CHAPTER 7

Module - 5

Clustering

Syllabus

CURE Algorithm, Stream-Computing, A Stream-Clustering, Algorithm, Initializing & Merging Buckets, Answering Queries

7.1 Introduction

Q. What is clustering algorithm?

Cluster Using REpresentative i.e. CURE is very efficient data clustering algorithm for specifically large databases. CURE is robust to outliers.

Traditional clustering algorithm :

- In traditional clustering, it selects for any one point and it is only point considered as a cluster i.e. clusters centroid approach.
- Points in a cluster appear close to each other compared to other data points of any other clusters. It works in eclipse shape in better way.
- Drawback of traditional clustering algorithm is all-points approach makes algorithm highly sensitive to outliers and a minute change in position of data points.
- Cluster centroid and all points approach not work on arbitrary shape.

7.2 CURE Algorithm

MU - May 16, Dec. 16, May 18, May 19

Q. Clearly explain how the CURE algorithm can be used to cluster big data sets. (May 16, May 19, 10 Marks)

Q. Explain CURE algorithm for large scale clustering. (Dec. 16, 10 Marks)

Q. Explain the CURE algorithm for clustering large datasets. Please illustrate the algorithm using appropriate figures. (May 18, 10 Marks)

- CURE algorithm works better in spherical as well as non-spherical clusters.
- CURE : An efficient clustering algorithm for large database : sudipto Guha, Rajeev Rastogi, Kyuseok Shim.
- It prefers a set of points which are scattered as representative cluster than all-points or centroid approach.
- CURE uses random sampling and partitioning to speed up clustering.

7.2.1 Overview of CURE (Cluster Using REpresentative)

Data

↓

Make random sample

↓

Make partitioning of sample

↓

Partially cluster partitions

↓

Eliminate outliers

↓

Cluster partial clusters

↓

Label data in disk

7.2.2 Hierarchical Clustering Algorithm

Q. Write procedure of CURE cluster algorithm?

- A centroid-based point 'c' is chosen. All remaining scattered points are just at a fraction distance of α to get shrunk towards centroid.
- Such multiple scattered points help to discover in non spherical cluster i.e. elongated cluster.
- Hierarchical clustering algorithm uses such space which is linear to input size n.
- Worst-case time complexity is O (n^2 long n) and it may reduce to O(n^2) for lower dimensions.

CURE algorithm : CURE cluster procedure

- It is similar to hierarchical clustering approach. But it use sample point variant as cluster representative rather than every point in the cluster.
- First set a target sample number C. Then we try to select C well scattered sample points from cluster.
- The chosen scattered points are shrunk towards the centroid in a fraction of α where $0 \leq \alpha \leq 1$.

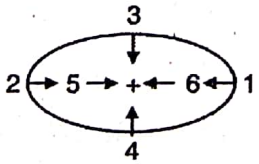


Fig. 7.2.1

- These points are used as representative of clusters and will be used as point in d_{min} cluster merging approach.
- After each merging, C sample points will be selected from original representative of previous clusters to represent new cluster.

- Cluster merging will be stopped until target K cluster is found.

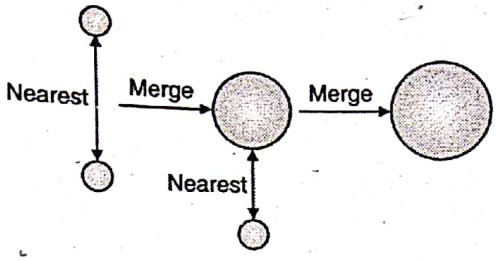


Fig. 7.2.2

7.2.2(A) Random Sampling and Partitioning Sample

MU - May 16

Q. Describe any two sampling techniques for big data with the help of examples. **(May 16, 10 Marks)**

Q. What is sampling? Explain random sampling and partition sampling.

- To reduce size of input to CURE's clustering algorithm random sampling is used in case of large data sets.
- Good clusters can be obtained by moderate size random samples, it provides tradeoff between efficiency and accuracy.
- Portioning sample reduces time required for execution because before final cluster made each partition get clustered whenever it is in pre-clustered data format at eliminated outliers.

7.2.2(B) Eliminate Outlier's and Data Labelling

Q. Write pseudo function of cluster algorithm.

Q. Write procedure for merging cluster in CURE.

- Outliers points are generally less than number in cluster.
- As random sample gets clustered, multiple representative points from each cluster are labelled with data set remainders.
- Clustering based on scattered point i.e. CURE approach found most efficient compared to centroid or all-points approach of traditional clustering algorithm.

