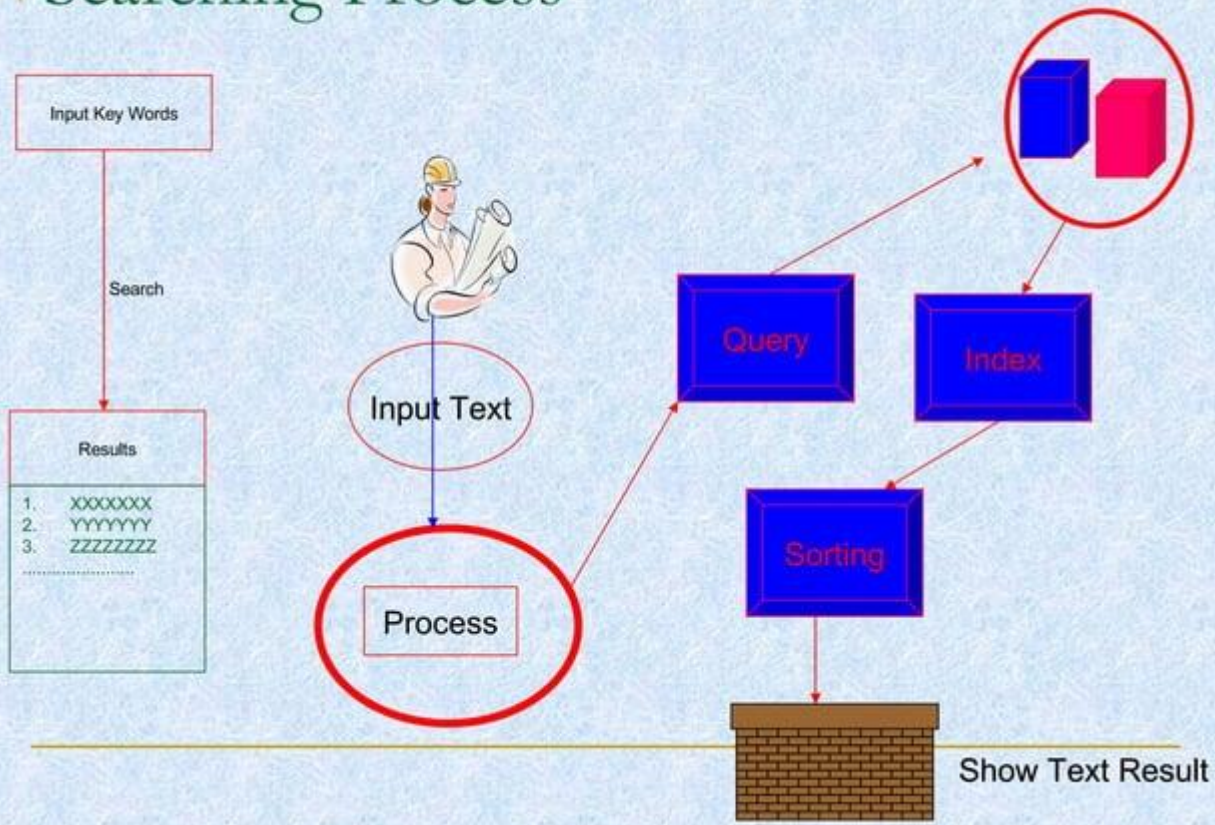
SIMILARITY MEASUREMENTS

Searching Process



Similarity Measures

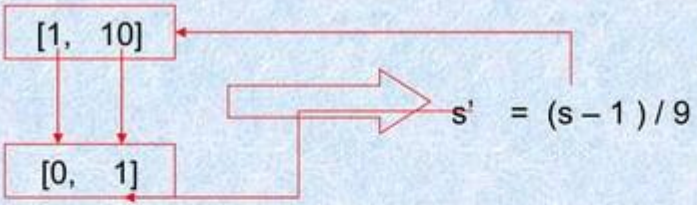
- A similarity measure can represent the similarity between two documents, two queries, or one document and one query
- It is possible to rank the retrieved documents in the order of presumed importance
- A similarity measure is a function which computes the degree of similarity between a pair of text objects
- There are a large number of similarity measures proposed in the literature, because the best similarity measure doesn't exist (yet!)

