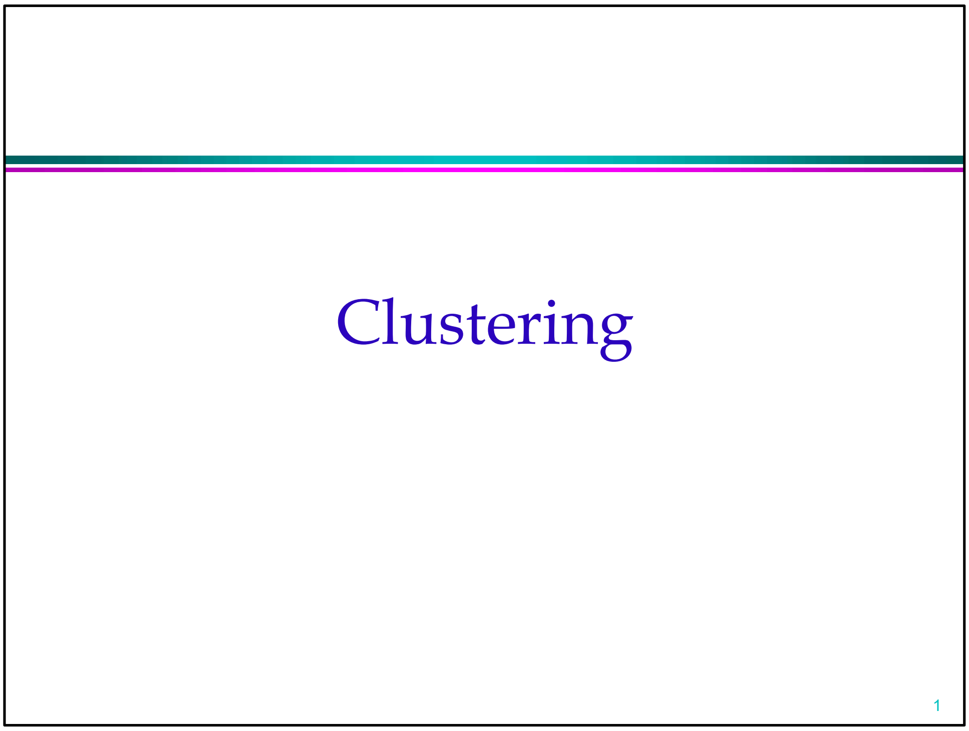
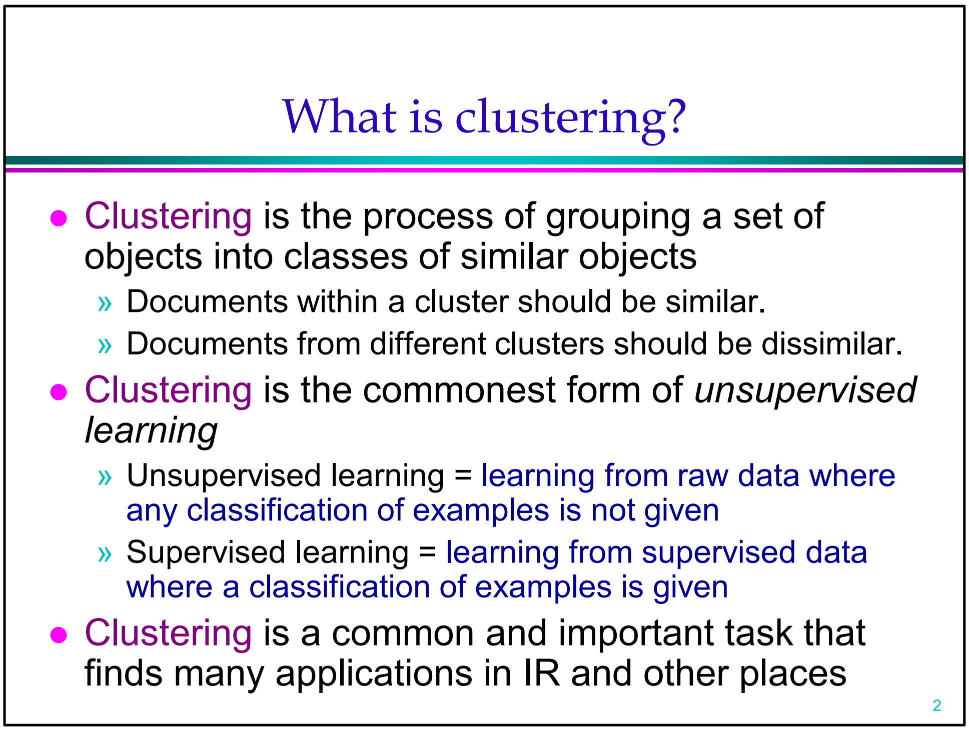


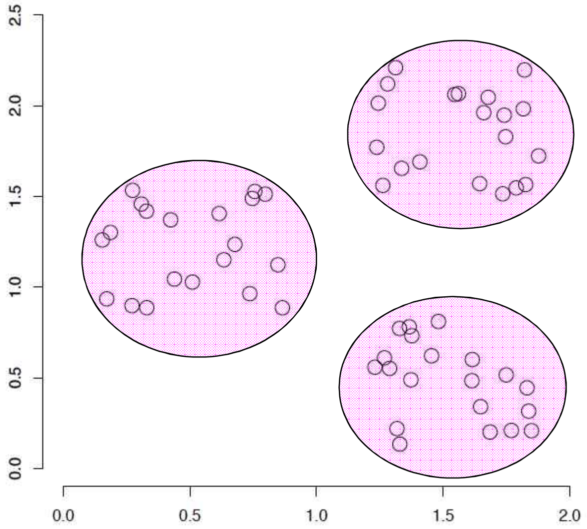
Clustering



What is clustering?

- Clustering is the process of grouping a set of objects into classes of similar objects
  - » Documents within a cluster should be similar.
  - » Documents from different clusters should be dissimilar.
- Clustering is the commonest form of *unsupervised learning*
  - » Unsupervised learning = learning from raw data where any classification of examples is not given
  - » Supervised learning = learning from supervised data where a classification of examples is given
- Clustering is a common and important task that finds many applications in IR and other places

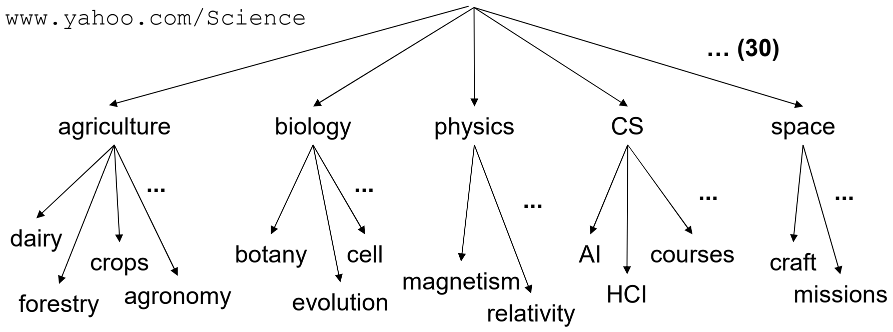
A data set with clear cluster structure



- How would you design an algorithm for finding these three clusters in this case?
- We are going to learn such clustering algorithms in this lecture!

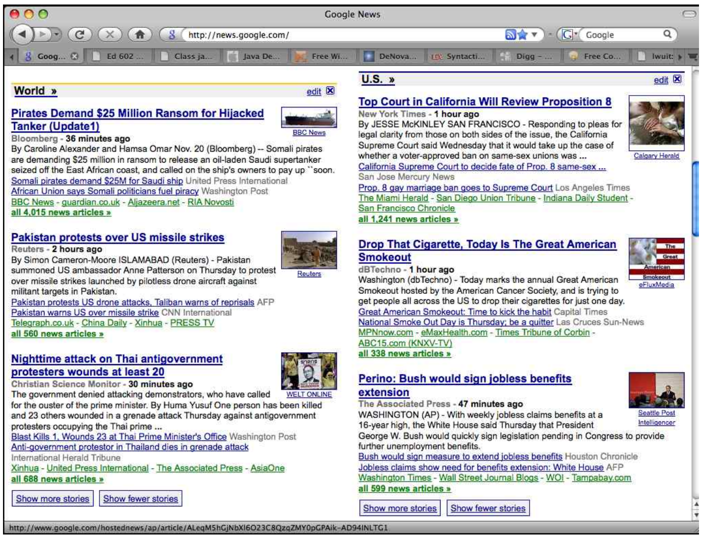
Clustering is widely used in many IR applications

Topic Hierarchy of Yahoo!
: the kind of output we obtain from clustering



Yahoo Science hierarchy, consisting of web pages that were gathered in summer of 1997 and pointed to by Yahoo
This hierarchy has totally 264 classes, but only samples are shown here

Google News: automatic clustering gives an effective news presentation style



World »