Pseudo function of CURE (clustering algorithm)

```
Procedure cluster (s, k)
Begin
    T : =  build – kd – tree (s)
    Q : =  build – heap (s)
While size (Q) > k
do  {
    u : = extract – min (Q)
    v : = u – closest
    delete (Q, v)
    w : =  merge (u, v)
    delete – rep (T, u) ;
```

```
delete – rep (T, v) ;
insert – rep (T, w) ;
w – closest : = x
for each x ∈ Q
do {
    if dist (w, x) < dist (w, w – closest)
    w – closest : = x
    if x – closest is either u or v {
    if dist (x, x – closest) < dist (x, w)
    x – closest : = closest – cluster
    (T, x, dist (x, w))
else
x – closest : = w
relocate (Q, x)
 }
else if dist (x, x–closest) > dist (x, w){
x – closest : = w
relocate (Q, x)
 }
}
insert (Q, w)}
```

Procedure for merging clusters

```
Procedure merge (u, v)
being
w : = u U v
w. mean : = |u| u.mean + |v| v-mean/|u| + |v|
'tmpset : = ϕ
For i : = 1 to c do {
maxDist : = 0
for each point p in cluster w do {
if i = 1
min Dist : = dist (p, w, mean)
else
min Dist : = min {dist (p, q) : q ∈ tmpset}
if (min Dist > max Dist)
{ max Dist : = min Dist
Max point : = P
}
}

tmpset : = tmpset U {maxpoint}
}
```

```
For each point P in tempest do
w_rep := w_rep ∪ {p + α * (w-mean-p)}
return w'
end
```

7.3 Stream Computing

Q. What is stream computing?

- Stream computing is useful in real time system like count of items placed on a conveyor belt.
- 1BM announced stream computing system in 2007, which runs 800 microprocessors and it enables to software applications to get split to task and rearrange data into answer.
- AT1 technologies derives stream computing with Graphical Processors (GPUs) working with high performance with low latency CPU to resolve computational issues.
- AT1 preferred stream computing to run application on GPU instead of CPU.

Fig. 7.3.1 : Standard stream for input, output and error

7.3.1 A Stream - Clustering Algorithm

Q. Explain BDMO algorithm.

- BDMO Algorithm has complex structures and it is designed in approach to give guaranteed performance even in worst case.
- BDMO designed by B. Bahcock, M. Datar, R. Motwani and L. OCallaghan.

Details of BDMO algorithm

(i) Stream of data are initially partitioned and later summarized with help of bucket size and bucket is a power of two.

(ii) Bucket size has few restrictions size of buckets are one or two of each size within a limit. Required bucket may start with sized or twice to previous for example bucket size required are 3, 6, 12, 24, 48 and so on.

(iii) Bucket size are restrained in some scenario, buckets mostly O (log N).

(iv) Bucket consists with contents like size, timestamp, number of points in cluster, centriod etc.

Few well – known algorithm for data stream clustering are :

(a) Small – Spaces algorithm
(b) BIRCH
(c) COBWEB
(d) C2ICM

7.4 Initializing and Merging Buckets

Q. What is bucket, how it is used for clustering?

Q. Explain in brief initializing and merging of bucket.

- A small size 'p' is chosen for bucket where p is power of 2. Timestamp of this bucket belongs to a timestamp of most recent points of bucket.
- Clustering of these points done by specific strategy. Method preferred for clustering at initial stage provide the centriod or clustroids, it becomes record for each cluster.

  Let,

  * 'p' be smallest bucket size.

  * Every p point, creates a new bucket, where bucket is time stamped along with cluster points.

  * Any bucket older than N is dropped

  * If number of buckets are 3 of size p

  p → merge oldest two
- Then propagated merge may be like (2_p, 4_p, ...).
- While merging buckets a new bucked created by review of sequence of buckets.
- If any bucket with more timestamp than N time unit prior to current time, at such scenario nothing will be in window of the bucket such bucket will be dropped.
- If we created p bucket then two of three oldest bucket will get merged. The newly merged bucket size nearly z_p, as we needed to merge buckets with increasing sizes.
- To merge two consecutive buckets we need size of bucket twice than size of 2 buckets going to merge. Timestamp of newly merged bucket is most recent timestamp from 2 consecutive buckets. By computing few parameters decision of cluster merging is taken.
- Let, k-means Euclidean. A cluster represent with number of points (n) and centriod (c).

  Put p = k, or larger – k-means clustering while creating bucket

  To merge, $n = n_1 + n_2$, $c = \frac{n_1c_1 + n_2c_2}{n_1 + n_2}$
- Let, a non Euclidean, a cluster represented using clusteroid and CSD. To choose new clusteroid while merging, k-points furthest are selected from clusteroids.

  $CSD_m(P) = CSD_1(P) + N_2(d^2(P, c_1) + d_2(c_1, c_2)) + CSD_2(c_2)$

7.5 Answering Queries

- Given m, choose the smallest set of bucket such that it covers the most recent m points. At most 2m points.
- Bucket construction and solution generation are the two steps used for quarry rewriting in a shared – variable bucket algorithm, one of the efficient approaches for answering queries.

