COM6115: Text Processing

Information Retrieval: retrieval models — ranked retrieval methods

Mark Hepple

Department of Computer Science
University of Sheffield

Overview

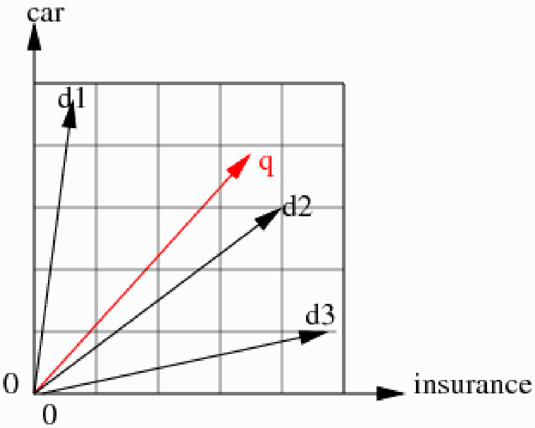
- Definition of the information retrieval problem
- Approaches to document indexing
  - manual approaches
  - automatic approaches
- Automated retrieval models
  - boolean model
  - ranked retrieval methods (e.g. vector space model)
- Term manipulation:
  - stemming, stopwords, term weighting
- Web Search Ranking
- Evaluation

The Vector Space model

- Documents are also represented as "bags of words":
  - ◇ "John is quicker than Mary" = "Mary is quicker than John"
- Documents are points in high-dimensional vector space
  - ◇ each term in index is a dimension → sparse vectors
  - ◇ values are frequencies of terms in documents, or variants of frequency
- Queries are also represented as vectors (for terms that exist in index)
- Approach
  - ◇ Select document(s) with highest document–query similarity
  - ◇ Document–query similarity is a model for relevance (ranking)
  - ◇ With ranking, the number of returned documents is less relevant → users start at the top and stop when satisfied

2 dimensions:

Query: car insurance



The Vector Space Model (contd)

- Approach: compare vector of query against vector of each document
  - to rank documents according to their similarity to the query

	Term_1	Term_2	Term_3	...	Term_n
Doc_1	9	0	1	...	0
Doc_2	0	1	0	...	10
Doc_3	0	1	0	...	2
...	...	...	...	...	...
Doc_N	4	7	0	...	5

Q	0	1	0	...	1

How to measure similarity between vectors?

- Each document and the query are represented as a vector of n values:

$$\vec{d^i} = (d^i_1, d^i_2, \ldots, d^i_n), \qquad \vec{q} = (q_1, q_2, \ldots, q_n)$$

- Many metrics of similarity between 2 vectors, e.g.: Euclidean

$$\sqrt{\sum_{k=1}^{n}(q_k - d_k)^2}$$

- E.g.: Distance between:

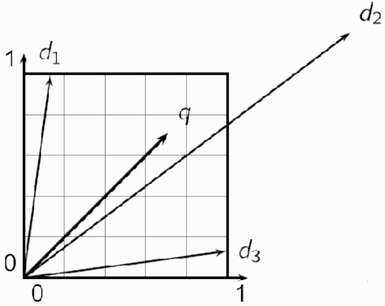
$$Doc_1 \text{ and } Q = \sqrt{(9-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2} = \sqrt{84} = 9.15$$
$$Doc_2 \text{ and } Q = \sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (10-1)^2} = \sqrt{81} = 9$$
$$Doc_3 \text{ and } Q = \sqrt{(0-0)^2 + (1-1)^2 + (0-0)^2 + (2-1)^2} = \sqrt{1} = 1$$

Doc 3 is the closest (shortest distance)

How to measure similarity between vectors? (contd)

Is it a good idea?

- distance is large for vectors of different lengths, even if by only one term (e.g. Doc_2 and Q)
- means frequency of terms given *too much impact*



How to measure similarity between vectors? (contd)

- Better similarity metric, used in *vector-space* model: **cosine** of the angle between two vectors $\vec{x}$ and $\vec{y}$:

$$cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

- It can be interpreted as the normalised correlation coefficient:
  - i.e. it computes how well the x_i and y_i correlate, and then divides by the length of the vectors, to scale for their magnitude
    - ◇ The vector $\vec{x}$ is normalised by dividing its components by its length:

$$|\vec{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}$$

How to measure similarity between vectors? (contd)

- The cosine value ranges from:
  - 1, for vectors pointing in the same direction, to
  - 0, for orthogonal vectors, to
  - -1, for vectors pointing in opposite directions
- Specialising the equation to comparing a query q and document d:

$$sim(\vec{q}, \vec{d}) = cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\sqrt{\sum_{i=1}^{n} d_i^2}}$$

i.e. computes how well occurrences of each term i correlate in query and document, then scales for the magnitude of the overall vectors