Classic Similarity Measures

- All similarity measures should map to the range [-1,1] or [0,1],
- 0 or -1 shows minimum similarity. (incompatible similarity)
- 1 shows maximum similarity. (absolute similarity)

Conversion

- For example
- 1 shows incompatible similarity, 10 shows absolute similarity.



Linear
Non-linear

Generally, we may use:

$s' = (s - min\_s) / (max\_s - min\_s)$

Vector-Space Model-VSM

- 1960s Salton etc provided VSM, which has been successfully applied on SMART (a text searching system).

  (System for the Mechanical Analysis and Retrieval of Text)

Example

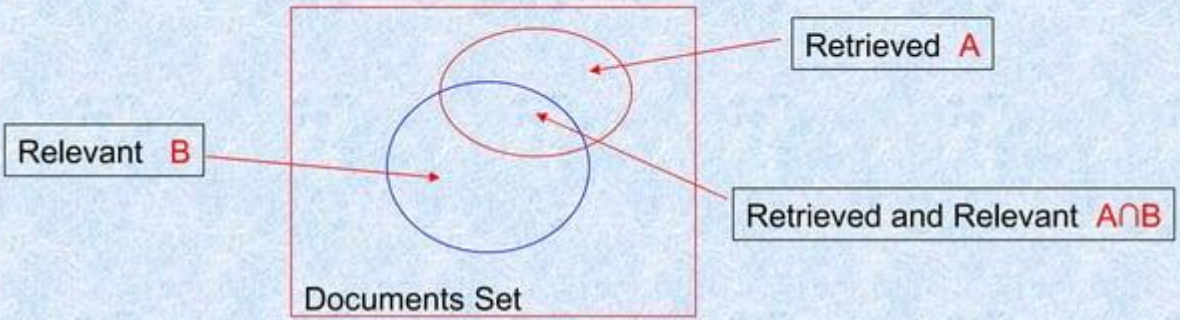
- D is a set, which contains m Web documents;
  $D=\{d_1, d_2,\ldots d_i\ldots d_m\}$ i=1,2…m

- There are n words among m Web documents. $di=\{w_{i1},w_{i2},\ldots w_{ij},\ldots w_{in}\}$ i=1,2…m , j=1,2,….n

- $Q=\{q_1,q_2,\ldots q_n\}$ i=1,2,…n

If similarity(q,d_i) > similarity(q,d_j)
We may get the result d_i is more relevant than d_j

Simple Measure Technology



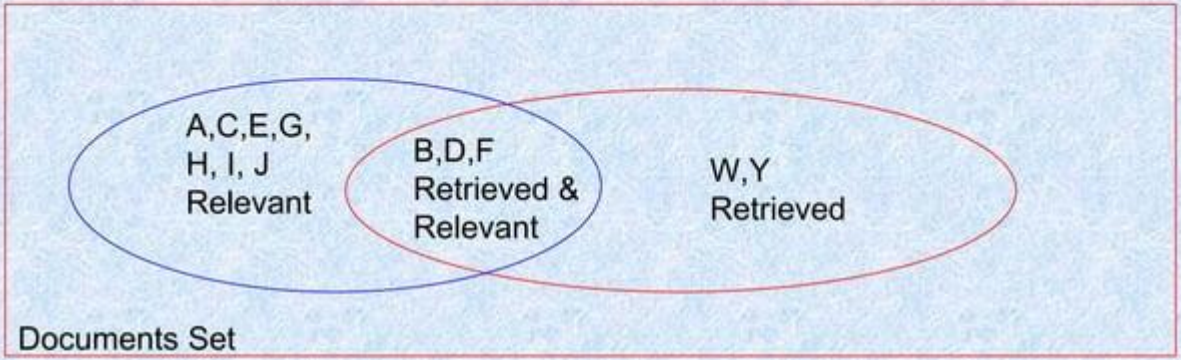
Precision = Returned Relevant Documents / Total Returned Documents