Review Questions

Q. 1 What is clustering algorithm ?

Q. 2 What is CURE ?

Q. 3 Write procedure of CURE cluster algorithm.

Q. 4 What is sampling ? Explain random sampling and partition sampling.

Q. 5 Write pseudo function of cluster algorithm.

Q. 6 Write procedure for merging cluster in CURE.

Q. 7 What is stream computing ?

Q. 8 What is stdin, stdout, stddir ?

Q. 9 Explain BDMO algorithm.

Q. 10 What is bucket, how it is used for clustering ?

Q. 11 Explain in brief initializing and merging of bucket.

CHAPTER 8

Module - 6

Link Analysis

Syllabus

PageRank Overview, Efficient computation of PageRank : PageRank Iteration Using MapReduce, Use of Combiners to Consolidate the Result Vector

8.1 Page Rank Definition

MU - May 17

Q. Explain Page Rank with Example. (May 17, 2 Marks)

Q. What is Page Rank ? Explain the Inverted Index.

- Google™ is one of the giants in Information Technology. The major product of the Google i.e. search engines, dominate the all other web services.
- Before Google's search engine there were many search engines available but the algorithm they exhibit is not up to the mark. These search engines were worked equivalent to a "web-crawler".
- Web-crawler is the web component whose responsibility is to identify, and list down the different terms found on every web page encountered by it.
- This listing of different terms will be stored inside the specialized data structure known as an **"inverted Index".**
- An inverted index data structure has listing of different non-redundant terms and it issues an individual pointer to all available sources to which given term is related.
- Fig. 8.1.1 shows inverted index functionality.

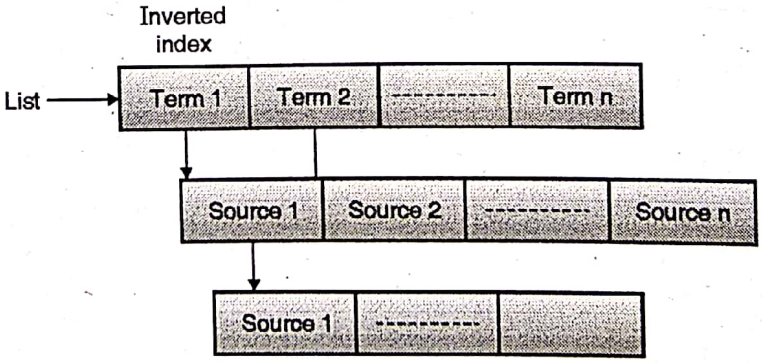


Fig. 8.1.1 : Inverted Index

- Every term from the inverted index will be extracted and analyzed for the usage of that term within the web page.

- Every term has some usage percentage within the given web page According to percentage of usage of terms in a web page it will be ranked. This will be achieved by firing a search query.

Example : If a given term 'x' is appeared in the Header of a web page then the page in which this term occurs proves to be much relevant rather than the term appeared in a paragraph text.

8.1.1 Importance of Page Ranks

Q. Explain Importance of Page Rank.

- In the Information Technology world, the storage and retrieval of the information becomes a crucial activity as Data generation from every sector increases exponentially.
- Every day if world will face new challenges for managing the different category web pages their arrangement by category, their ranking by search criteria etc.
- According to statistics in 1998 World Wide Web has around 150 million web pages and by today we have 1500 + million web pages.
- It is very much difficult to manage these huge number of web pages because every page contains following associated parameters :
  - (i) Number of terms involved
  - (ii) Category of web page
  - (iii) Topics involved in a given web page
  - (iv) Usage of involved topics by other web pages
  - (v) Quality of web pages etc.

Page Ranking : The term page Ranking can be defined as, "A classical method used to arrange the web pages according to its objective and the usage of terms involved in it on the world wide web by using any link data structure."

- The page Ranking mechanism was developed by Larry page and sergey Brin. This page-Ranking mechanism was a part of their research project which was started in 1995 and result into a functional prototype in 1998.
- After that, shortly they founded the Google.

8.1.2 Links in Page Ranking

MU - May 19

Q. List down the steps in modified Page Rank Algorithm to avoid spider trap with one example. **(May 19, 10 Marks)**

If we consider that, there are 150 million web pages exists in the some part of world wide web then all pages may have approximately 1.7 billion links to different web page.

Example

Suppose we have 3 pages A, B, C in a given domain of web sites. They have interconnection links between them as shown in Fig. 8.1.2

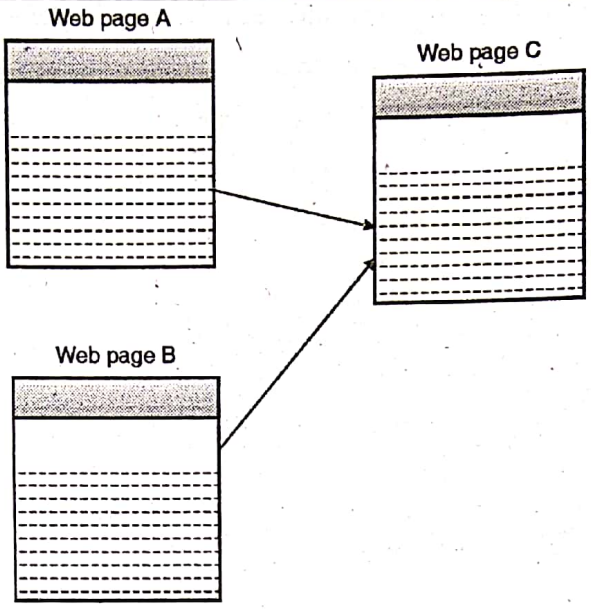


Fig. 8.1.2

The number of links exists between two or more web pages can be categorize as follows :

1. Back links
2. Forward links

1. Back links

With reference to Fig. 8.1.2 A and B are the Back links of web page 'C' i.e. Back link indicates given web page is referred by how many number of other web pages.

