Search Model

- Extended Boolean Model

Contents

Definitions

- is based on Boolean Model and added Vector Space Model.

- The weak point of Boolean Model is not give Weight.

- Query used Boolean Model and Results used Vector Space Model.

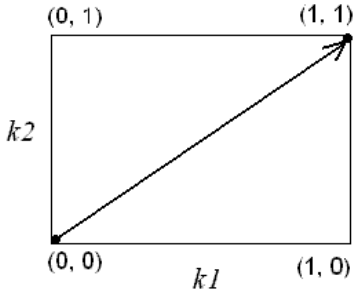
- find similarity how calculate Euclid distance between Document Term Weight

Boolean OR

- Assume Query has K1, K2

- So, We can find two extreme points in Term Space.
  - Document include two terms(k1, k2), and complete similarity → (1,1)
  - Document hasn't two terms(k1, k2), and similarity is 0 → (0,0)

$$dmax = \sqrt{(1-0)^2 + (1-0)^2} = \sqrt{2} = 1.41$$



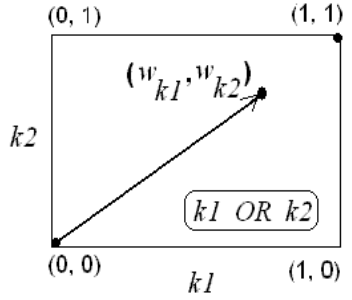
- The max distance (dmax) of two point is 1.41
- Therefore, the similarity of documents is between 0 and 1.41

- We knows that Euclid distance of OR Query (W_{k1}, W_{k2}) is less than 1.41

- $W_{k1} = TF_{(k1, d)} X (IDF_{k1} / (max_d X IDF_d))$

  - $TF_{(k1, d)}$ is TF of k1 in document 'd'
  - max_d is TF of term has max TF in document 'd'
  - IDF_d is inverse DF(IDF) of term has max TF in document 'd'
  - max_d X IDF_d is TF*IDF of term has max TF in document 'd'

$$d_{OR} = \sqrt{(w_{k1} - 0)^2 + (w_{k2} - 0)^2} = \sqrt{w_{k1}^2 + w_{k2}^2}$$

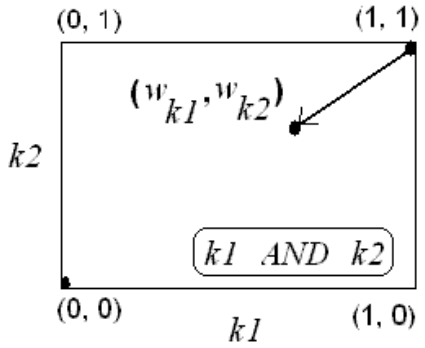


Boolean AND

- Assume Query has K1, K2

- We calculate Euclid distance between dmax(1,1) and (W_{k1}, W_{k2})

$$d_{AND} = \sqrt{2} - \sqrt{(1 - w_{k1})^2 + (1 - w_{k2})^2}$$



Normalized : Similarity Score

- We need normalized for similarity
- So, divide into term's number

$$Sim(Q_{k1\ OR\ k2}, D) = \sqrt{\frac{w_{k1}^2 + w_{k2}^2}{2}}$$

$$Sim(Q_{k1\ AND\ k2}, D) = 1 - \sqrt{\frac{(1 - w_{k1})^2 + (1 - w_{k2})^2}{2}}$$

P-Norm

- Add independent p-parameter(p-norm) into Normalized Similarity

$$Sim(Q_{k1\ OR\ k2}, D) = \left(\frac{w_{k1}^{p} + w_{k2}^{p} + \cdots + w_{km}^{p}}{m} \right)^{1/p}$$

$$Sim(Q_{k1\ AND\ k2}, D) = 1 - \left(\frac{(1 - w_{k1})^{p} + (1 - w_{k2})^{p} + \cdots + (1 - w_{km})^{p}}{m} \right)^{1/p}$$

- if p is 1, then get effectiveness of **Vector Space Model**
  - $Sim(Q_{OR}, D) = Sim(Q_{AND}, D) = (W_{k1} + W_{k2} + \ldots + W_{m}) / m$
- if p is ∞, then get effectiveness of **Boolean Model**

AND OR 조합

- p-norm has between 1 and ∞

$$Sim(Q_{k1\ AND\ K2\ OR\ k3}, D) = \left(\frac{\left(1 - \left(\frac{(1 - w_{k1})^p + (1 - w_{k2})^p}{2} \right)^{1/p}\right)^p + w_{k3}^p}{2} \right)^{1/p}$$

- Upper Similarity's formula can apply using recursive method unrelated Operator's number

Features

- We can change p-norm 1 to ∞, so can see Vector Ranking Result to Boolean Ranking Result

- We can use different p-norm

  - for example

$$Sim(Q_{k1\ AND\ K2\ OR\ k3}, D) = \left(\frac{\left(1 - \left(\frac{(1 - w_{k1})^p + (1 - w_{k2})^p}{2}\right)^{1/p}\right)^p + w_{k3}^p}{2} \right)^{1/p}$$

P = 2 P =∞

so, can get Vector Space Model's effectiveness at specified Operator (p=2)
also, get Boolean Model's effectiveness at specified Operator (p = ∞)