Vector Space Model
Cosine Similarity

COMP3009J: Information Retrieval

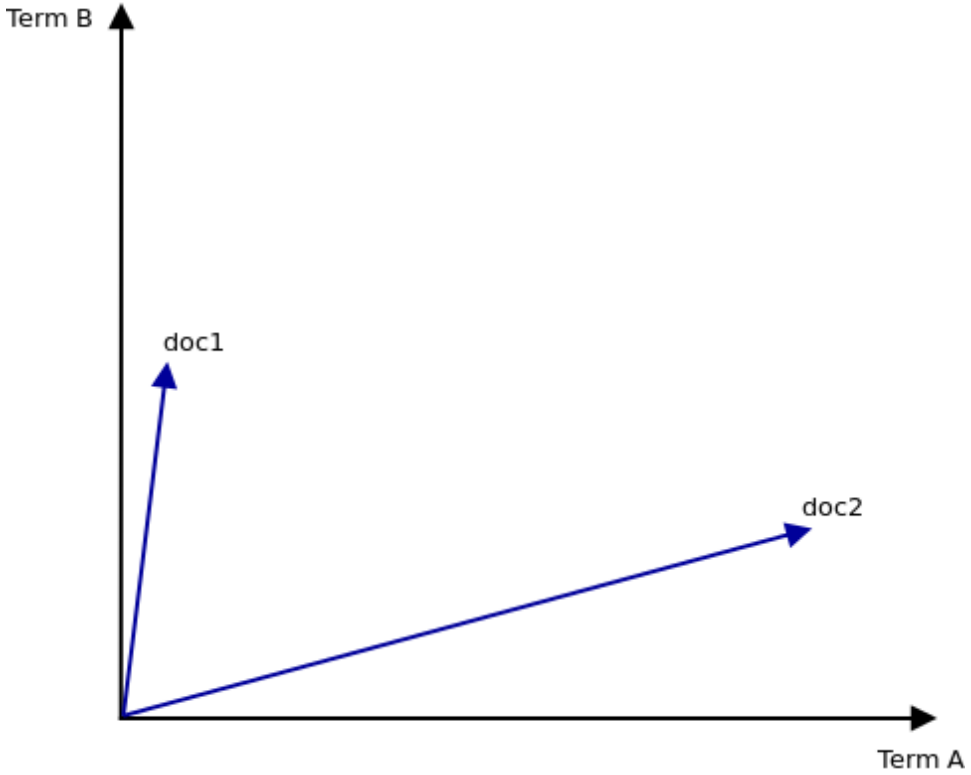
Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science
Beijing Dublin International College

Retrieval

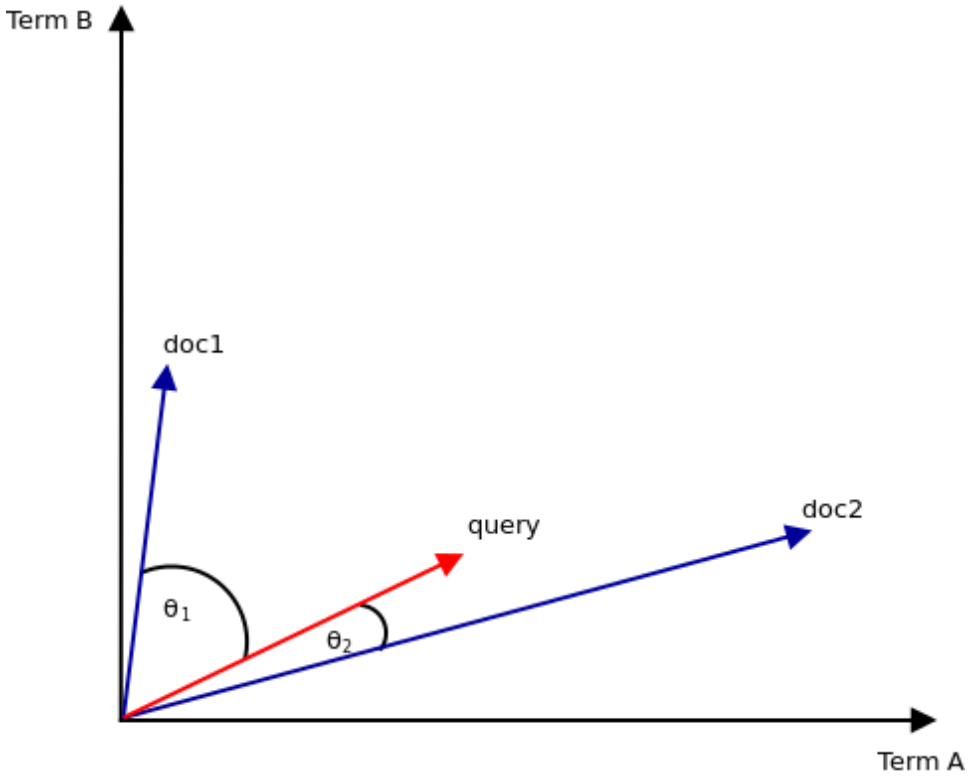
- The theory behind retrieval says that the closer the vectors are in the N-dimensional space, the more similar the items they represent are.
- In retrieval, the query vector is compared to **all of the document vectors** in the index (unless this has been filtered beforehand using postings lists).
- Vector algebra contains an operator known as the **dot product**, defined for two vectors $\vec{a}$ and $\vec{b}$ as:
  - $\vec{a}.\vec{b} = cos.\theta \times |\vec{a}| \times |\vec{b}|$
  - Here, $|\vec{a}|$ and $|\vec{b}|$ are the lengths of vectors $\vec{a}$ and $\vec{b}$ respectively.
  - θ is the angle between the two vectors.