2. Forward link

- Forward link represents the fact that, how many web pages will be referred by a given web pages.
- Clearly, out of these two types of links back links are very important from Ranking of documents perspective.
- A web page which contains number of back links is said to be important web page and will get upper position in Ranking.
- A page Ranking in mathematical format can be represented as,

$$R(u) = c \sum_{V \in B_u} \frac{R(V)}{N_V}$$

Where,

μ : Represents the web page Bμ

N_V : It represents number of forward links of page v.

C : It represents the Normalizations factor to make

$$|R|_{L1} = 1 \; (|R|_{L1} = |R_1 + R_2 + \ldots R_n|)$$

- A world wide web can be considered as the 'Di-graph' i.e. Directed graph Any graph 'G' is composed of two fundamental components vertices and Edges.

G ≡ (V, E)
- Edges
- Vertices/nodes

Here, vertices or Nodes can be mapped to pages.

- If we consider a small part of world wide web containing 4 web pages named as P_1, P_2, P_3, P_4.
- Every page i has Back links and forward links to other pages.
- Fig. 8.1.3 shows the above mentioned structure.

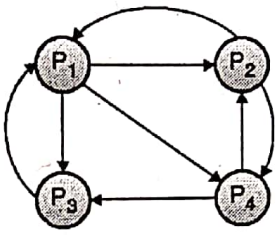


Fig. 8.1.3

- In Fig. 8.1.3 page P_1 has Forward links to page P_2, P_3 and P_4 respectively.
- Page P_2 has links to page 1 and page P_3.
- Page 3 has link to page 1 and
- Page 4 has links to page 2 and page 3.
- If a user starts surfing with page P_1 in above web page P_1 has links to page P_2, P_3 and P_4

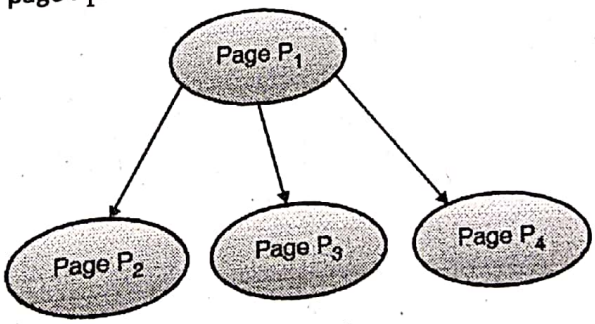


Fig. 8.1.4

- Probability that user will be at page P_2 / P_3 / P_4 is equal to 1/3.
- Probability that user will be at page 1 itself is '0'.
- Suppose user has chosen page P_2 then
- Probability that user will be at page P_1 is 1/2,
- Probability that user will be at page P_4 is 1/2.
- Probability that user will be at page P_2 or P_3 is 'D'.
- Probability that user will be at page P_2 or P_3 is 'D'.

These possibilities of web surfing by a given user can be represented using special structure known as "Transition Matrix".

In general, the transition Matrix 'M' is composed of 'n' pages 'n' rows and 'n' columns. Two pointer c and j will be to represent the current row and columns respectively.

Any given element can be represent as m_{ij}.

m_{ij} = 1/k if and only if the page at j^{th} column has k forward links.

- Additionally one of the forward links to same page itself.

∴ The transition Matrix for above web can be represented as,

$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \end{array}$$

Matrix should be seen column wise Example 2

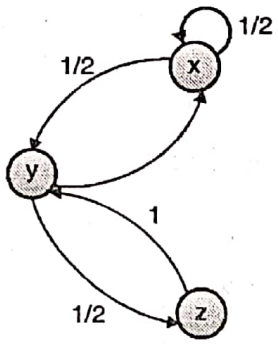


Fig. 8.1.5

∴ The transition Matrix for above graph can be represented as,

$$M = \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \end{array}$$

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$ ...For first iteration

$$\begin{bmatrix} 5/12 \\ 1/3 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix}$$ ...For second iteration

- Hence, with simplified page Rank algorithm a critical problem has evolved i.e. during each iteration, the loop accumulates the rank but never distributes rank to other pages.
- To identify the location at which the user will in near future one must have a probability with a specialized function known as "A page Rank".
- All Transition Matrixes will work on column vector or j^{th} component.
- Consider a user "xyz" want to surf the web consists of 'n' web pages. Every page in a web has equal probability that user will visit that page at next instance.

- Consider a vector V_0 as an initial vector component. The probability that the user will be at n^{th} page will be 1/n with same initial vector V_0.
- At next instance, user will at one of 'n' available pages. The probability distribution of this situation can be represented as, MV_0 on next instance it will be $M(M.V_0)$ and process continues.
- The probability that a user will be present at i^{th} node on given instance is equal to elements position into vectors value.

$$\text{Probability } (x_i) = \sum m_{ij} : V_j$$

Where,

(i) m_{ij} represents the probability of user movement at given instance from j^{th} location to i^{th} location.

(ii) V_j represents the probabilities that user is at j^{th} position for previous instance.

- Traditionally, this process is known as "Marker Principle"
- To have Markov distribution a system of graph should satisfy following constraints.

(i) Graph under consideration should be "strongly connected" i.e.

Every node is accessible other available nodes.

(ii) There should not be any dead ends.