$P(A,B) = |A \cap B| / |A|$

Recall = Returned Relevant Documents / Total Relevant Documents

$R(A,B) = |A \cap B| / |B|$

Example--Simple Measure Technology



|A| = {retrieved} = {B, D, F,W,Y} = 5

|B| = {relevant} ={A,B,C,D,E,F,G,H,I,J} = 10

|A∩B| = {relevant} ∩ {retrieved} ={B,D,F} = 3

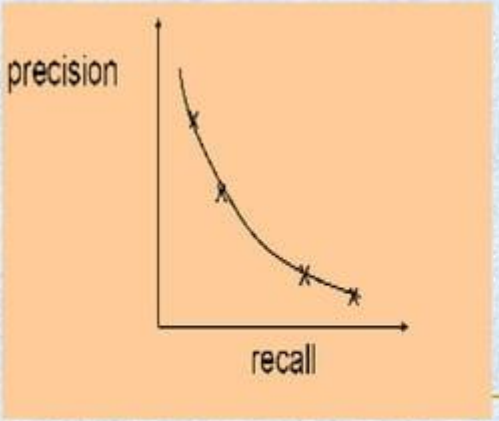
P = precision = 3/5 = 60%

R = recall = 3/10 = 30%

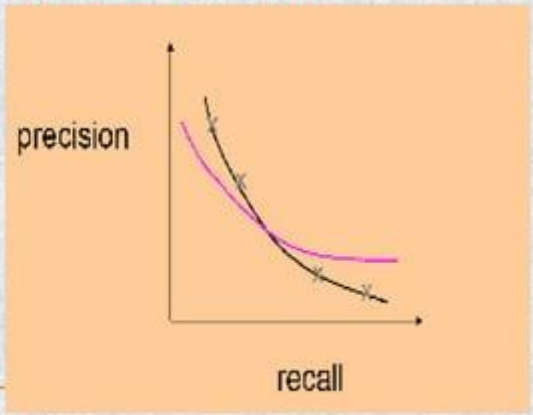
Precision-Recall Graph-Curves

- There is a tradeoff between Precision and Recall
- So measure Precision at different levels of Recall

Difficult to determine which of these two hypothetical results is better



One Query



Two Queries

Similarity measures based on VSM

- Dice coefficient
- Overlap Coefficient
- Jaccard
- Cosine
- Asymmetric
- Dissimilarity
- Other measures

Dice Coefficient-Cont'

- Definition of Harmonic Mean:
- To $X_1, X_2, \ldots, X_n$,their harmonic mean E equals n divided by$(1/x_1+1/x_2+\ldots+1/X_n)$, that is

$$E = \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\ldots+\frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

- To Harmonic Mean (E) of Precision (P) and Recall (R)

$$E = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2}{\frac{|A\cap B|}{|A|}+\frac{|A\cap B|}{|B|}} = \frac{2\times|A\cap B|}{|A|+|B|}$$

Dice Coefficient-Cont'

- Denotation of Dice Coefficient:

$$sim(q,d_j) = D(A,B) = \frac{|A \cap B|}{\alpha|A| + (1-\alpha)|B|}$$

$$\cong \frac{\sum_{k=1}^{n} w_{kq} w_{kj}}{\alpha \sum_{k=1}^{n} w_{kq}^2 + (1-\alpha) \sum_{k=1}^{n} w_{kj}^2} (\alpha \in [0,1])$$