Pirates Demand $25 Million Ransom for Hijacked Tanker (Update1)
Bloomberg - **36 minutes ago**
By Caroline Alexander and Hamsa Omar Nov. 20 (Bloomberg) -- Somali pirates are demanding $25 million in ransom to release an oil-laden Saudi supertanker seized off the East African coast, and called on the ship's owners to pay up ``soon.
Somali pirates demand $25M for Saudi ship United Press International
African Union says Somali politicians fuel piracy Washington Post
BBC News - guardian.co.uk - Aljazeera.net - RIA Novosti
all 4,015 news articles »

Pakistan protests over US missile strikes
Reuters - **2 hours ago**
By Simon Cameron-Moore ISLAMABAD (Reuters) - Pakistan summoned US ambassador Anne Patterson on Thursday to protest over missile strikes launched by pilotless drone aircraft against militant targets in Pakistan.
Pakistan protests US drone attacks, Taliban warns of reprisals AFP
Pakistan warns US over missile strike CNN International
Telegraph.co.uk - China Daily - Xinhua - PRESS TV
all 560 news articles »

Nighttime attack on Thai antigovernment protesters wounds at least 20
Christian Science Monitor - **30 minutes ago**
The government denied attacking demonstrators, who have called for the ouster of the prime minister. By Huma Yusuf One person has been killed and 23 others wounded in a grenade attack Thursday against antigovernment protesters occupying the Thai prime ...
Blast Kills 1, Wounds 23 at Thai Prime Minister's Office Washington Post
Anti-government protestor in Thailand dies in grenade attack International Herald Tribune
Xinhua - United Press International - The Associated Press - AsiaOne
all 688 news articles »

Show more stories | Show fewer stories

U.S. »

Top Court in California Will Review Proposition 8
New York Times - **1 hour ago**
By JESSE McKINLEY SAN FRANCISCO - Responding to pleas for legal clarity from those on both sides of the issue, the California Supreme Court said Wednesday that it would take up the case of whether a voter-approved ban on same-sex unions was ...
California Supreme Court to decide fate of Prop. 8 same-sex ... San Jose Mercury News
Prop. 8 gay marriage ban goes to Supreme Court Los Angeles Times
The Miami Herald - San Diego Union Tribune - Indiana Daily Student - San Francisco Chronicle
all 1,241 news articles »

Drop That Cigarette, Today Is The Great American Smokeout
dBTechno - **1 hour ago**
Washington (dbTechno) - Today marks the annual Great American Smokeout hosted by the American Cancer Society, and is trying to get people all across the US to drop their cigarettes for just one day.
Great American Smokeout: Time to kick the habit Capital Times
National Smoke Out Day is Thursday; be a quitter Las Cruces Sun-News
MPNnow.com - eMaxHealth.com - Times Tribune of Corbin - ABC15.com (KNXV-TV)
all 338 news articles »

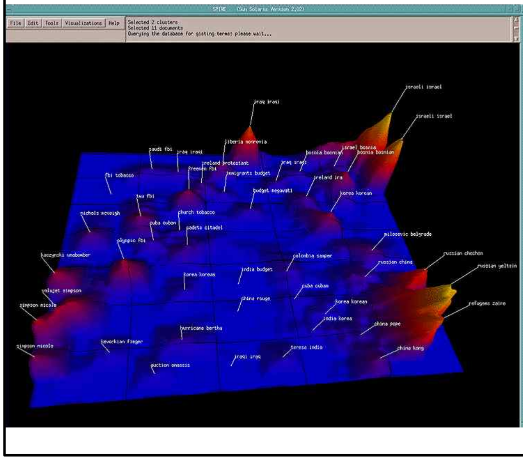
Perino: Bush would sign jobless benefits extension
The Associated Press - **47 minutes ago**
WASHINGTON (AP) - With weekly jobless claims benefits at a 16-year high, the White House said Thursday that President George W. Bush would quickly sign legislation pending in Congress to provide further unemployment benefits.
Bush would sign measure to extend jobless benefits Houston Chronicle
Jobless claims show need for benefits extension: White House AFP
Washington Times - Wall Street Journal Blogs - WOI - Tampabay.com
all 599 news articles »

Show more stories | Show fewer stories

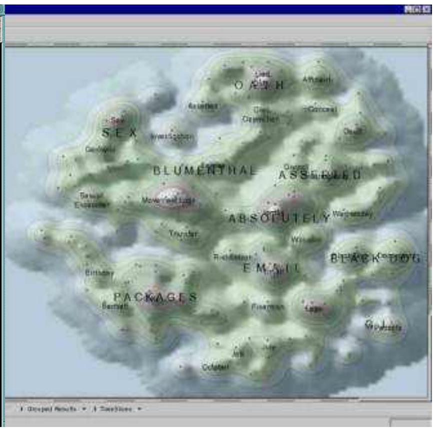
http://www.google.com/hostednews/ap/article/ALeqM5hGjNbXl6O23C8QzqZMY0pGPAik-AD94INLTG1

Clustering is utilized for visualizing a document collection based on its themes

» In this contour map, each mountain stands for a certain cluster (with its most frequent term) and mountain height stands for its cluster size



Selected 2 clusters
Selected 11 documents
Querying the database for gisting terms: please wait...



Clustering can be used for improving recall in IR system

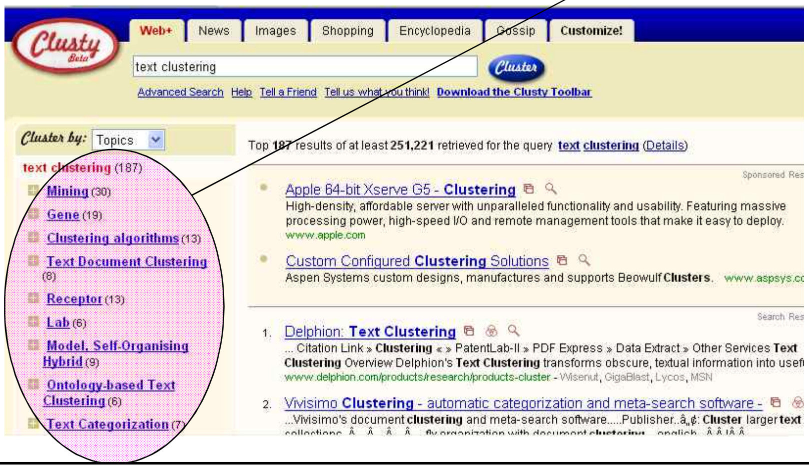
- *Basic Cluster hypothesis* : Documents in the same cluster behave similarly with respect to relevance to information needs
- Therefore, to improve recall in IR system
  - » Prior process: we can cluster all the documents in collection
  - » When a query matches a document *D*(that is, D is relevant to that query), IR system return other documents in the cluster containing that document *D*
- By doing this, we hope that querying with “car” will also return documents containing “*automobile*”
  - » Because clustering groups documents containing “*car*” together with documents containing “*automobile*”

Why might this happen?

A document containing “*car*” is likely to have many common terms(or words) with a document containing “*automobile*” so that the similarity between the two documents is apt to be very high in vector space model

Clustering can be also used for better navigation of search results in IR system

- As shown in the picture, this search engine gives well-categorized search results that are grouped thematically by clustering



text clustering (187)
- Mining (30)
- Gene (19)
- Clustering algorithms (13)
- Text Document Clustering (8)
- Receptor (13)
- Lab (6)
- Model, Self-Organising Hybrid (9)
- Ontology-based Text Clustering (6)
- Text Categorization (7)

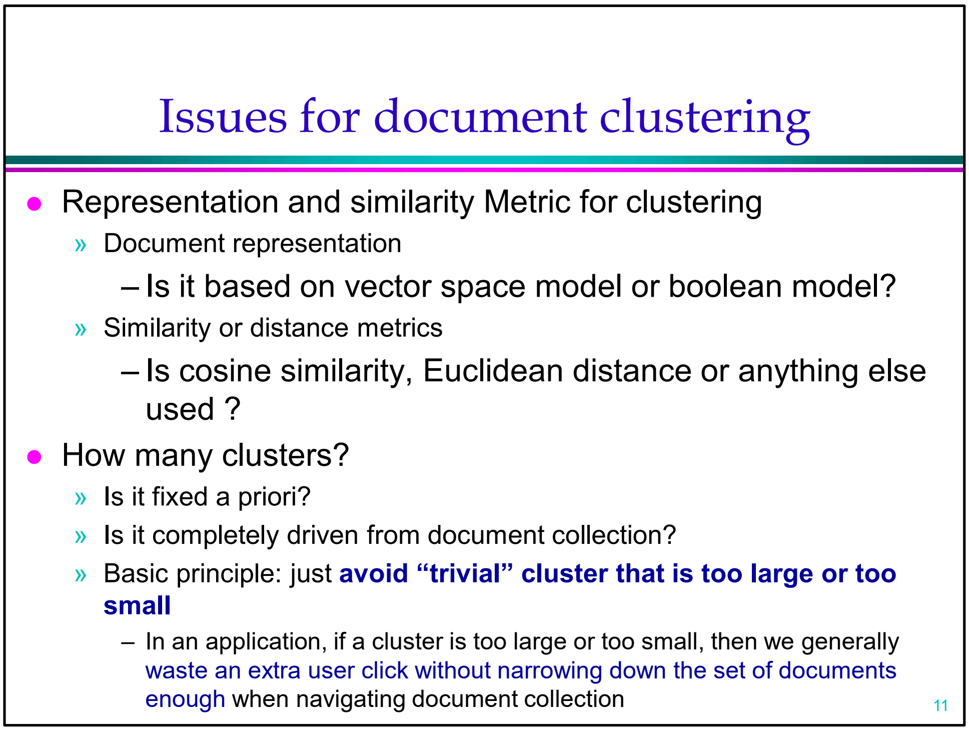
Top 187 results of at least 251,221 retrieved for the query text clustering (Details)