8.1.3 Structure of the Web

Q. Explain Structure of Web? Explain Spider trap in detail.

- The web is nothing but the composition of number of individual independent nodes. The nodes can be also termed as workstations. We can imagine the web as set of different distributed systems together.
- In every distributed systems we have clusters, cluster means a group of similar objects for the clustering in the context of distributed systems, the term object can be replaced with node i.e. A cluster in a given distributed system is the collection of similar nodes.
- Additionally the term similarity can be determined by analyzing the concept such as :

(i) Nodes having same hardware configuration

(ii) Nodes having same operating system and other system and application software configuration.

- It is always recommended that, all the nodes in web should always be connected. This can be achieved theoretically but practically this is not the case always.
- Fig. 8.1.6 shows the part of World Wide Web where, every node is connected with other nodes.

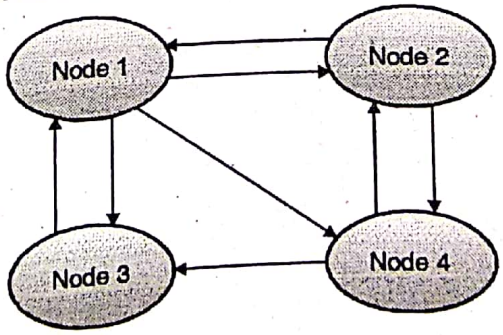


Fig. 8.1.6

– In practice any given web structure is composed of 4 types of components :

1. Strongly Connected Components (SCC)
2. In - components
3. Out components
4. Disconnected components

1. **A strongly connected components :** Is nothing but the components which are directly connected to each other for the data exchange and they also has forward and backward link to each other.

2. **In-components :** In-components are the integral part of where it exhibit the relation with SCC such that,

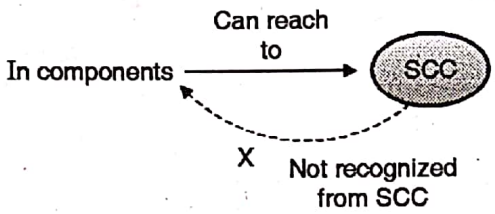


Fig. 8.1.7

3. **Out-components :** Out-components are the structures which shows following properties.

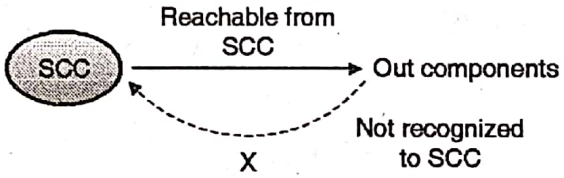


Fig. 8.1.8

The in-component and out components can have tendrils which represents in and out components.

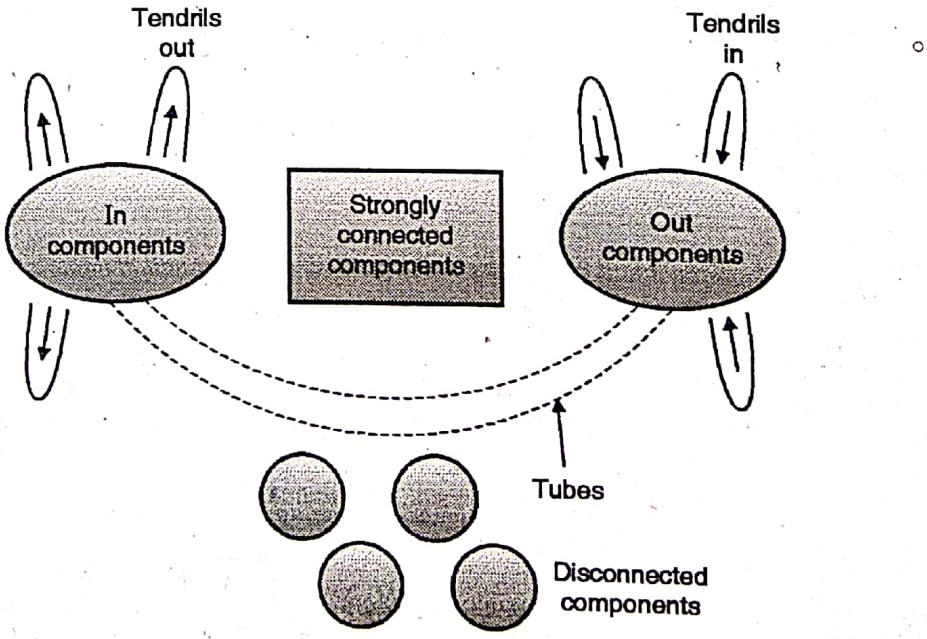


Fig. 8.1.9

In real time two major issues we will encountered :

(i) Dead end	(ii) Spider traps

(i) **Dead ends**

MU - Dec. 16

Q. Explain how dead ends are handled in Page Rank. (Dec. 16, 5 Marks)

- In a given part of web, if we encountered with a page which doesn't have links which are going out from that page or component.
- This will affect the transition matrix directly as, the column containing entry for dead end page will lead to sum = 0 instead of 1.
- The property of having sum = 1 for most of the columns in a given transition matrix is known as **'Stochasticity'** and if there are dead ends then some of the columns have '0' entries.
- Consider the Fig. 8.1.10.