α>0.5 : precision is more important

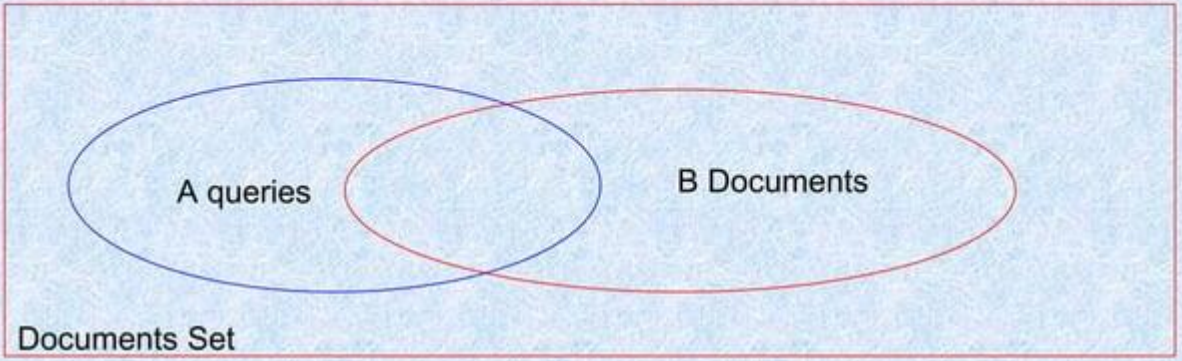
α<0.5 : recall is more important

Usually α=0.5

$$if \alpha = \frac{1}{2} then D(A,B) = \frac{|A \cap B|}{\alpha|A| + (1-\alpha)|B|}$$

$$= \frac{2|A \cap B|}{|A| + |B|} = E$$

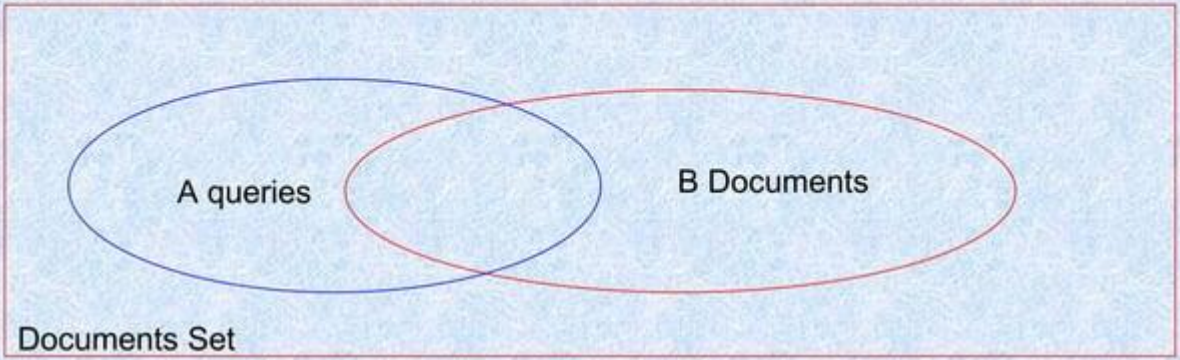
Overlap Coefficient



$$sim(q, d_j) = O(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$$

$$\cong \frac{\sum_{k=1}^{n} w_{kq} w_{kj}}{\min(\sum_{k=1}^{n} w_{kq}^2, \sum_{k=1}^{n} w_{kj}^2)}$$

Jaccard Coefficient-Cont'



$$sim(q, d_j) = J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

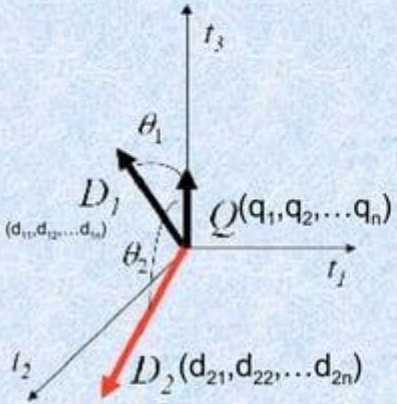
$$\cong \frac{\sum_{k=1}^{n} w_{kq} w_{kj}}{\sum_{k=1}^{n} w_{kq}^2 + \sum_{k=1}^{n} w_{kj}^2 - \sum_{k=1}^{n} w_{kq} w_{kj}}$$

Example- Jaccard Coefficient

$$sim(q, d_j) \cong \frac{\sum_{k=1}^{n} w_{kq} w_{kj}}{\sum_{k=1}^{n} w_{kq}^2 + \sum_{k=1}^{n} w_{kj}^2 - \sum_{k=1}^{n} w_{kq} w_{kj}}$$