Similarity Example



- Here, we see a VERY simple vector model with only two terms.
- In the document “doc1”, Term B is far more important than Term A.
- The opposite is the case with document “doc2”.
- (This is for illustration only: a realistic example will have many more dimensions).

Similarity Example



- We now introduce a query, **also represented by a vector**.
- The key to the Vector Space Model is the angle between two vectors (θ_1 and θ_2).
- Here, θ_2 is clearly the smaller angle, suggesting that the query is more similar to "doc2".

Retrieval

- Consider the dot product formula again:
  - $\vec{a}.\vec{b} = cos.\theta \times |\vec{a}| \times |\vec{b}|$
- This can be rewritten as:
  - $cos.\theta = \frac{\vec{a}.\vec{b}}{|\vec{a}| \times |\vec{b}|}$
- This means that if we can calculate the dot product of the two vectors, and the lengths of the two vectors, we can easily find the cosine of the angle between them.
- Why would we want to do this?

Cosine Similarity: Why?

- Principle: the smaller the angle between two vectors, the more similar they are.
- We do not use negative weights, therefore: $0° \leq \theta \leq 90°$
  - The angle is $90°$ if the vectors have nothing in common.
  - The angle is $0°$ if the vectors are identical (or near-identical)
- $cos.0° = 1$ and $cos.90° = 0$
  - Therefore, the cosine function gives a value between 0 (for very different documents) and 1 (for identical documents).
  - This is ideal for a measure of **similarity**.

Cosine Similarity Formulae

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \cdot \vec{q}}{|\overrightarrow{d_j}| \times |\vec{q}|}$$

- $\overrightarrow{d_j}$ is a vector representing a document d_j
- $\vec{q}$ is a vector representing the query q
- $\overrightarrow{d_j} \cdot \vec{q}$ is the dot product of $\overrightarrow{d_j}$ and $\vec{q}$
- $|\overrightarrow{d_j}|$ is the length of $\overrightarrow{d_j}$
- $|\vec{q}|$ is the length of $\vec{q}$

Cosine Similarity Formulae

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|}$$

$$= \frac{\sum_{i=1}^{T} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{T} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{T} w_{i,q}^2}}$$

- T is the number of distinct terms in the entire document collection
- $w_{i,j}$ is the weight of term i in document j
- $w_{i,q}$ is the weight of term i in the query q

Vector Definitions

- The top line of this formula is the dot product in *N*-dimensions of $\vec{a}$ and $\vec{b}$, $\vec{a}.\vec{b}$, which is defined as:
  - $\vec{a}.\vec{b} = \sum_{i=1}^{T}(w_{i,a} \times w_{i,b})$
  - where T is the number of dimensions and $w_{i,a}$ is the weight of term i in $\vec{a}$ (i.e. the i^{th} component of the vector $\vec{a}$).
- Or in pseudocode:
  - Begin with a total of 0
  - For each dimension in the vectors:
    - Multiply the value in that dimension of vector $\vec{a}$ by the value in that dimension of vector $\vec{b}$.
    - Add this to the total.
  - The final total is the dot product.

Dot Product: Example

- Example: dot product of (1,1,1,1,0,0) and (1,0,0,0,0,1)
- (1×1)+(1×0)+(1×0)+(1×0)+(0×0)+(0×1) =1

Vector Definitions

- The bottom line of the similarity formula requires us to calculate the length of a vector.
- The length of a vector $\vec{a}$, given by $|\vec{a}|$ is defined as:
- $|\vec{a}| = \sqrt{\sum_{i=1}^{T} w_{i,a}^2}$
- Or in pseudocode:
  - Begin with a total of 0
  - For each dimension in vector $\vec{a}$
    - Take the value in that dimension of vector $\vec{a}$ and square it.
    - Add this to the total.
  - The length of vector $\vec{a}$ is the square root of the total

Vector Length: Example

- Example: length of (1, 1, 1, 1, 0, 0)
- $\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2}$
- $= \sqrt{4} = 2$

Cosine Similarity: Example

- Returning to the documents and the query we saw earlier:
  - D1: 1, 1, 1, 1, 0, 0
  - D2: 1, 1, 0, 0, 1, 1
  - Q: 1, 0, 0, 0, 0, 1
- The similarities between these documents and the query are calculated as:
  - $sim(d_1, q) = \frac{\vec{d_1} \cdot \vec{q}}{|\vec{d_1}| \times |\vec{q}|} = \frac{1}{2 \times \sqrt{2}} = 0.3535$
  - $sim(d_2, q) = \frac{\vec{d_2} \cdot \vec{q}}{|\vec{d_2}| \times |\vec{q}|} = \frac{2}{2 \times \sqrt{2}} = 0.7071$
- Hence document D2 is most similar to the query.
- For a larger corpus, we would return a **ranked list** starting with the document most similar to the query, and descending from there until some cutoff point (we would not return all documents)

Cosine Similarity: Some Questions

- Some questions that will help us when planning to implement this:
  - What effect on the similarity score does a term have if it does not appear in the query?
  - What effect on the similarity score does a term have if it does not appear in the document?
  - What effect on the length of a document vector does a term have if that term is not contained in the document?
- Earlier, we said:
  - … **EVERY DOCUMENT** is mathematically represented by six dimensions (for this tiny corpus), regardless of how many terms it contains.
  - In practice, this leads to very **sparse vectors** (i.e. lots of zeros).
  - "**mathematically**" represented ≠ how we represent using a computer program.