문서 유사도 계산

국민대학교 컴퓨터공학부

강 승 식

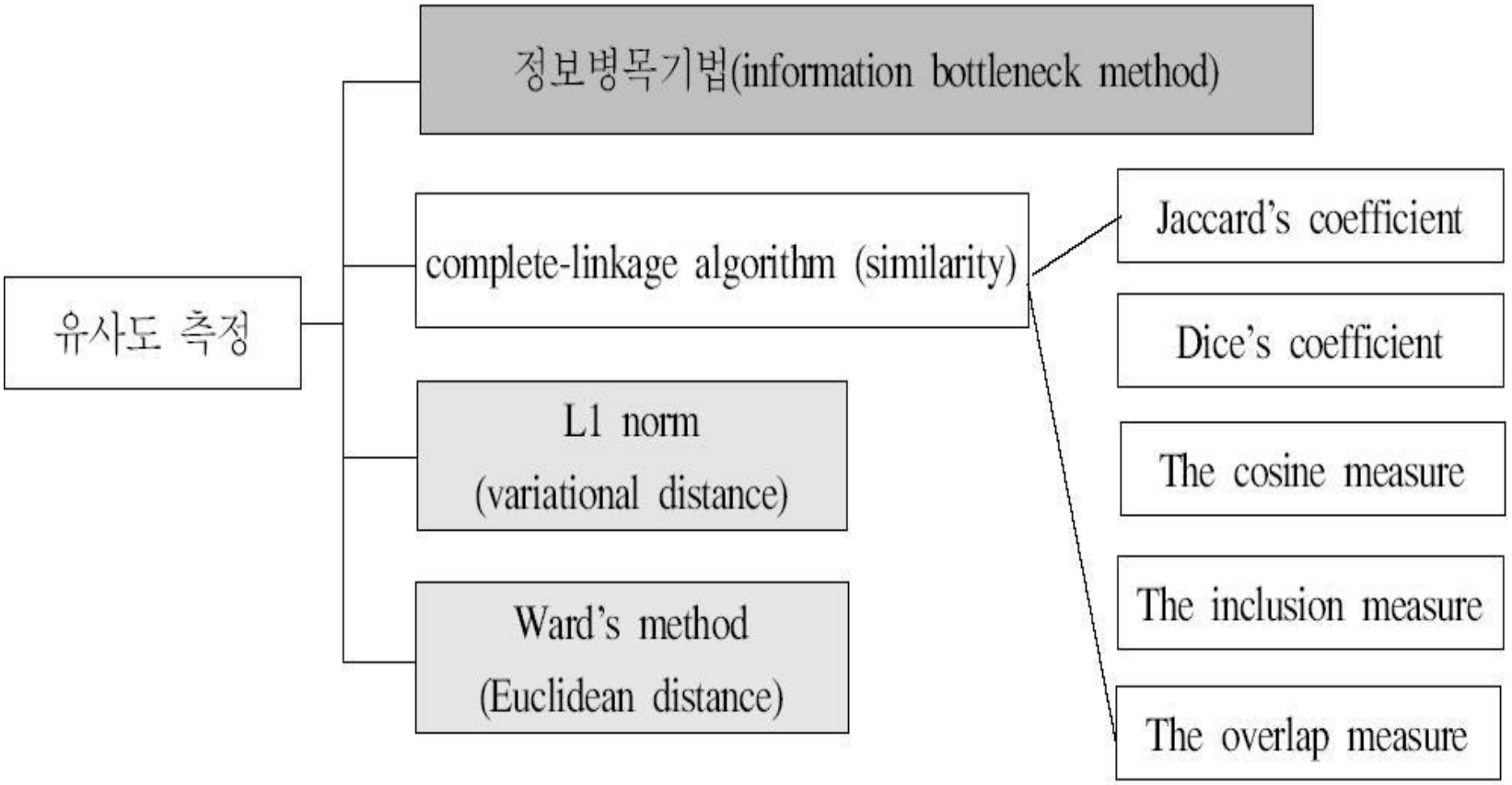
문서-문서간의 관련성 평가 방법

- 문서의 인덱싱 과정이 끝나면 이를 바탕으로 문서-문서간의 관련성을 평가할 수 있는 척도를 설정한다. 관련성에 대한 평가는 관련 정보의 추출, 여과, 분류 등의 기능을 수행하기 위한 중요한 기본 자료가 된다

- 일반적인 분류 시스템에서의 접근 방법에서는 N차원 벡터 공간에 존재하는 문서 벡터들 사이의 거리를 사용하여 문서간의 관련성을 평가한다

- 질의 또한, 이를 구성하는 어휘를 통해 벡터 형태로 표현될 수 있다.