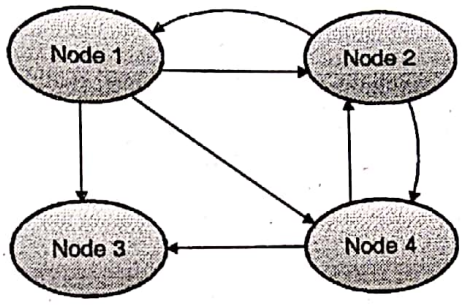


Fig. 8.1.10 : Node 3 is dead end

- By referring Fig. 8.1.10 we came to know that, Mode 3 is dead end and while searching if we encountered on the Node 3. i.e. at dead end then web surfing will struck at that page or node as there is not a single out link from Node 3. Hence, transition matrix for Fig. 8.1.10 is,

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \end{bmatrix}$$

Following are the ways to deal with dead ends :

1. **First approach to deal with dead ends we can delete that node by removing their incoming links.**
   - Disadvantage of this approach is it will introduce more dead ends which has to be solved with the same approach in recursive manner.
   - Though we delete the node but total page rank for a given graph to or web will be kept as it is the Nodes which are not available in graph G, but we can consider the set of other nodes which acts as predecessors for the calculation of page rank.
   - Additionally we can consider the successor nodes and have a division operation.
   - After this procedure, some nodes might be there which are not available in graph G but they are done with their predecessors calculations.
   - After some iterations all nodes has their page rank the order in which page ranks are calculated is exactly opposite to node deletion order.

- Suppose we have graph containing nodes and these nodes are arranged in following manner as shown in Fig. 8.1.11.

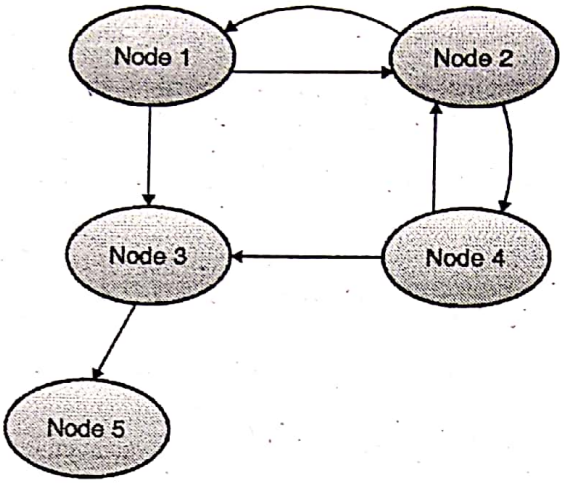


Fig. 8.1.11

- If we observe the Fig. 8.1.11 to calculate the page rank. We find that, Node 5 is the dead end as it doesn't have any forward links i.e. the links going out from Node 5.
- So hence, to avoid the dead ends, delete the Node 5 and its corresponding are coming from Node 3. So now the graph G becomes.

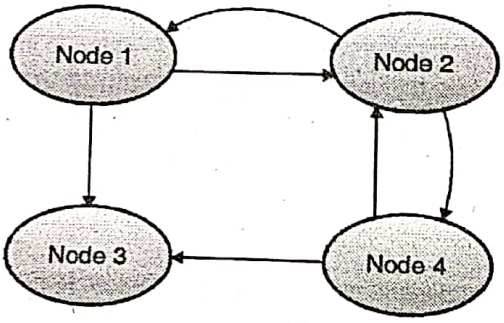


Fig. 8.1.12 : Graph a after deletion of node 5

- By observe the Fig. 8.1.12 we came to know that now 'node 3' is **'dead end'.**
- Now as we are avoiding the dead ends. Hence delete 'Node 3' and it is respective in coming edges. Now, Graph G becomes,

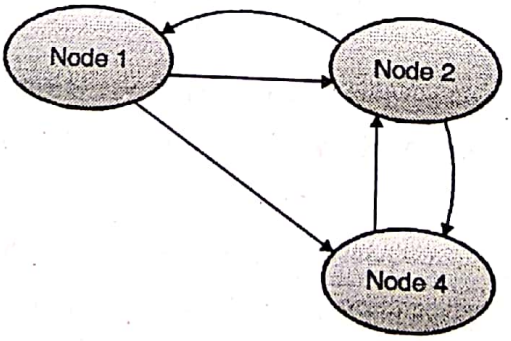


Fig. 8.1.13 : After deletion of node 3

- Now every node : Node 1, Node 2, and Node 3 has edges coming out i.e. forward links. Hence, is no dead end in graph 'G'.

