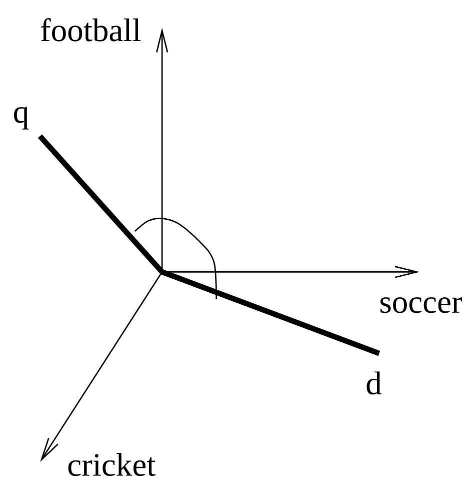
## Terms as Basis Vectors

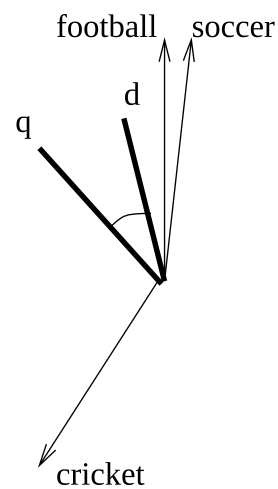


- Typically terms from the document set are chosen as orthogonal basis vectors
- But terms are clearly not orthogonal (?)
- If we believe terms are “not orthogonal”, we must have some pre-defined notion of a space in which terms exist
- What are the basis vectors of this space?
  - concepts?
    - ∗ documents, queries, terms are linear combinations of concepts?
  - Latent Semantic Indexing is an example of a technique which considers this question

## The Problem with Orthogonal Terms






- Document d mentions soccer and cricket; query q mentions football and cricket
- Relaxing the orthogonality assumption brings d closer to q in the space

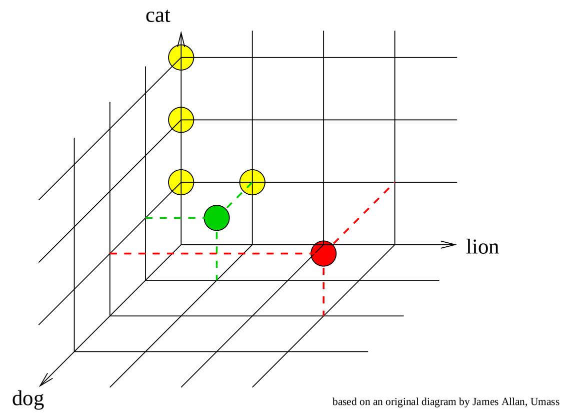
## The Problem with Orthogonal Terms



- Synonymy: two documents with similar content can contain different words and be far apart in the space
  - problem of synonymy may adversely affect recall
- Polysemy: two documents can share many words, and hence be close in the space, but have very different content
  - problem of polysemy may adversely affect precision
- However, despite many attempts to improve upon the orthogonality assumption, systems which make this assumption are hard to beat

## Document Representation using Term Frequency






- yellow: {cat}, {cat cat}, {cat cat cat}, {cat lion}
- green: {cat lion dog}
- red: {cat dog dog lion lion lion}

## Weighting Schemes



- Weighting scheme determines position of documents and queries in the space
  - ideally we want similar objects clustered together, and dissimilar objects far apart
- Individual vector components for an object determine:
  - the degree to which the object embodies that dimension
  - possibly the usefulness of that dimension in distinguishing the object

e.g. TF $\times$ IDF

- Small IDF for a term effectively shrinks the space in that dimension, making it less important

## TF $\times$ IDF 

- Term frequency: (monotonic function of) the number of times a term appears in a document
  - can be thought of as a recall enhancing device
- Inverse document frequency: (monotonic function of) the number of documents in which a term appears
  - can be thought of as a precision enhancing device
- Why not use inverse collection frequency: the total number of occurrences of a term across all documents?
  - a term may have a high collection frequency but still be concentrated in a small set of documents

## A Possible TF $\times$ IDF Scheme 

- $\mathrm{TF}_{i,j} = 1 + \log(\mathit{tf}_{i,j})$ $(\mathit{tf}_{i,j} > 0)$
  - $\mathit{tf}_{i,j}$ is frequency of $i$th term in $j$th document
- $\mathrm{IDF}_i = \log \frac{N}{\mathit{df}_i}$
  - $N$ is number of documents in collection
  - $\mathit{df}_i$ is the number of documents in which $i$th term appears
- Many variations of TF $\times$ IDF exist
  - Salton and Buckley (1988) give generalisations about effective weighting schemes

## Query Representation



- Queries are also vectors in the space, based on the terms in the query
- Query term weights need not be the same as document term weights
- A possible query-term weighting scheme (Salton and Buckley, 1998):
  - TF $\times$ IDF = $(0.5 + \frac{0.5_{tf}}{\max_{tf}}) \log \frac{N}{n}$
    - *tf* is the number of times term appears in query
    - $\max$ tf is the highest number of occurrences for any term in the query
    - $N$ is the total number of documents
    - $n$ is number of documents in which query term appears