- Apple 64-bit Xserve G5 - Clustering
  High-density, affordable server with unparalleled functionality and usability. Featuring massive processing power, high-speed I/O and remote management tools that make it easy to deploy.
  www.apple.com
- Custom Configured Clustering Solutions
  Aspen Systems custom designs, manufactures and supports Beowulf Clusters. www.aspsys.cc

1. Delphion: Text Clustering
   ... Citation Link » Clustering « » PatentLab-II » PDF Express » Data Extract » Other Services Text Clustering Overview Delphion's Text Clustering transforms obscure, textual information into usef
   www.delphion.com/products/research/products-cluster - Wisenut, GigaBlast, Lycos, MSN
2. Vivisimo Clustering - automatic categorization and meta-search software -
   ...Vivisimo's document clustering and meta-search software.....Publisher..â„¢: Cluster larger text

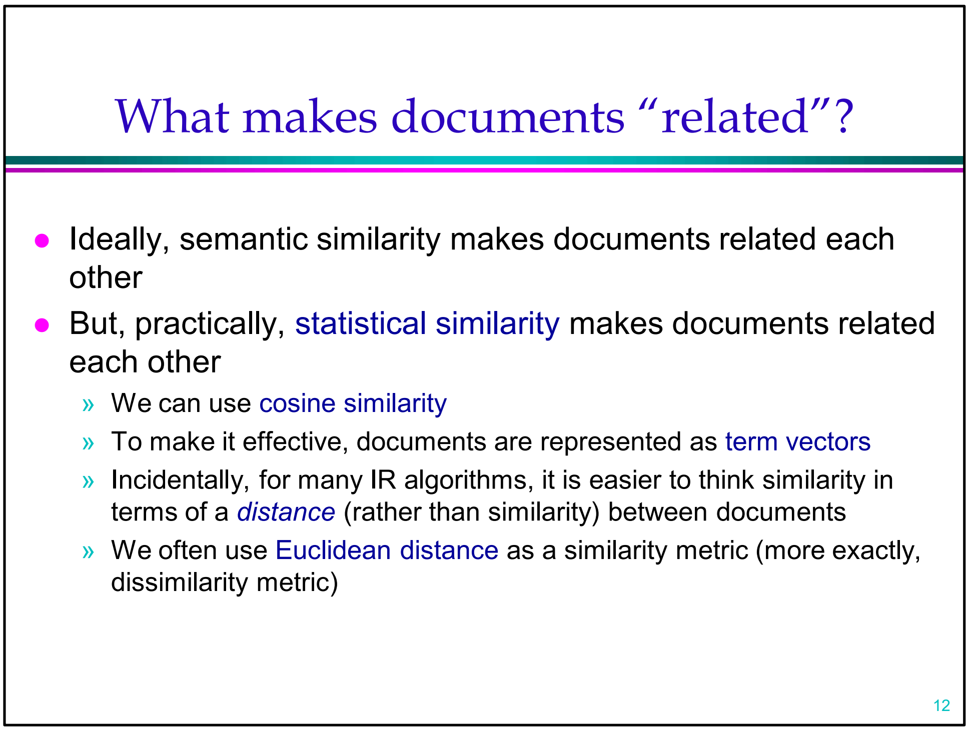
Document clustering

- Given a set of documents, document clustering groups those documents into a number of clusters based on their content similarity



Issues for document clustering

- Representation and similarity Metric for clustering
  - » Document representation
    - – Is it based on vector space model or boolean model?
  - » Similarity or distance metrics
    - – Is cosine similarity, Euclidean distance or anything else used ?
- How many clusters?
  - » Is it fixed a priori?
  - » Is it completely driven from document collection?
  - » Basic principle: just **avoid "trivial" cluster that is too large or too small**
    - – In an application, if a cluster is too large or too small, then we generally waste an extra user click without narrowing down the set of documents enough when navigating document collection

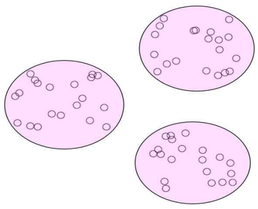


What makes documents "related"?

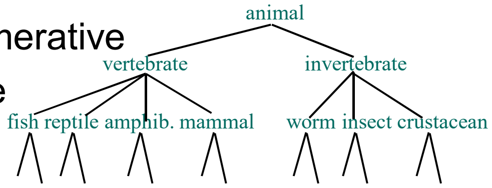
- Ideally, semantic similarity makes documents related each other
- But, practically, statistical similarity makes documents related each other
  - » We can use cosine similarity
  - » To make it effective, documents are represented as term vectors
  - » Incidentally, for many IR algorithms, it is easier to think similarity in terms of a *distance* (rather than similarity) between documents
  - » We often use Euclidean distance as a similarity metric (more exactly, dissimilarity metric)

Clustering Algorithms

- Flat algorithm
  - » Usually, it start with a random partitioning
  - » Gradually, it is refined iteratively
    - – *e.g. K*-means clustering



- Hierarchical algorithms
  - » Bottom-up, agglomerative
  - » Top-down, divisive



K-Means (flat clustering algorithm)

- It assumes that documents are real-valued term vectors
- Cluster is based on *centroid* (the center of gravity or mean) of term vectors:

  In a cluster *c* , $\vec{\mu}(c) = \frac{1}{|c|}\sum_{\vec{x}\in c}\vec{x}$

- Re-assigning instances to clusters is based on Euclidean distance to each cluster centroid.

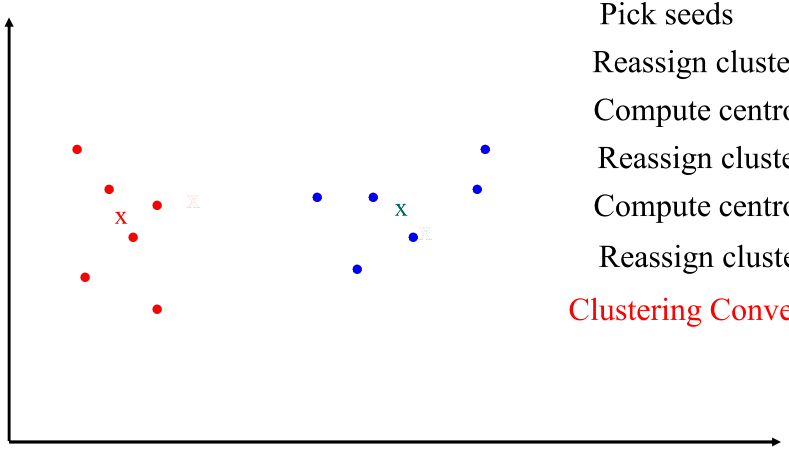
Note that henceforth, we can equivalently phrase Euclidean distance in terms of dissimilarity)

K-Means Algorithm for making *K* clusters

```
select $K$ random docs $\{s_1, s_2, \ldots s_K\}$ as seeds.
do
{
    for each doc $d_i$
    {
        assign $d_i$ to the cluster $c_j$ such that $dist(d_i, s_j)$ is minimal ;
    }
    for each cluster $c_j$        // update the seeds to the centroid of each cluster
    {
        $s_j = \mu(c_j)$ ;
    }
} until(clustering is converged enough or termination condition is met)
```

Much simpler and easier than expected !

K Means Example (*K*=2)



Pick seeds

Reassign clusters

Compute centroids

Reassign clusters

Compute centroids

Reassign clusters

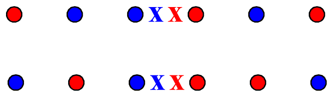
Clustering Converged!

Termination conditions

- We may have several options for termination condition of K-means algorithm
- e.g.
  - » just terminate in a fixed number of iterations
  - » terminate when all clusters are unchanged at all
  - » terminate when all cluster centroid positions are unchanged at all

Does the second termination condition is equivalent to the third one?