- *D1 = 2T1 + 3T2 + 5T3,* *(2,3,5)*
- *D2 = 3T1 + 7T2 + T3 ,* *(3,7,1)*
- *Q = 0T1 + 0T2 + 2T3,* *(0,0,2)*
- J(*D1* , *Q*) = 10 / (38+4-10) = 10/32 = 0.31
- J(*D2* , *Q*) = 2 / (59+4-2) = 2/61 = 0.04

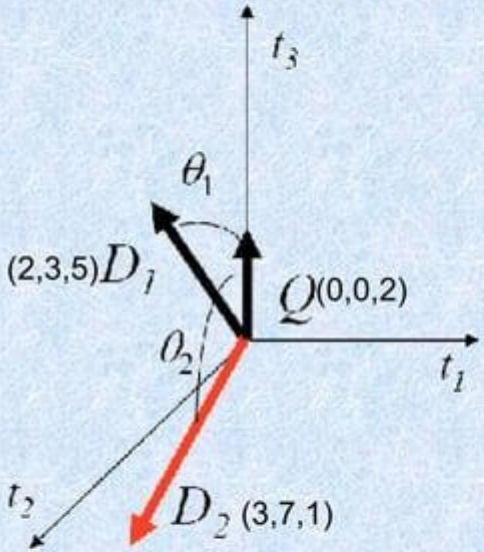
Cosine Coefficient-Cont'



$$sim(q,d_j) = C(A,B) = \sqrt{PR} = \frac{|A \cap B|}{\sqrt{|A||B|}}$$

$$\cong \frac{\vec{q} \cdot \vec{d}_j}{\left|\vec{q}\right| \times \left|\vec{d}_j\right|} = \frac{\sum_{k=1}^{n} w_{kq} w_{kj}}{\sqrt{\sum_{k=1}^{n} w_{kq}^2 \sum_{k=1}^{n} w_{kj}^2}}$$

Example-Cosine Coefficient



- $Q = 0T1 + 0T2 + 2T3$
- $D1 = 2T1 + 3T2 + 5T3$
- $D2 = 3T1 + 7T2 + T3$

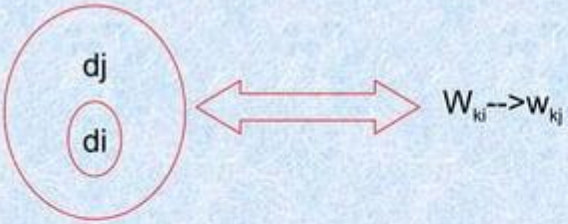
$$C(D1, Q) = \frac{0\times2+0\times3+2\times5}{\sqrt{(2^2+3^2+5^2)\times(0^2+0^2+2^2)}}$$