## Similarity Measures



- How similar are the query and document vectors?
- Inner product
  - D $\cdot$ Q = $\Sigma_i\, d_i.q_i$
  - documents containing many instances of informative query terms score highly
- Cosine
  - cosine(D,Q) = $\frac{D \cdot Q}{|D||Q|}$
  - length normalised inner product
    - measures angle between vectors
  - prevents longer documents scoring highly simply because of length
- There are alternative similarity measures

## Summary of Vector Space Model



- Simple model
  - map documents and queries to vectors
  - compare using angle between the vectors
  - intuitive geometric interpretation is appealing
- Main challenge is finding a good weighting scheme
  - model provides no guidance
  - TF $\times$ IDF schemes have been found to work well empirically
- Vector Space Model: References
  - Relevant sections from the course textbook
  - Sparck Jones and Willett, eds., Introduction to Chapter 5
  - Baeza-Yates & Ribeiro-Neto, Chapter 2
  - Term-Weighting Approaches in Automatic Text Retrieval, Salton and Buckley, Ch.6 in Sparck Jones and Willett

## Document and Text Representation: the Index



- Manually searching a book for a desired topic is possible, but tedious and time-consuming
  - indexes help a reader quickly locate a topic
- Exhaustive automatic search of very large document collections is expensive
  - indexes greatly speed-up the search process
- Indexing:
  - the process of building the index
    - inverted file, signature file, . . .
  - deciding what will be used to represent the documents
    - need to facilitate good matching of queries and relevant documents

## Inverted files



| Doc 1 |
|---|
| Except Russia and Mexico no country had had the decency to come to the rescue of the government. |

| Doc 2 |
|---|
| It was a dark and stormy night in the country manor. The time was past midnight. |

| Term | Doc no | Freq | Offset |
|---|---|---|---|
| a | 2 | 1 | 2 |
| and | 1 | 1 | 2 |
| and | 2 | 1 | 4 |
| come | 1 | 1 | 11 |
| country | 1 | 1 | 5 |
| country | 2 | 1 | 9 |
| dark | 2 | 1 | 3 |
| decency | 1 | 1 | 9 |
| except | 1 | 1 | 0 |
| government | 1 | 1 | 17 |
| had | 1 | 2 | 6,7 |
| in | 2 | 1 | 7 |
| it | 2 | 1 | 0 |
| manor | 2 | 1 | 10 |
| mexico | 1 | 1 | 3 |
| midnight | 2 | 1 | 17 |
| night | 2 | 1 | 6 |
| no | 1 | 1 | 4 |
| of | 1 | 1 | 15 |
| past | 2 | 1 | 15 |
| rescue | 1 | 1 | 14 |
| russia | 1 | 1 | 1 |
| stormy | 2 | 1 | 5 |
| the | 1 | 2 | 8,13 |
| the | 2 | 2 | 8,12 |
| time | 2 | 1 | 14 |
| to | 1 | 2 | 10,12 |
| was | 2 | 2 | 16 |

Information kept for each term:

- Document ID where this term occurs
- Frequency of occurrence of this term in each document
- Possibly: Offset of this term in document

## Manual Indexing



- Controlled vocabularies can be used to determine index terms
- Examples: MeSH, Library of Congress, Gene Ontology, ...
- e.g. Decubitus Ulcer could also be referred to using *Bedsore, Pressure Ulcer, Pressure Sore* ...
  – MeSH: Bacterial Infections -surgery; Cross Infection -surgery; Decubitus Ulcer -complications; ...
- Single term can describe an ambiguous concept
- Human indexers determine what the important concepts are, and what terms can denote those concepts
- Ontologies (e.g., MeSH, GO) organise concepts hierarchically
  – can contain many relations between concepts, e.g. hyponymy, meronymy
  – Structure of ontology can be used to aid search (specificity)

## Automatic Indexing



- Manual creation of controlled vocabulary, and maintenance of associated document collection, is very expensive
- Automatic indexing program decides which words or phrases to use as terms from the documents themselves
- Program may even determine concepts and synonymous terms automatically (automatic thesaurus construction)
- Cranfield experiments in the 60s (Cleverdon papers in Sparck Jones and Willett): automatic indexing can be at least as effective as manual indexing with controlled vocabularies
- perhaps counter-intuitive
- general message: much can be achieved with "shallow" content representations

## Document Representation



- Represent a document using index terms – what should these be?
- Example 1: "The colinearity of genes in Hox clusters suggests a role for chromosome structure in gene regulation."
  - should "gene" and "genes" be separate terms?
  - should "the", "of", "in", "a", 'for", "in" be terms?
  - should "chromosome structure", "gene regulation", "Hox clusters" be single terms?
- Example 2: "By using retinoic acid (RA) to induce regulated expression of Hoxb genes in mouse embryonic stem (ES) cells, . . ."
  - should "regulation" and "regulated" be separate terms?
  - should "retinoic acid" and "RA" be terms?
  - should "embryonic stem (ES) cells", 'embryonic stem cells", "stem cells", "ES cells" be single terms?