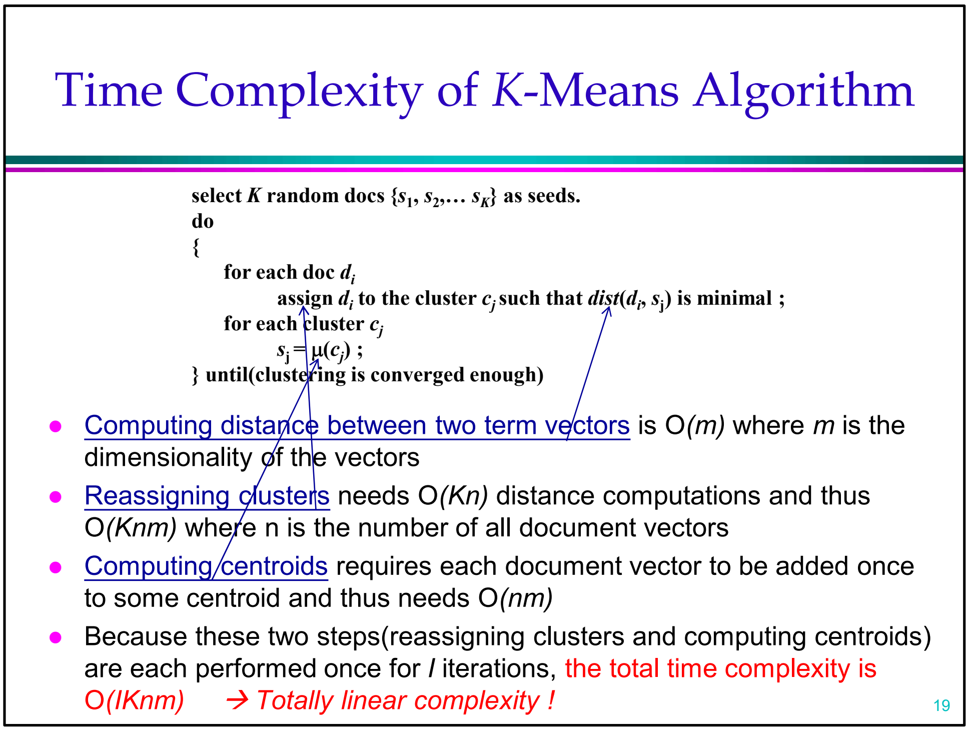
yes or **no** ?



Convergence

- Why should the *K*-means algorithm ever reach a *fixed point*?
  - » There is always a state in which all clusters don't change
- Indirect Answer: *K*-means algorithm is a special case of a general procedure known as *Expectation Maximization (EM) algorithm*
  - » EM is proved to converge to a stable state

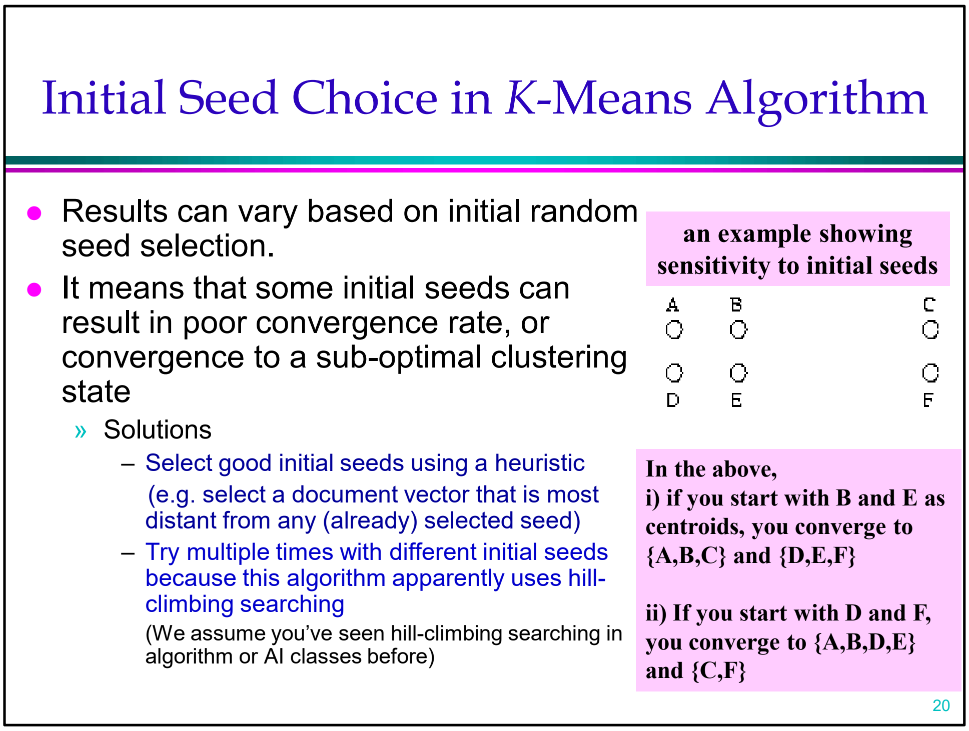
    (We assume you've seen EM algorithm in AI class before)
  - » Number of iterations for convergence can be large, but in practice it is usually not be large (much smaller than expected)



Time Complexity of *K*-Means Algorithm

```
select $K$ random docs $\{s_1, s_2, \ldots s_K\}$ as seeds.
do
{
    for each doc $d_i$
        assign $d_i$ to the cluster $c_j$ such that $dist(d_i, s_j)$ is minimal ;
    for each cluster $c_j$
        $s_j = \mu(c_j)$ ;
} until(clustering is converged enough)
```

- Computing distance between two term vectors is O*(m)* where *m* is the dimensionality of the vectors
- Reassigning clusters needs O*(Kn)* distance computations and thus O*(Knm)* where n is the number of all document vectors
- Computing centroids requires each document vector to be added once to some centroid and thus needs O*(nm)*
- Because these two steps(reassigning clusters and computing centroids) are each performed once for *I* iterations, the total time complexity is O*(IKnm)*   → *Totally linear complexity !*



Initial Seed Choice in *K*-Means Algorithm

- Results can vary based on initial random seed selection.
- It means that some initial seeds can result in poor convergence rate, or convergence to a sub-optimal clustering state
  - » Solutions
    - – Select good initial seeds using a heuristic (e.g. select a document vector that is most distant from any (already) selected seed)
    - – Try multiple times with different initial seeds because this algorithm apparently uses hill-climbing searching
      (We assume you've seen hill-climbing searching in algorithm or AI classes before)

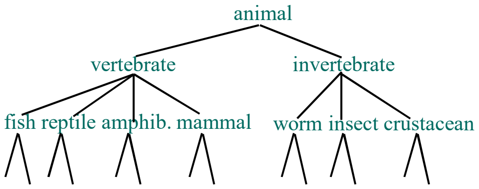
an example showing sensitivity to initial seeds

In the above,
i) if you start with B and E as centroids, you converge to {A,B,C} and {D,E,F}

ii) If you start with D and F, you converge to {A,B,D,E} and {C,F}

Hierarchical Clustering Algorithm

- Two ways
  - » Divisive Clustering (top down)
    - – The complete document collection is assumed to represent one complete cluster
    - – Then, each cluster is subsequently and recursively broken down into smaller clusters using flat clustering algorithm like *k*-means algorithm



  - » Hierarchical Agglomerative Clustering (bottom up)
    - – Individual document similarities are used as a starting point.
    - – Based on it, a gluing operation agglomeratively collects similar documents (or similar document clusters) into a super-ordinate (or larger) cluster

Hierarchical Agglomerative Clustering

- Basic procedure

1. Place each of *N* document vectors into a cluster of its own ;
2. Compute all pairwise cluster-to-cluster similarities ;
3. Form a new larger cluster by combining the pair of most similar clusters ;
4. Goto step 2 while the number of clusters left is greater than 1 ;

How to Combine Clusters?

- Basic procedure

  1. Place each of *N* document vectors into a cluster of its own ;
  2. Compute all pairwise cluster-to-cluster similarities ;
  3. Form a new larger cluster by combining the pair of most similar clusters ;
  4. Goto step 2 if the number of clusters left is greater than 1 ;

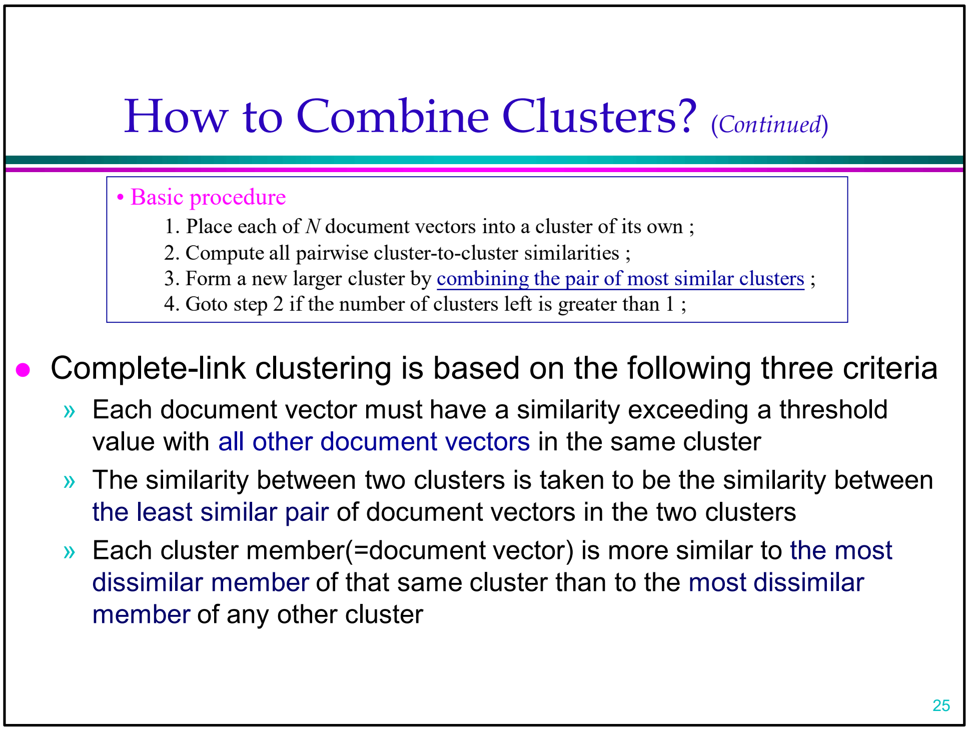
- Just combine two clusters using intercluster similarity
  - » Three options
    - – Single-link clustering
    - – Complete-link clustering
    - – (Group) average-link clustering

How to Combine Clusters?