The transition matrix for above graph will be,

$$M = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$

We can have component vector representation for above matrix as follows :

$$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \text{ – (1) Iteration 1}$$

$$\begin{bmatrix} 1/6 \\ 3/6 \\ 2/6 \end{bmatrix} \text{ – (2) Iteration 2}$$

$$\begin{bmatrix} 3/12 \\ 5/12 \\ 4/12 \end{bmatrix} \text{ – (3) Iteration 3}$$

Final value for component vector will be, $\begin{bmatrix} 2/9 \\ 4/9 \\ 3/9 \end{bmatrix}$

Page rank for node 1 $= \frac{2}{9}$

Page rank for node 2 $= \frac{4}{9}$

Page rank for node 4 $= \frac{3}{9}$

- We have to calculate the page rank for Node 3 and Node 5 with the exact opposite order of node deletion. Here Node 1, Node 2, Node 4 are in the role of predecessors.
- Number of successor to Node 1 = 3, Hence, the contribution from Node 1 for calculating the page rank of Node 3 is 1/3
- For Node 5 it has 2 successors. Hence the contribution from Node 5 for calculating the page rank of node 3 is 1/2.

$$\text{Page rank of Node 3} = \left(\frac{1}{3} \times \frac{2}{9}\right) + \left(\frac{1}{2} \times \frac{3}{9}\right) = \frac{13}{54}$$

- For calculating the page rank of Node 5, Node 3 plays a crucial role. As Node 3 has, number of successors = 1 and node 1 has node 3 as its predecessor. Hence, we can conclude that Node 5 was page rank same as that of Node 3.
- As the aggregate of their page rank is greater than 1, so it doesn't indicate the distribution for a given user who is surfing through that web page. Still it highlights the importance of web page relatively.

2. **Another way to deal with dead ends is configure the process for a given user by having assumption that it is assumed to be moved through web known as "Taxation".**

- Taxation method points to other problem also which is known as "spider traps".

(ii) Spider traps

- Spider traps is nothing but set of web pages all of them containing the out links but they never going to link with any other page.

i.e. Spider Trap = Set of web pages with no dead ends but no edge going outside also (no forward link)

- Spider traps can be sowed in the web with or without intention. There can be multiple spider traps in real time in a given web page. Set but for demonstration purpose, consider Fig. 8.1.14 which shows the part of web containing only one spider trap.

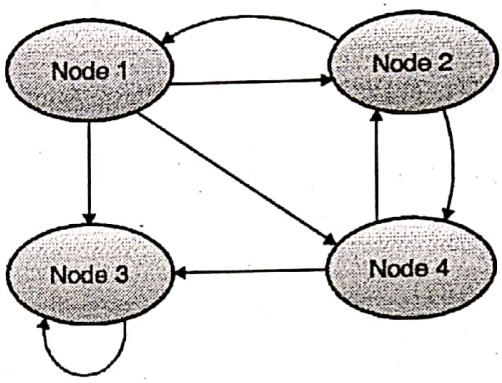


Fig. 8.1.14 : Graph with one spider - trap

The transition Matrix for the Fig. 8.1.14

$$M = \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

- If we proceed further by the same method stated in previous section for calculating the page rank then the ultimate result that we get will be,

$$\begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} - (1) \text{ Iteration (1)}$$

$$\begin{bmatrix} 3/24 \\ 5/24 \\ 11/24 \\ 5/24 \end{bmatrix} - (2) \text{ Iteration (2)}$$

$$\begin{bmatrix} 5/48 \\ 7/48 \\ 29/48 \\ 7/48 \end{bmatrix} - (3) \text{ Iteration (3)}$$

$$\begin{bmatrix} 21/288 \\ 31/288 \\ 205/288 \\ 31/288 \end{bmatrix} - (4) \quad \text{Iteration (4)}$$

:

:

At last iteration we get,

$$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

- The highest page rank will be given to Node 3 as, there is no link which goes out from it, but if has the link to inside it. So, user is going to struck at Node 3. As it is represented that number of user are there at Node 3 so Node 3 has greater importance.
- To have a remedy to this problem we just configure the method of calculating the page rank by injecting a new concept known as 'teleporting' or more specifically a probability distribution of "teleporting" and we are not following the links going out from given node.
- To calculate the teleporting probability – calculate new component vector V_{new} for estimating page ranks such that,

$$V_{new} = \beta M \cdot V + (1 - \beta) \cdot e/n$$

Where,

β = Constant (Ranges from 0.8 to 0.9)

e = Aggregate vector of all vectors with value 1

n = Number of web pages/ nodes in a graph

- If we do not dead ends in the graph then,

Probability of introduction of New user to a given web = Probability of not to choose out-link for the current page by same user

- Another possibility is user will not be able to move to any page as $(1 - \beta)$ e/n term is independent of $\sum V$.

i.e. when we don't encountered with dead ends then,

$\sum V < 1$ but $\sum V = 0$.

8.1.4 Using Page Rank in a Search Engine

- The page rank calculation plays crucial role in deciding the overall efficiency of the search engine.
- A basic crawling is done to have page ranking to fetch the required information and the page.
- When a user submit some query or a request to the given search engine then at background a secret algorithm is triggered for the execution which fetches different web pages in some order which is based on a pre defined criteria.
- The user query generally in the form of single word or collection of words.

