Alternative Models for IR

D. Thenmozhi

Associate Professor

SSNCE

Alternative Models for IR

- Set-Based Model
- Extended Boolean Model
- Fuzzy Set Model
- The Generalized Vector Model
- Latent Semantic Indexing
- Neural Network for IR
- Cluster Model

Cluster Model

- This model is an attempt to reduce the number of matches during retrieval.
- Hypothesis: Closely associated documents tend to be relevant to the same clusters.
- Instead of matching the query with every documents in the collection, it is matched with representatives of the class.

Cluster Model - Algorithm

- Let D={d1,d2,...,dm} – finite set of documents
- Let E=(eij)n,n – similarity matrix (or distance matrix)
- Let T be the threshold
- Any pair of documents di and dj $(i \neq j)$ whose similarity measures exceeds T or whose distance is less than T is grouped to form a cluster
- The remaining document form a single cluster.
- A representative vector of each class is constructed by computing the centroid of the document vector (mean)
- During retrieval, the query is compared with the cluster vectors based on similarity or distance

Steps with Examples

- Let term-document matrix

$$A = \begin{array}{ccc} d1 & d2 & d3 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{array}$$

- The similarity matrix

$$\begin{array}{cccc} & d1 & d2 & d3 \\ d1 & 1.0 & & \\ d2 & 0.9 & 1.0 & \\ d3 & 0.4 & 0.4 & 1.0 \end{array}$$

Steps with Examples

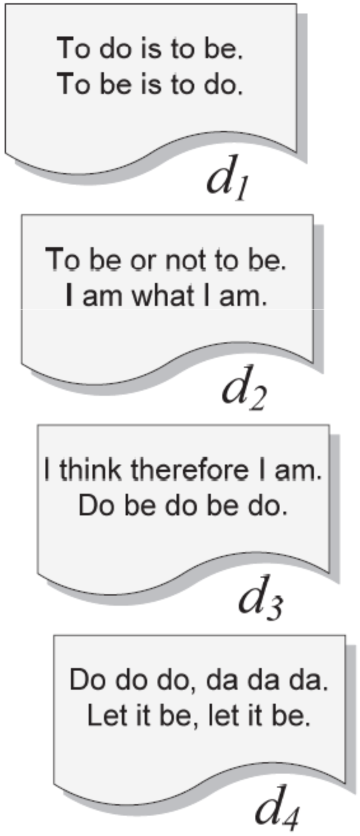
- Threshold T = 0.7
- We get two clusters

  C1 ={d1,d2}   C2={d3}
- The cluster vectors (representatives) for C1 and C2

  r1=(1   0.5   1   0   1)
  r2=(0   0   1   1   0)
- Retrieval is performed by matching the query vector with r1 and r2

Exercise

- Given term-document matrix, find document cluster which is relevant for the query “to do”



To do is to be.
To be is to do.

d_1

To be or not to be.
I am what I am.

d_2

I think therefore I am.
Do be do be do.

d_3

Do do do, da da da.
Let it be, let it be.

d_4

		d_1	d_2	d_3	d_4
1	to	3	2	-	-
2	do	0.830	-	1.073	1.073
3	is	4	-	-	-
4	be	-	-	-	-
5	or	-	2	-	-
6	not	-	2	-	-
7	I	-	2	2	-
8	am	-	2	1	-
9	what	-	2	-	-
10	think	-	-	2	-
11	therefore	-	-	2	-
12	da	-	-	-	5.170
13	let	-	-	-	4
14	it	-	-	-	4

wt(to) – 1, wt(do) – 0.415

Fuzzy Model

- In fuzzy model, the document is represented as a fuzzy set of terms [ti, µ(ti)] where µ is the membership function
- µ assigns a membership degree to each term of the document
- The membership degree expresses the significance (weights) of term to the information contained in the documents

Fuzzy Model - Algorithm

- Each term ti is represented by a fuzzy set fi
- Fuzzy set operators are applied to obtain the desired result
- For single term query q=tq, documents from the fuzzy set fq are retrieved
- For AND query q=tq1 ^ tq2
  - Fuzzy sets fq1 and fq2 are obtained
  - Fuzzy intersection operator is used to obtain the resultant set

    fq1 ^ fq2 = min{(dj,wq1), (dj,wq2)}
  - The documents in this set are returned
- For OR query q=tq1 V tq2
  - Fuzzy sets fq1 and fq2 are obtained
  - Fuzzy intersection operator is used to obtain the resultant set

    fq1 V fq2 = max{(dj,wq1), (dj,wq2)}
  - The documents in this set are returned

Example

- D1={information, retrieval, query}
- D2={retrieval, query, model}
- D3={information, retrieval}

- Vocabulary = {information, model, query, retrieval}

- The fuzzy sets induced by these terms are

  f1={(d1, 1/3), (d2,0), (d3, ½)}
  f2={(d1, 0), (d2,1/3), (d3, 0)}
  f3={(d1, 1/3), (d2,1/3), (d3, 0)}
  f4={(d1, 1/3), (d2,1/3), (d3, ½)}

Try : q= t1 v t4

- Query q= t2 ^ t4 (model retrieval)
  - Fuzzy sets f2 and f4 are considered
  - Min(f2(d1), f4(d1)), Min(f2(d2), f4(d2)), Min(f2(d3), f4(d3))
  - = {(d1, 0), (d2,1/3), (d3, 0)}
  - d2 is returned