- Basic procedure

  1. Place each of *N* document vectors into a cluster of its own ;
  2. Compute all pairwise cluster-to-cluster similarities ;
  3. Form a new larger cluster by combining the pair of most similar clusters ;
  4. Goto step 2 if the number of clusters left is greater than 1 ;

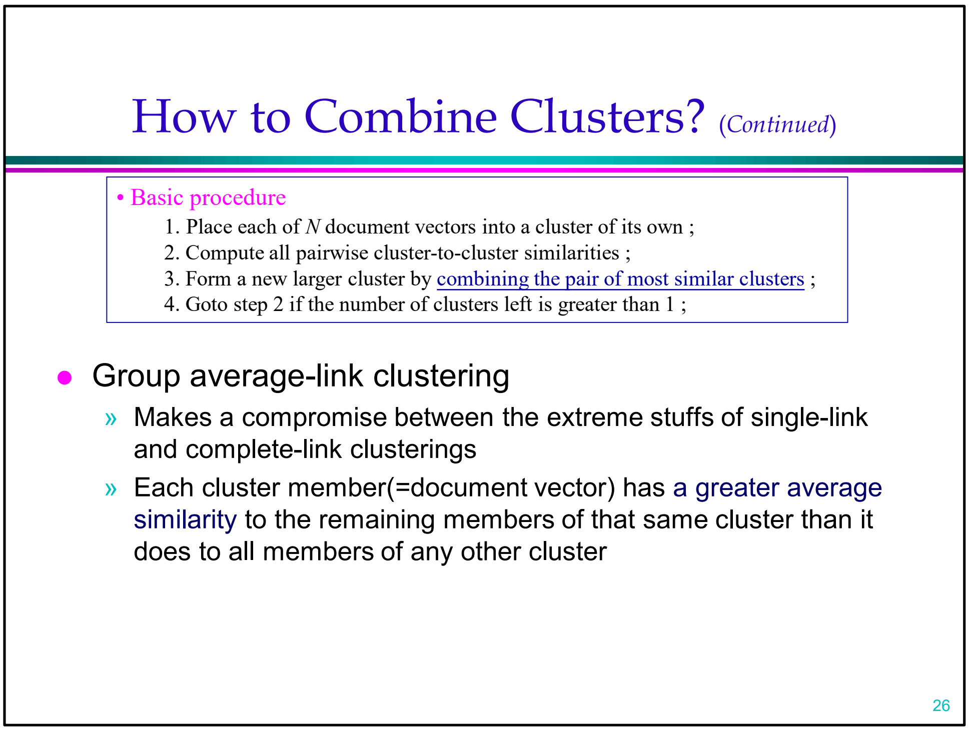
- Single-link clustering is based on the following three criteria
  - » Each document vector must have a similarity exceeding a threshold value with at least one other document vector in the same cluster
  - » The similarity between two clusters is taken to be the similarity between the most similar pair of document vectors in the two clusters
  - » Each cluster member(=document vector) will be more similar to at least one member in that same cluster than to any member of another cluster



How to Combine Clusters? *(Continued)*

- Basic procedure
  1. Place each of *N* document vectors into a cluster of its own ;
  2. Compute all pairwise cluster-to-cluster similarities ;
  3. Form a new larger cluster by combining the pair of most similar clusters ;
  4. Goto step 2 if the number of clusters left is greater than 1 ;

- Complete-link clustering is based on the following three criteria
  - Each document vector must have a similarity exceeding a threshold value with all other document vectors in the same cluster
  - The similarity between two clusters is taken to be the similarity between the least similar pair of document vectors in the two clusters
  - Each cluster member(=document vector) is more similar to the most dissimilar member of that same cluster than to the most dissimilar member of any other cluster



How to Combine Clusters? *(Continued)*

- Basic procedure
  1. Place each of *N* document vectors into a cluster of its own ;
  2. Compute all pairwise cluster-to-cluster similarities ;
  3. Form a new larger cluster by combining the pair of most similar clusters ;
  4. Goto step 2 if the number of clusters left is greater than 1 ;

- Group average-link clustering
  - Makes a compromise between the extreme stuffs of single-link and complete-link clusterings
  - Each cluster member(=document vector) has a greater average similarity to the remaining members of that same cluster than it does to all members of any other cluster

Single-link, complete-link, and group average-link clusterings

- Note: for easy understanding of single-link, complete-link, and group average-link clusterings, we may need some examples and try out each of them for ourselves

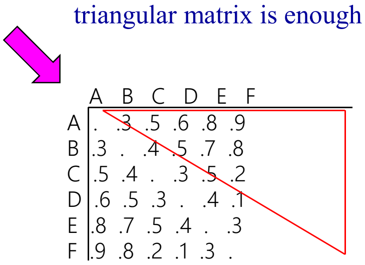
Detailed examples for single-link clusterings in terms of run-time data structure

A-F (6 items) 6(6-1)/2 (15) pairwise similarities

Rank	Pair	Similarity	Rank	Pair	Similarity
1	AF	0.9	9	BC	0.4
2	AE	0.8	10	DE	0.4
3	BF	0.8	11	AB	0.3
4	BE	0.7	12	CD	0.3
5	AD	0.6	13	EF	0.3
6	AC	0.5	14	CF	0.2
7	BD	0.5	15	DF	0.1
8	CE	0.5			

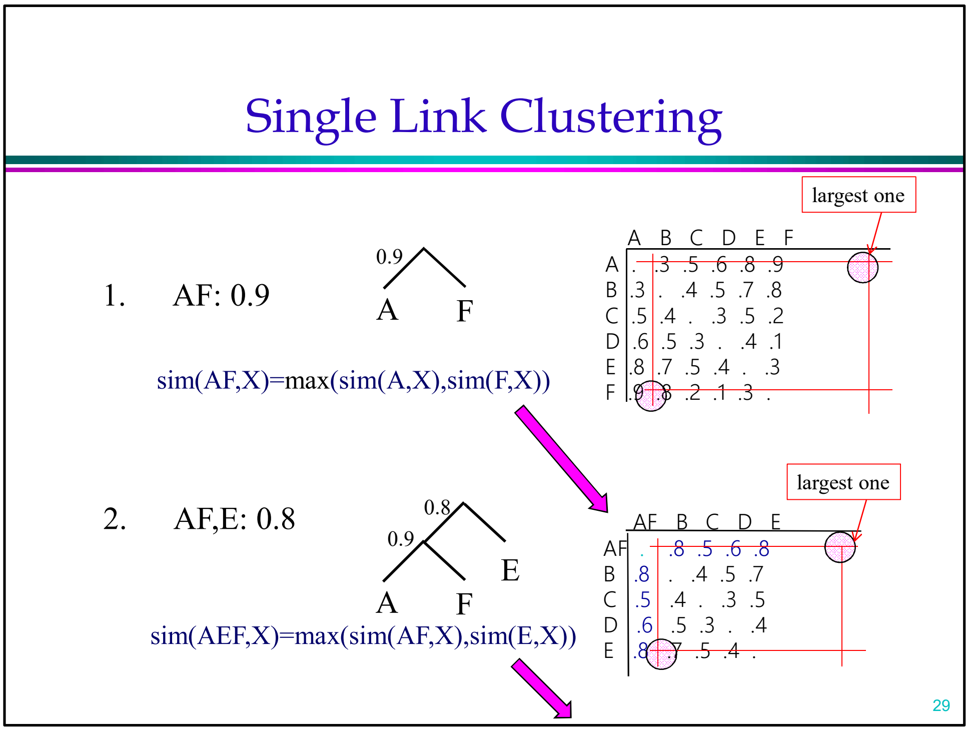
decreasing order

Actually only a upper or lower triangular matrix is enough !