## Tokenisation



- Tokenisation: dividing a character stream into a sequence of distinct word forms (tokens)
- Just separate on white-space?
  - end-of-sentence punctuation:
    "role for chromosome structure in gene regulation."
    "Apple Computer, Inc."
  - bracketing: "By using retinoic acid (RA) . . ."
  - hyphenation: "By using this state-of-the-art technique . . ."
  - apostrophes: "The biologist's hypothesis doesn't imply . . ."
  - slashes: "The body has a potassium/sodium mixture . . ."
  - . . .
- Languages other than English may need additional processing, e.g. segmentation for Chinese

## Stop Words



- A stop word is a high-frequency word which is not useful for distinguishing between documents
- Some of the PubMed stop words:
  *a, did, it, perhaps, these, about, do, its, quite, they, again, does, itself, rather, this, all, done, just*
- Substantially reduces the size of an inverted file index
  - Statistics for TREC documents: (Witten et al.)
    - 33 terms appear in more than 38% of documents
    - 33 terms account for 30% of all term appearances
    - and acount for 11% of pointers in inverted file
- Use of stop words can be problematic
  - Stop words can be names ("The Who"), or frequent words with other meanings (*may, can, will*)

## Terms and Equivalence Classes



- Can be useful to put tokens into equivalance classes, and treat a group of terms as the same term
  - reduces size of index
  - may lead to improved retrieval; e.g. if query is {*gene*, *regulation*} may help to retrieve pages containing *regulate, regulated, regulates*, . . .
  - the combined frequencies of class terms may better reflect the content than the individual frequencies

## Morphology and Stemming



- Stem: the core of a word (its main morpheme) to which inflectional and derivational morphology applies
  - inflectional morphology deals with such things as plurality and tense: *employ* $\rightarrow$ *employed, employs, employing*, . . .
  - derivational morphology deals with obtaining nouns from verbs, adjectives and verbs from nouns, etc.: *fool* $\rightarrow$ *foolish*, *advert* $\rightarrow$ *advertise*, . . .
- Stemming attempts to remove inflectional (and some) derivational morphology
- Lemmatisation just attempts to remove inflectional morphology
- Morphology is a serious issue for e.g. Arabic, Turkish

## Porter Stemmer



- Well known rule-based stemmer
- Example original sentence: Document will describe marketing stategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, perdicted sales, market share, stimulate demand, price cut, volume of sales
- Porter-stemmed (minus stop words): market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale stimul demand price cut volum sale

(Example from James Allan, Umass)

## Porter Stemmer



- Output appears non-sensical – why does it work?
  - representation of root word only needs to be unique for the relevant class of words
  - and transformation needs to be repeatable
- Porter stemmer is sometimes too agressive
  - e.g. *policy/police, execute/executive, organize/organic*
- Porter stemmer sometimes misses good conflations
  - e.g. *European/Europe, matrices/matrix, machine/machinery*
- Literature gives contrasting evidence about whether stemming helps
  - but improving stemmers still an active research area

## Term Similarity Metrics



- Terms can be compared using a string similarity metric, e.g. edit distance
  - More general way to obtain morphological variants
  - Also deals with other variations, e.g. typos/mis-spellings
  - Variants of *Britney Spears* entered as Google queries over a 3-month period:

    *488941 britney spears 40134 brittany spears 36315 brittney spears 24342 britany spears 7331 britny spears 6633 briteny spears 2696 britteny spears 1807 briney spears 1635 brittny spears 1479 brintey spears 1479 britanny spears 1338 britiny spears 1211 britnet spears 1096 britiney spears 991 britaney spears 991 britnay spears 811 brithney spears 811 brtiney spears 664 birtney spears*

## Phrases, Multi-Word Terms



- Why use multi-word phrases as index terms?
  - search for *New York* may be improved if *New York* is in the index . . .
  - since we don't want documents about York and New Jersey, for example
- How do we determine the multi-word terms?
  - could do it manually, but expensive
  - automatically:
    - ∗ observe word combinations in a large corpus of text
    - ∗ extract multi-word terms on the basis of some statistic, e.g. frequency
    - ∗ accounting for syntax may help, e.g. looking for consecutive nouns in a complex noun phrase
- Why not just use quotes?

## Multi-Word Terms from TREC Data



| | |
|---|---|
| 65824 United States | 5778 long time |
| 61327 Article Type | 5776 Armed Forces |
| 33864 Los Angeles | 5636 Santa Ana |
| 17788 North Korea | 5527 Bosnia-Herzegovnia |
| 17308 New York | 5458 words indistinct |
| 15513 San Diego | 5452 international community |
| 15009 Orange County | 5443 vice president |
| 12869 prime minister | 5247 Security Council |
| 12799 first time | 5098 North Korean |
| 12067 Soviet Union | 5023 Long Beach |
| . . . | . . . |

(Data from James Allan, Umass)

## References for Today



- Relevant parts of the course textbook
- Ch. 7, Modern Information Retrieval
- Introduction to Ch. 6, Readings in Information Retrieval
- Managing Gigabytes, 3.1, 3.2, 4.1, 4.3, 4.6