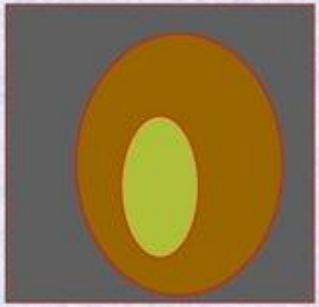
$$= 10/\sqrt{(38*4)} = 0.81$$

$$C(D2, Q) = 2/\sqrt{(59*4)} = 0.13$$

Asymmetric

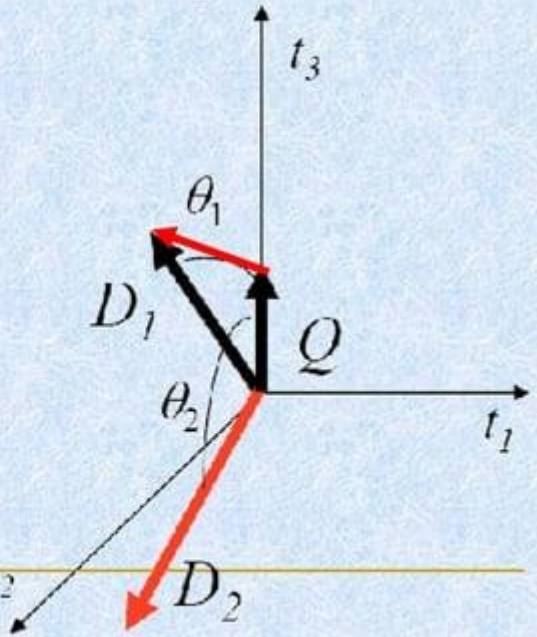
$$sim(q,d_j) = A(q,d_j) = \frac{\sum_{k=1}^{n} \min(w_{kq}, w_{kj})}{\sum_{k=1}^{n} w_{kq}}$$





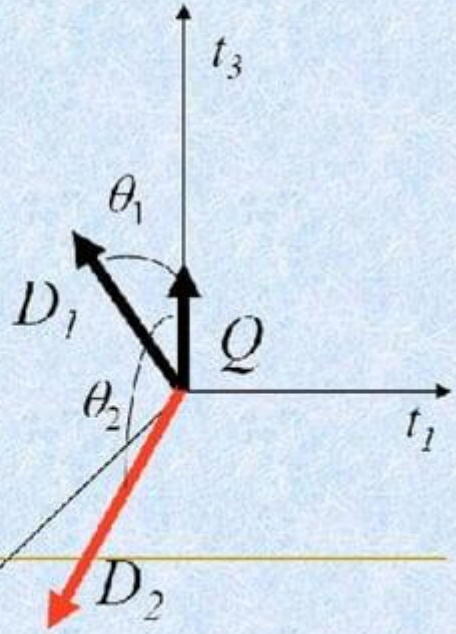
Euclidean distance

$$dis(q, d_j) = d_E(q, d_j) = \sqrt{\sum_{k=1}^{n} (w_{kq} - w_{kj})^2}$$



Manhattan block distance

$$dis(q, d_j) = d_M(q, d_j) = \sum_{k=1}^{n} \left| w_{kq} - w_{kj} \right|$$



Other Measures

- We may use priori/context knowledge
- For example:
- $Sim(q,d_j) = \alpha$[content identifier similarity]+
  β [objective term similarity]+
  γ [citation similarity]

Comparison

Inner Product: $\sum_{k=1}^{t}(x_k \bullet y_k)$ $\quad |x \cap y|$

Cosine: $\dfrac{\sum_{k=1}^{t}(x_k \bullet y_k)}{\sqrt{\sum_{k=1}^{t} x_k^2 \bullet \sum_{k=1}^{t} y_k^2}}$ $\quad \dfrac{|x \cap y|}{\sqrt{|x|} \cdot \sqrt{|y|}}$

Jaccard : $\dfrac{\sum_{k=1}^{t}(x_k \bullet y_k)}{\sum_{k=1}^{t} x_k^2 + \sum_{k=1}^{t} y_k^2 - \sum_{k=1}^{t}(x_k \bullet y_k)}$ $\quad \dfrac{|x \cap y|}{|x \cup y|}$

$\dfrac{|x \cap y|}{|x| + |y| - |x \cap y|}$

x and y here are vectors

x and y here are sets of keywords

Comparison

$\lvert A \cap B \rvert$	Simple matching
$D = 2 * \frac{\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$	Dice's Coefficient
$J = \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	Jaccard's Coefficient
$C = \frac{A \bullet B}{\lvert A \rvert^{\frac{1}{2}} \times \lvert B \rvert^{\frac{1}{2}}}$	Cosine Coefficient
$O = \frac{\lvert A \cap B \rvert}{\min(\lvert A \rvert, \lvert B \rvert)}$	Overlap Coefficient

$|A| \geq |A \cap B|$
$|B| \geq |A \cap B|$

$|A| + |B| - |A \cap B| \geq (|A| + |B|)/2$

$(|A| + |B|)/2 \geq \sqrt{(|A| * |B|)}$

$\sqrt{(|A| * |B|)} \geq \min(|A|, |B|)$

$O \geq C \geq D \geq J$

Example-Documents-Term-Query-Cont'