	A	B	C	D	E	F
A	.	.3	.5	.6	.8	.9
B	.3	.	.4	.5	.7	.8
C	.5	.4	.	.3	.5	.2
D	.6	.5	.3	.	.4	.1
E	.8	.7	.5	.4	.	.3
F	.9	.8	.2	.1	.3	.

initial data structure



Single Link Clustering

1. AF: 0.9

sim(AF,X)=max(sim(A,X),sim(F,X))

largest one

	A	B	C	D	E	F
A	.	.3	.5	.6	.8	.9
B	.3	.	.4	.5	.7	.8
C	.5	.4	.	.3	.5	.2
D	.6	.5	.3	.	.4	.1
E	.8	.7	.5	.4	.	.3
F	.9	.8	.2	.1	.3	.

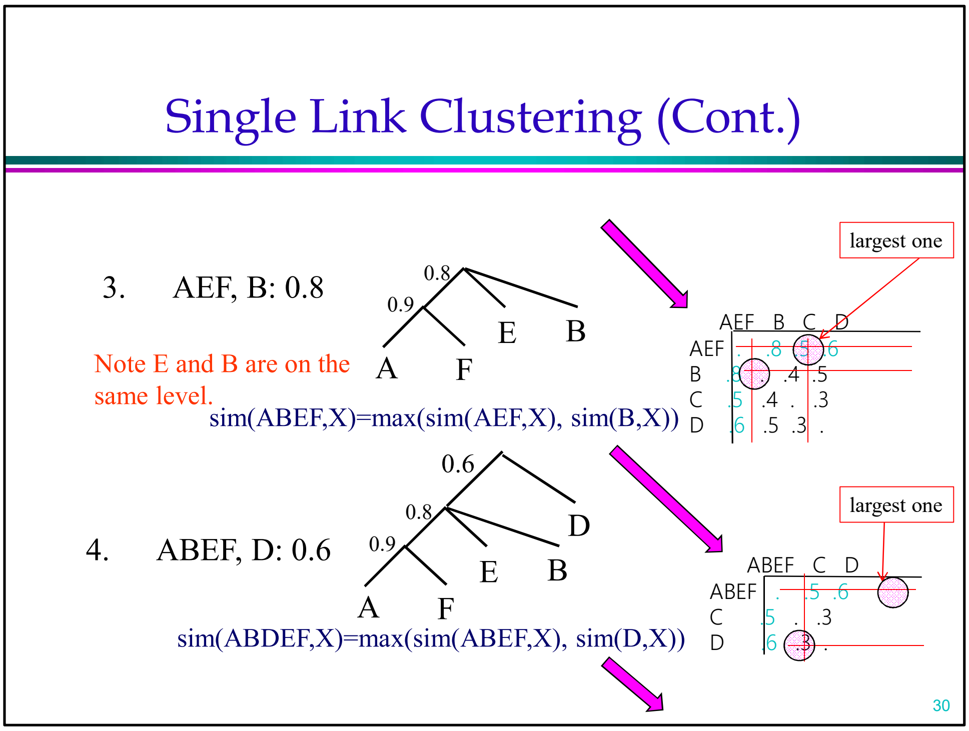
2. AF,E: 0.8

sim(AEF,X)=max(sim(AF,X),sim(E,X))

largest one

	AF	B	C	D	E
AF	.	.8	.5	.6	.8
B	.8	.	.4	.5	.7
C	.5	.4	.	.3	.5
D	.6	.5	.3	.	.4
E	.8	.7	.5	.4	.

29



Single Link Clustering (Cont.)

3. AEF, B: 0.8

Note E and B are on the same level.

sim(ABEF,X)=max(sim(AEF,X), sim(B,X))

largest one

	AEF	B	C	D
AEF	.	.8	.5	.6
B	.8	.	.4	.5
C	.5	.4	.	.3
D	.6	.5	.3	.

4. ABEF, D: 0.6

sim(ABDEF,X)=max(sim(ABEF,X), sim(D,X))

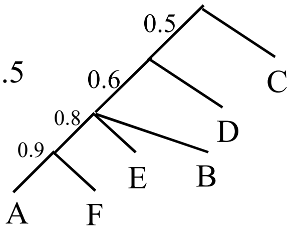
largest one

	ABEF	C	D
ABEF	.	.5	.6
C	.5	.	.3
D	.6	.3	.

30

Single Link Clustering (Cont.)

5. ABDEF, C: 0.5



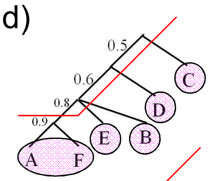
final one

	ABDEF	C
ABDEF	.	.5
C	.5	.

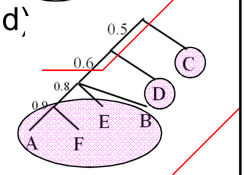
ABCDEF	.

Single-Link Clusters

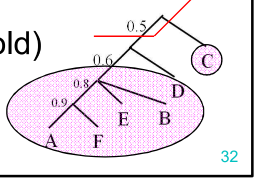
- Similarity level 0.85 (i.e., similarity threshold)
  - » AF, E, B, D, C



- Similarity level 0.65 (i.e., similarity threshold)
  - » AFEB, D, C



- Similarity level 0.55 (i.e., similarity threshold)
  - » AFEBD, C

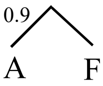


Detailed examples for complete-link clusterings in terms of run-time data structure

Rank	Pair	Similarity	Rank	Pair	Similarity
1	AF	0.9	9	BC	0.4
2	AE	0.8	10	DE	0.4
3	BF	0.8	11	AB	0.3
4	BE	0.7	12	CD	0.3
5	AD	0.6	13	EF	0.3
6	AC	0.5	14	CF	0.2
7	BD	0.5	15	DF	0.1
8	CE	0.5			

Step Number | Similarity Pair | **Checked List** | Link Structure

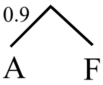
1. AF 0.9 AF



A and F are clustered

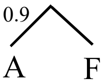
sim(AF,x)=**min**(sim(A,x), sim(F,x))

2. AE 0.8 AE, AF



Note: EF is required in checked list for clustering AF and E

3. BF 0.8 AE, AF, BF



Note: AB is required in checked list for clustering AF and B

Similarity Matrix

	A	B	C	D	E	F
A	.	.3	.5	.6	.8	.9
B	.3	.	.4	.5	.7	.8
C	.5	.4	.	.3	.5	.2
D	.6	.5	.3	.	.4	.1
E	.8	.7	.5	.4	.	.3
F	.9	.8	.2	.1	.3	.

	AF	B	C	D	E
AF	.	.3	.2	.1	.3
B	.3	.	.4	.5	.7
C	.2	.4	.	.3	.5
D	.1	.5	.3	.	.4
E	.3	.7	.5	.4	.

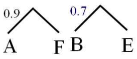
	AF	B	C	D	E
AF	.	.3	.2	.1	.3
B	.3	.	.4	.5	.7
C	.2	.4	.	.3	.5
D	.1	.5	.3	.	.4
E	.3	.7	.5	.4	.

Complete-Linke Cluster Generation (Cont.)

Rank	Pair	Similarity	Rank	Pair	Similarity
1	AF	0.9	9	BC	0.4
2	AE	0.8	10	DE	0.4
3	BF	0.8	11	AB	0.3
4	BE	0.7	12	CD	0.3
5	AD	0.6	13	EF	0.3
6	AC	0.5	14	CF	0.2
7	BD	0.5	15	DF	0.1
8	CE	0.5			

Step Number | Similarity Pair | **Checked List** | Link Structure

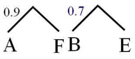
4. BE 0.7 AE, AF, **BE**, BF



B and E are clustered

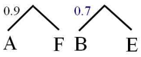
sim(BE,x)=**min**(sim(B,x), sim(E,x))

5. AD 0.6 **AD**, AE, AF, BE, BF



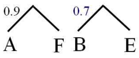
Note: DF is required in checked list for clustering AF and D

6. AC 0.6 **AC**, AD, AE, AF, BE, BF



Note: CF is required in checked list for clustering AF and C

7. BD 0.5 AC, AD, AE, AF, **BD**, BE, BF



Note: DE is required in checked list for clustering BE and D

Similarity Matrix

	AF	B	C	D	E
AF	.	.3	.2	.1	.3
B	.3	.	.4	.5	.7
C	.2	.4	.	.3	.5
D	.1	.5	.3	.	.4
E	.3	.7	.5	.4	.

	AF	BE	C	D
AF	.	.3	.2	.1
BE	.3	.	.4	.4
C	.2	.4	.	.3
D	.1	.4	.3	.

	AF	BE	C	D
AF	.	.3	.2	.1
BE	.3	.	.4	.4
C	.2	.4	.	.3
D	.1	.4	.3	.

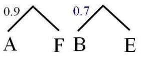
	AF	BE	C	D
AF	.	.3	.2	.1
BE	.3	.	.4	.4
C	.2	.4	.	.3
D	.1	.4	.3	.

Complete-Linke Cluster Generation (Cont.)

