http://gogamza.cafe24.com

백터 공간 모델(vector space model)에 대한 설명을 해보고자 한다.

불린 모델이 가지는 이진가중치에 대한 대안으로 나온 검색 모델링 방법으로 이 부분의 색인어 가중치는 나름대로의 알고리즘에 따라서 도출되어야 하지만 (일반적으로 tf/idf 방법을 많이 쓴다.) 이곳에는 1, 0의 값으로만 가중치를 표현하도록 하겠다.

여기서 검색의 대상이 되는 문서를 $D_1, D_2, D_3, D_4, \ldots, D_n$ 이라고 하고, 이와 같은 문서 집합 전체에 걸쳐 전부 m개의 색인어 $w_1, w_2, w_3, \ldots, w_m$ 이 있다고 한다. 이때, 문서는 다음과 같은 벡터로 표현된다. 이것을 문서벡터(Document Vector)라고 부른다.

$$d_j = \begin{bmatrix} d_{1j} \\ d_{2j} \\ \vdots \\ d_{mj} \end{bmatrix}$$

여기서 d_{ij} 는 색인어 w_i 의 문서 d_j 에서의 가중치 이다.
따라서 문서 집합 전체는 다음과 같은 $m*n$ 행렬 D에 의해 표현할 수 있다.
그럼 색인어 – 문서행렬(term-document matrix)을 만들어 보자. 여기서 각 행이 색인어(term vector라고 한다.

$$D = \begin{bmatrix} d_1 & d_2 & \cdots & d_n \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{bmatrix}$$

검색 쿼리도 문서와 동일하게 벡터로 표시하 주자.
검색질문에 포함되는 색인어 w_i 의 가중치를 q_i 라고 하면 검색질문 벡터 q는 다음과 같이 표현된다.

$$q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix}$$

실제의 문서검색에서는 벡터 q와 d_j 사이의 유사도를 계산하여 비슷한 문서를 찾아 내는데 여기서 자주 사용되는 방법이 벡터의 내적이다.(두 벡터가 이루는 각도)

● 코사인 척도

http://gogamza.cafe24.com

$$\cos(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \|q\|} = \frac{\sum_{i=1}^{m} d_{ij} q_i}{\sqrt{\sum_{i=1}^{m} d_{ij}^2} \sqrt{\sum_{i=1}^{m} q_i^2}}$$

책에 아주 좋은 예제가 있어서 첨부한다.
이 예는 유전자정보 처리와 관계가 있는 문헌의 검색을 가정하고 있으며 8개의 단어를 색인어로 이용하고 있다. 색인어에 밑줄을 그어 나타내고 있다.
그리고 색인어의 가중치로는 색인어 빈도를 이용하고 있다.
검색질문으로써 “Genes and Genomes”가 주어졌을때 검색 예를 나타낸다.

색 인 어

w1 : Bioinformatics
w2 : Biology
w3 : Chemistry
w4 : Enzymes
w5 : Evolution
w6 : Gens
w7 : Genome(s)
w8 : Proteins

문 서
D1 Bioinformatics : A Practical Guide to the Analysis of Genes and Proteins

D2 Proteins. Enzymes. Genes : The Interplay of Chemistry and Biology

D3 Adaptive Evolution of Genes and Genomes
D4 Advanced in Genome Biology : Genes and Genomes
D5 Bioinformatics and Genome Research
D6 Data Analysis in Molecular Biology and Evolution

색인어, 문서행렬 (색인어의 가중치 색인어 빈도) : $D = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$

http://gogamza.cafe24.com

검색쿼리 : Genes and Genomes

검색질문 백터 : $q = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$

유사도 계산

$\cos(d_1, q) = 0.408$
$\cos(d_2, q) = 0.316$
$\cos(d_3, q) = 0.816$
$\cos(d_4, q) = 0.866$
$\cos(d_5, q) = 0.050$
$\cos(d_6, q) = 0.000$

검색결과

검색순위	*문서*	*유사도*
1	D_4	0.866
2	D_3	0.816
3	D_5	0.5
4	D_1	0.408
5	D_2	0.326

가중치에 관한 부분은 널리 쓰이는 역문헌 빈도수(inverted document frequency)를 쓰면 될 것이다.
간단하게 역문헌 빈도수를 설명하자면 한 문서내에서 많이 나온 색인어에 대한 가중치를 올리는 대신에 이 단어가 여러 문서에 걸쳐 나올경우에 그에 대한 패널티를 먹이는 가중치 부여 공식이다.

예를들어 신문 기사가 있을경우에 그곳에 마지막에 나오는 'OOO 기자'에서 '기자'라는 색인어는 모든 신문기사에 존재하기 때문에 패널티를 먹여서 순위를 낮출 필요가 있다는 것이다.