D1:A search Engine for 3D Models

D2:Design and Implementation of a string database query language

D3:Ranking of documents by measures considering conceptual dependence between terms

D4 Exploiting hierarchical domain structure to compute similarity

D5:an approach for measuring semantic similarity between words using multiple information sources

D6:determinging semantic similarity among entity classes from different ontologies

D7:strong similarity measures for ordered sets of documents in information retrieval

T1:search(ing)
T2:Engine(s)
T3:Models
T4:database
T5:query
T6:language
T7:documents
T8:measur(es,ing)
T9:conceptual
T10:dependence
T11: domain
T12:structure
T13:similarity
T14:semantic
T15: ontologies
T16:information
T17: retrieval

Query:

Semantic similarity measures used by search engines and other information searching mechanisms

Example-Term-Document Matrix-Cont'

Matrix[q][A]

||

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	13	T14	T15	T16	T17
D1	1	1	1														
D2				1	1	1											
D3							1	1	1	1							
D4											1	1	1				
D5								1					1	1		1	
D6													1	1	1		
D7							1	1					1			1	1
Q	2	1						1					1	1		1	

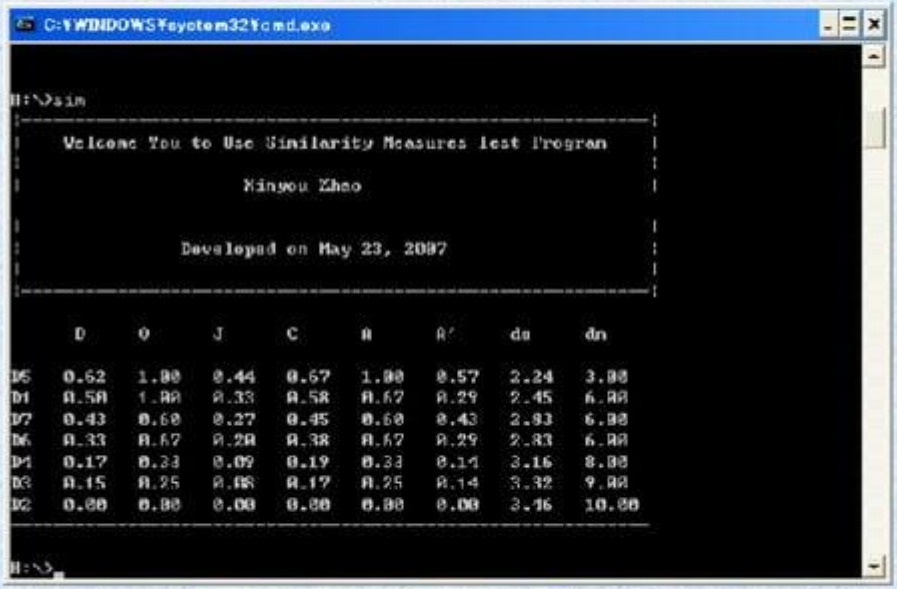
Dice coefficient

$$D(D1,q) \cong \frac{\sum_{k=1}^{n} w_{kq} w_{k1}}{\alpha \sum_{k=1}^{n} w_{kq}^2 + (1-\alpha) \sum_{k=1}^{n} w_{k1}^2} (\alpha = \frac{1}{2}) = \frac{2 \sum_{k=1}^{n} w_{kq} w_{k1}}{\sum_{k=1}^{n} w_{kq}^2 + \sum_{k=1}^{n} w_{k1}^2}$$

$$= \frac{2*(2*1+1*1+0*1+0*0+\ldots+0*1+0*0)}{(2*2+1*1+\ldots+0*0)+(1*1+1*1+\ldots+0*0)}$$

$$= \frac{2*(2+1)}{9+3} = \frac{6}{12} = 0.5$$

Final Results



O≥C≥D≥J

Current Applications

- Multi-Dimensional Modeling
- Hierarchical Clustering
- Bioinformatics