Rank	Pair	Similarity	Rank	Pair	Similarity
1	AF	0.9	9	BC	0.4
2	AE	0.8	10	DE	0.4
3	BF	0.8	11	AB	0.3
4	BE	0.7	12	CD	0.3
5	AD	0.6	13	EF	0.3
6	AC	0.5	14	CF	0.2
7	BD	0.5	15	DF	0.1
8	CE	0.5			

Step Number | Similarity Pair | **Checked List** | Link Structure | Similarity Matrix

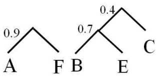
8. CE 0.5 — AC, AD, AE, AF, BD, BE, BF, **CE**



$$\begin{array}{c|cccc} & AF & BE & C & D \\ \hline AF & . & .3 & .2 & .1 \\ BE & .3 & . & .4 & .4 \\ C & .2 & .4 & . & .3 \\ D & .1 & .4 & .3 & . \end{array}$$

Note: BC is required in checked list for clustering BE and C

9. BC 0.4 — AC, AD, AE, AF, **BC**, BD, BE, BF, CE



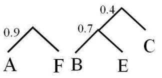
$$\begin{array}{c|cccc} & AF & BE & C & D \\ \hline AF & . & .3 & .2 & .1 \\ BE & .3 & . & .4 & .4 \\ C & .2 & .4 & . & .3 \\ D & .1 & .4 & .3 & . \end{array}$$

BE and C are clustered

sim(BCE,x)=**min**(sim(BE,x), sim(C,x))

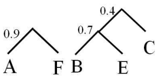
$$\begin{array}{c|ccc} & AF & BCE & D \\ \hline AF & . & .2 & .1 \\ BCE & .2 & . & .3 \\ D & .1 & .3 & . \end{array}$$

10. DE 0.4 — AC, AD, AE, AF, BC, BD, BE, BF, CE, **DE**



Note: CD is required in checked list for clustering BCE and D

11. AB 0.3 — **AB**, AC, AD, AE, AF, BC, BD, BE, BF, CE, DE



$$\begin{array}{c|ccc} & AF & BCE & D \\ \hline AF & . & .2 & .1 \\ BCE & .2 & . & .3 \\ D & .1 & .3 & . \end{array}$$

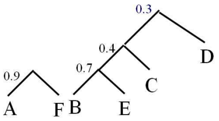
Note: CF and EF are required in checked list for clustering BCE and D

Complete-Linke Cluster Generation (Cont.)

Rank	Pair	Similarity	Rank	Pair	Similarity
1	AF	0.9	9	BC	0.4
2	AE	0.8	10	DE	0.4
3	BF	0.8	11	AB	0.3
4	BE	0.7	12	CD	0.3
5	AD	0.6	13	EF	0.3
6	AC	0.5	14	CF	0.2
7	BD	0.5	15	DF	0.1
8	CE	0.5			

Step Number | Similarity Pair | **Checked List** | Link Structure | Similarity Matrix

12. CD 0.3 — AB, AC, AD, AE, AF, BC, BD, BE, BF, **CD**, CE, DE



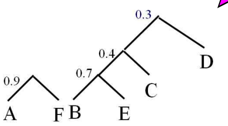
$$\begin{array}{c|ccc} & AF & BCE & D \\ \hline AF & . & .2 & .1 \\ BCE & .2 & . & .3 \\ D & .1 & .3 & . \end{array}$$

BCE and D are clustered

sim(BCDE,x)=**min**(sim(BCE,x), sim(D,x))

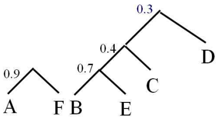
$$\begin{array}{c|cc} & AF & BCDE \\ \hline AF & . & .1 \\ BCDE & .1 & . \end{array}$$

13. EF 0.3 — AB, AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, DE, **EF**



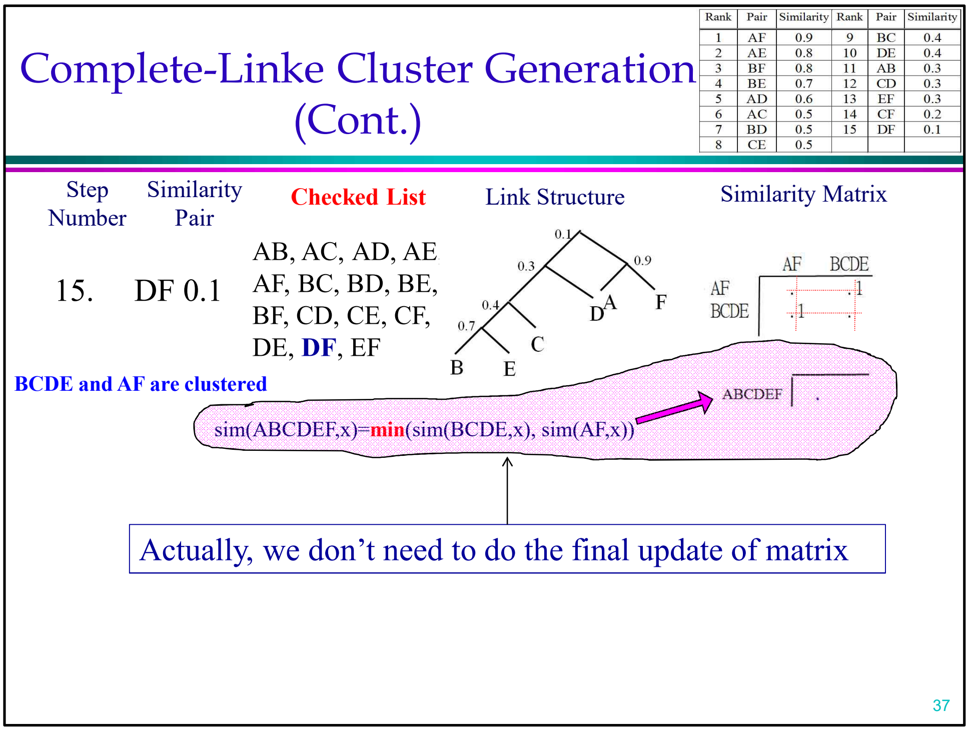
Note: CF and DF are required in checked list for clustering AF and BCDE

14. CF 0.2 — AB, AC, AD, AE, AF, BC, BD, BE, BF, CD, CE, **CF**, DE, EF



$$\begin{array}{c|cc} & AF & BCDE \\ \hline AF & . & .1 \\ BCDE & .1 & . \end{array}$$

Note: DF is required in checked list for clustering AF and BCDE



Complete-Linke Cluster Generation (Cont.)

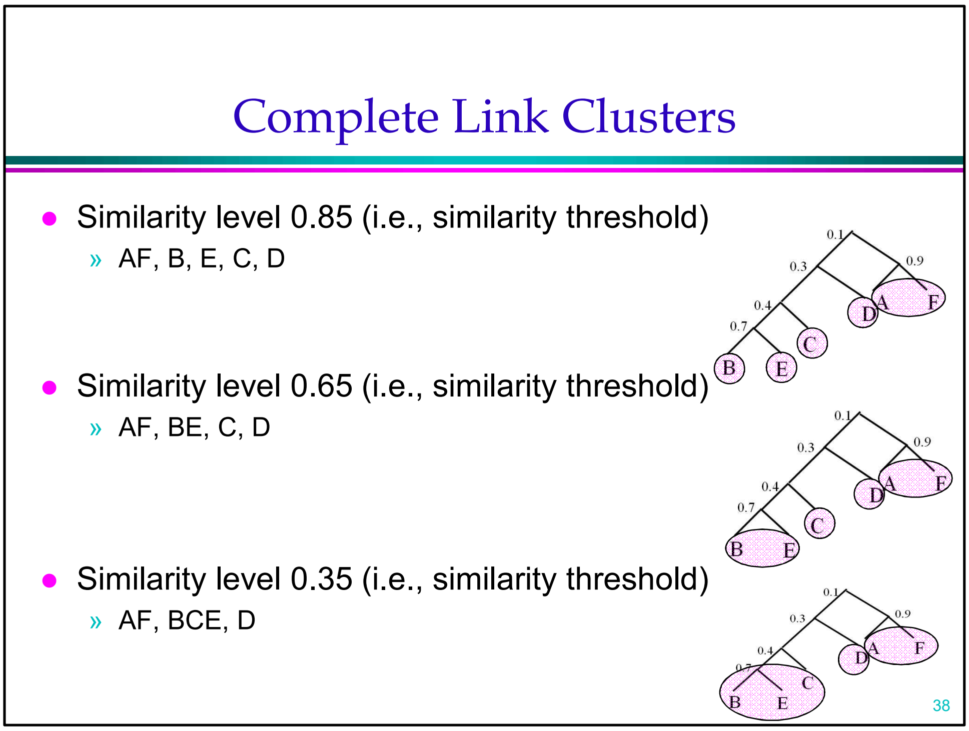
Rank	Pair	Similarity	Rank	Pair	Similarity
1	AF	0.9	9	BC	0.4
2	AE	0.8	10	DE	0.4
3	BF	0.8	11	AB	0.3
4	BE	0.7	12	CD	0.3
5	AD	0.6	13	EF	0.3
6	AC	0.5	14	CF	0.2
7	BD	0.5	15	DF	0.1
8	CE	0.5			