클러스터링 기법에서의 유사도 측정 방법



클러스터링 기법에서의 유사도 측정 방법 분류

문서 데이터의 형태 : 문서-색인어 행렬

	T_1	T_2	·····	T_m
D_1	t_{11}	t_{12}	·····	t_{1m}
D_2	t_{21}	t_{22}	·····	t_{2m}
⋮	⋮	⋮		⋮
D_n	t_{n1}	t_{n2}	·····	t_{nm}

n: 문서의 개수
m: 단어의 개수
$D_i(i=1,\cdots,n)$: i 번째 문서
$T_j(j=1,\cdots,m)$: j 번째 단어
t_{ij}: i 번째 문서에 나타난 j 번째 단어의 빈도수

문서 클러스터링에 사용되는 데이터의 형태 : 문서-색인어 행렬

Example

- M =6, n =5
- T = {computer, retrieval, archiving, hypertext, hypermedia, indexing}
- D_i = {computer, retrieval, archiving, hypertext, hypermedia}
  = {1, 1, 1, 1, 1, 0}
- q1 = {archiving, hypermedia}
  = {0, 0, 1, 0, 1, 0}
- q2 = {retrieval, indexing}
  = {0, 1, 0, 0, 0, 1}

Dice's coefficient

- $SIM_{Dice}(D_i, D_{i'}) =$

$$\frac{2\left[\sum_{j=1}^{m}\left(T_{ij}\cdot T_{i'j}\right)\right]}{\sum_{j=1}^{m}T_{ij}+\sum_{j=1}^{m}T_{i'j}}$$

- $SIM_{Dice}(q_1, D_i) = 2/3.5 = 0.5714$
- $SIM_{Dice}(q_2, D_i) = 1/3.5 = 0.286$

Jaccard's coefficient

- $SIM_{Jacc}(D_i, D_{i'}) =$

$$\frac{\sum_{j=1}^{m}\left(T_{ij}\cdot T_{i'j}\right)}{\sum_{j=1}^{m}T_{ij}+\sum_{j=1}^{m}T_{i'j}-\sum_{j=1}^{m}\left(T_{ij}\cdot T_{i'j}\right)}$$

- $SIM_{Jacc}(q_1, D_i) = 2/5 = 0.4$
- $SIM_{Jacc}(q_2, D_i) = 1/6 = 0.167$

Cosine measure

- $SIM_{cos}(D_i, D_{i'}) =$

$$\frac{\sum_{j=1}^{m} \left(T_{ij} \cdot T_{i'j}\right)}{\left[\sum_{j=1}^{m} \left(T_{ij}\right)^2 \cdot \sum_{j=1}^{m} \left(T_{i'j}\right)^2\right]^{1/2}}$$

- $SIM_{cos}(q_1, D) = 2/\sqrt{10} = 0.632$
- $SIM_{cos}(q_2, D) = 1/\sqrt{10} = 0.316$

Inclusion measure

- $SIM_{incl}(D_i, D_{i'}) =$

$$\frac{\sum_{j=1}^{m} (T_{ij} \cdot T_{i'j})}{\sum_{j=1}^{m} T_{ij}}$$

- $SIM_{incl}(q1, D_i) = 2/2 = 1.0$
- $SIM_{incl}(D_i, q1) = 2/5 = 0.4$
- $SIM_{incl}(q2, D_i) = 1/2 = 0.5$
- $SIM_{incl}(D_i, q2) = 1/5 = 0.2$

Overlap coefficient

- $SIM_{OVL}(D_i, D_{i'}) =$

$$\frac{\sum_{j=1}^{m}\left(Tij \cdot Ti'j\right)}{\min\left(\sum_{j=1}^{m} Tij, \sum_{j=1}^{m} Ti'j\right)}$$

- $SIM_{OVL}(q_1, D_i) = \frac{2}{\min(5,2)} = 1$

- $SIM_{OVL}(q_2, D_i) = \frac{1}{\min(5,2)} = 0.5$

여러 가지 관련성 척도

- ***Dice's Coefficient***

  $2|X \cap Y| / |X| + |Y|$

- ***Jaccard's Coefficient***

  $|X \cap Y| / |X \cup Y|$

- ***Cosine Coefficient***

  $|X \cap Y| / |X|^{1/2} * |Y|^{1/2}$

- ***Overlap Coefficient***

  $|X \cap Y| / min(|X|, |Y|)$