Step Number	Similarity Pair	**Checked List**	Link Structure	Similarity Matrix
15.	DF 0.1	AB, AC, AD, AE AF, BC, BD, BE, BF, CD, CE, CF, DE, **DF**, EF		

BCDE and AF are clustered

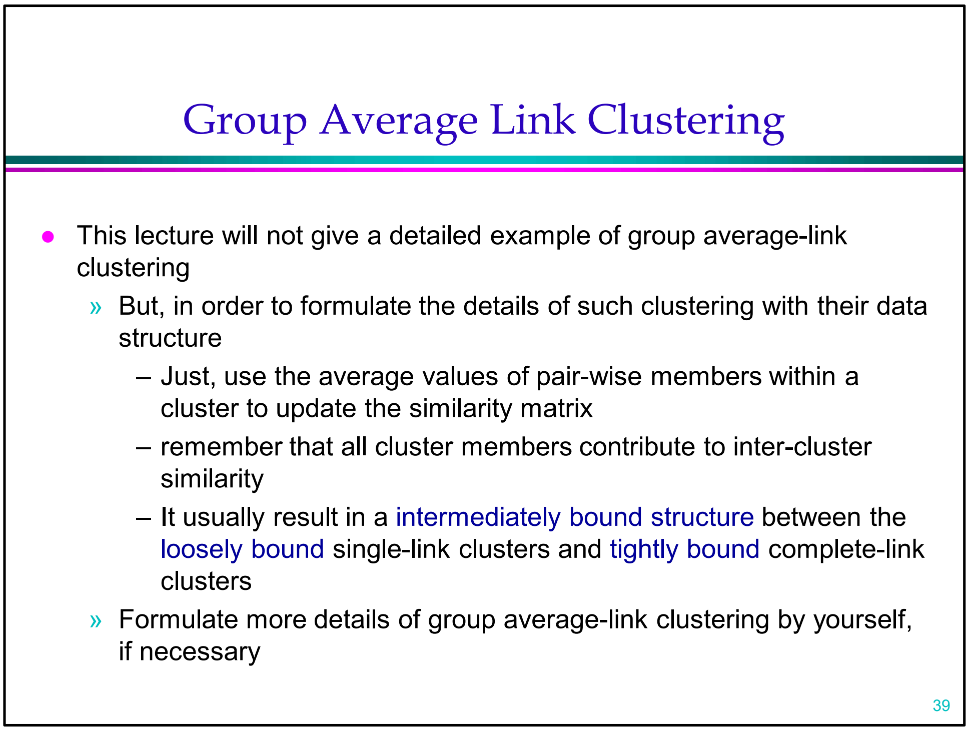
sim(ABCDEF,x)=**min**(sim(BCDE,x), sim(AF,x))

Actually, we don’t need to do the final update of matrix



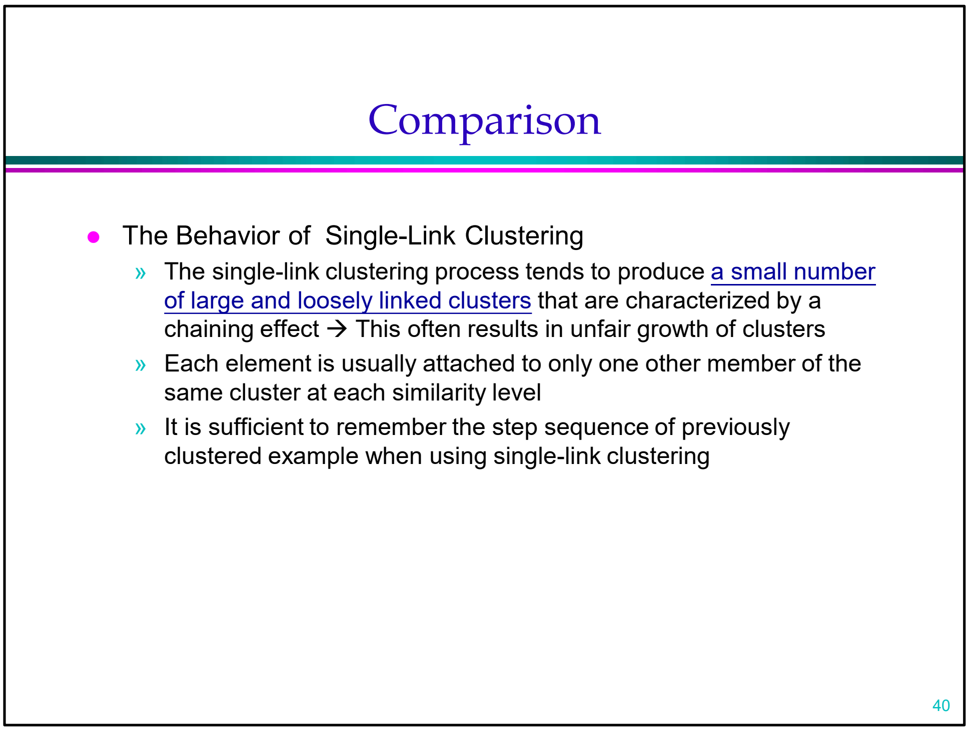
Complete Link Clusters

- Similarity level 0.85 (i.e., similarity threshold)
  - AF, B, E, C, D
- Similarity level 0.65 (i.e., similarity threshold)
  - AF, BE, C, D
- Similarity level 0.35 (i.e., similarity threshold)
  - AF, BCE, D



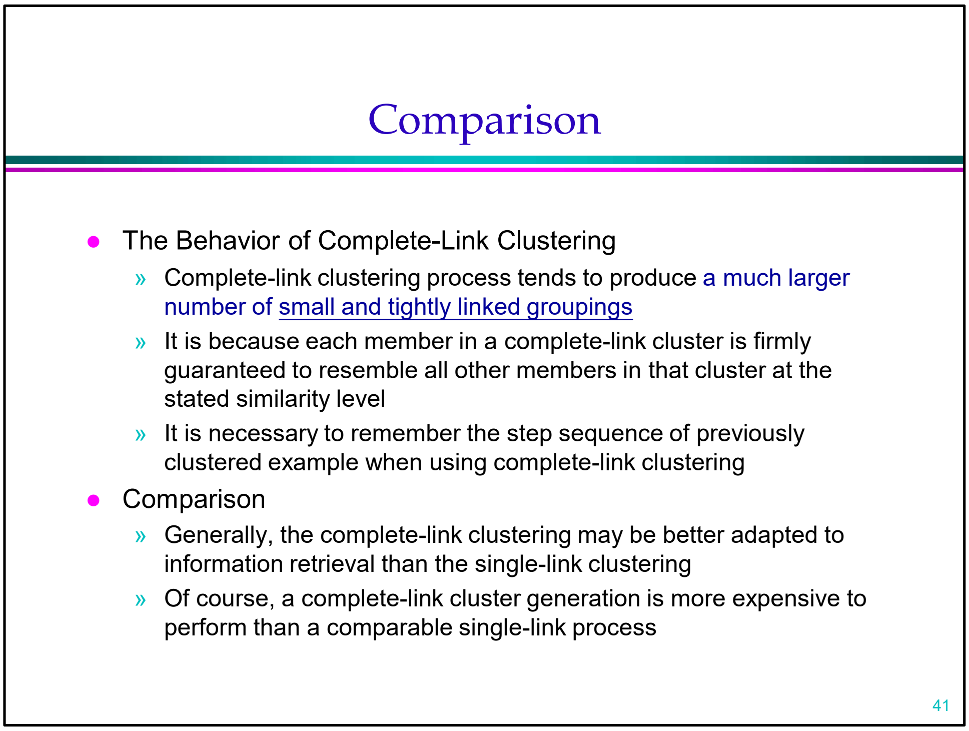
Group Average Link Clustering

- This lecture will not give a detailed example of group average-link clustering
  - » But, in order to formulate the details of such clustering with their data structure
    - – Just, use the average values of pair-wise members within a cluster to update the similarity matrix
    - – remember that all cluster members contribute to inter-cluster similarity
    - – It usually result in a intermediately bound structure between the loosely bound single-link clusters and tightly bound complete-link clusters
  - » Formulate more details of group average-link clustering by yourself, if necessary



Comparison

- The Behavior of Single-Link Clustering
  - » The single-link clustering process tends to produce <u>a small number of large and loosely linked clusters</u> that are characterized by a chaining effect → This often results in unfair growth of clusters
  - » Each element is usually attached to only one other member of the same cluster at each similarity level
  - » It is sufficient to remember the step sequence of previously clustered example when using single-link clustering



Comparison

- The Behavior of Complete-Link Clustering
  - Complete-link clustering process tends to produce a much larger number of small and tightly linked groupings
  - It is because each member in a complete-link cluster is firmly guaranteed to resemble all other members in that cluster at the stated similarity level
  - It is necessary to remember the step sequence of previously clustered example when using complete-link clustering
- Comparison
  - Generally, the complete-link clustering may be better adapted to information retrieval than the single-link clustering
  - Of course, a complete-link cluster generation is more expensive to perform than a comparable single-link process