- For example, the most popular search engine Google has 250+ such predefined criteria to arrange the fetched web pages in some particular order.
- Every page on the web should possess minimum one word or one phrase in it. Same as that of user's search query.
- If the given web page doesn't contain any word or phrase then there is less probability that a page will have the highest page rank.
- In page ranking calculation, the place on the web page where the phrase is appeared is also matters e.g. (phrase appears in the header will have more importance than in footer, the phrase appeared in paragraph will have average importance)

8.2 Efficient Computation of Page Rank

Q. What is thrashing? How it affect the Page ranking Mechanism?

- In previous sections we have studied that, how to calculate the page rank of the given web page in a given web structure.
- The efficiency in such complex calculation is achieved as we have taken a small part of web i.e. for 4-5 nodes or pages only.
- But, if scale-out this small for concept to a real-time condition billions of web pages V, a matrix – vector multiplication we have to compute of order atleast 70-80 times till a component vector will stop changing its value.
- For such real time complexity the solution proposed is use of Mapreduce technique studied in Section 3.2, but such usage is not that straight forward, it has two handles to cross.
  (i) The most important parameter that how to represent the transition matrix for such huge number of web pages. If we try to represent the matrix for all available web pages which are under consideration then it is absolutely in efficient for performing the calculations. One way to handle this situation is to indicate the non-zero elements only.
  (ii) One more thing is if we go for an alternative to mapreduce functionality for performance and efficiency concerns then we may think for 'combiners' explained in Section 3.2.4.
- The combiners generally used to minimize the data more specifically an intermediate data result to be transferred to reducer task.

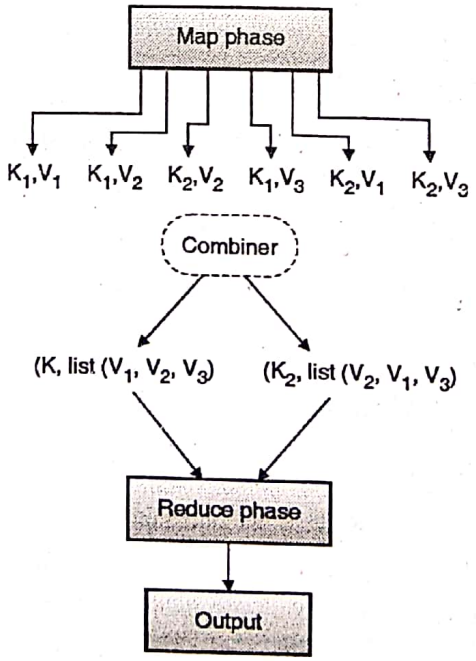


Fig. 8.2.1 : Use of combiners

- Also, the striping concept doesn't have much more impact to reduce the effect of "thrashing".

Thrashing

- All computations performed by the CPU itself irrespective of environment used i.e. Distributed Computing Environment (DCE) or standalone computing. Hence, the main task of processor / CPU is to execute the instructions and not the featuring of data from the secondary storage.
- If in some commonly occurred situation processor /CPU is busy in just fetching the data from secondary storage rather than executing the instruction then such situation is known as "Thrashing".

8.2.1 Representation of Transition Matrix

- As we know that, number of web pages that we are going to deal with are billions and number of links going out from a given web pages are 10 on an average.
- Entry '1' in billion pages is not zero. The best way to indicate the transition matrix is to have a list of different web page which has entries 'non-zero' with associated values.

The structure will look like,

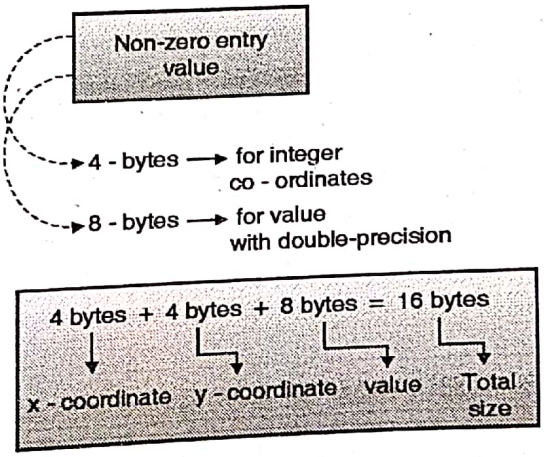


Fig. 8.2.2

- So, space required here has linear nature instead of quadratic.
- We can apply more compression by column wise representation of non zero entries i.e. 1/number of links going out from a given web page

A column is represented as ,

1 integer → to represent → out degree

1 integer → to represent → for every non-zero entry in that column

↓

yields a row number of entry location

8.2.2 Iterating Page Rank with MapReduce

Q. Explain Iterating Page Rank Process with MapReduce? Also comment on use of Combiners.

– A single pass of page rank calculation includes calculation two component vectors depicted by V and V_{new}.

$$V_{new} = \beta \cdot M \cdot V + (1 - \beta) \cdot e/n$$

Where, β = Constant (Ranges between 0.8 to 0.9)

e = components vector of entries 1

n = Number of web pages

M = Transition matrix

When 'n' has small value then V and V_{new} can be stored in primary memory or main memory for Map task.

– If in real – time V is big in size so that if can't be fit into main memory then we can go for striping method.

8.2.3 Use of Combiners to Aggregate the Result Vector

– The page ranking iteration with MapReduce task is not proven to be sufficient because

(i) If we want to add different terms V_{new} i.e. the i^{th} element of new resultant vector V, computed at Map phase

This is equivalent to the usage of special structure "combiner" which actually combines the different values according to their key as shown in the Fig 8.2.3.

(ii) If we are not going to use MapReduce at all.

– Hence depending on the requirement situation and complexity of problem a method to be used should be decided.

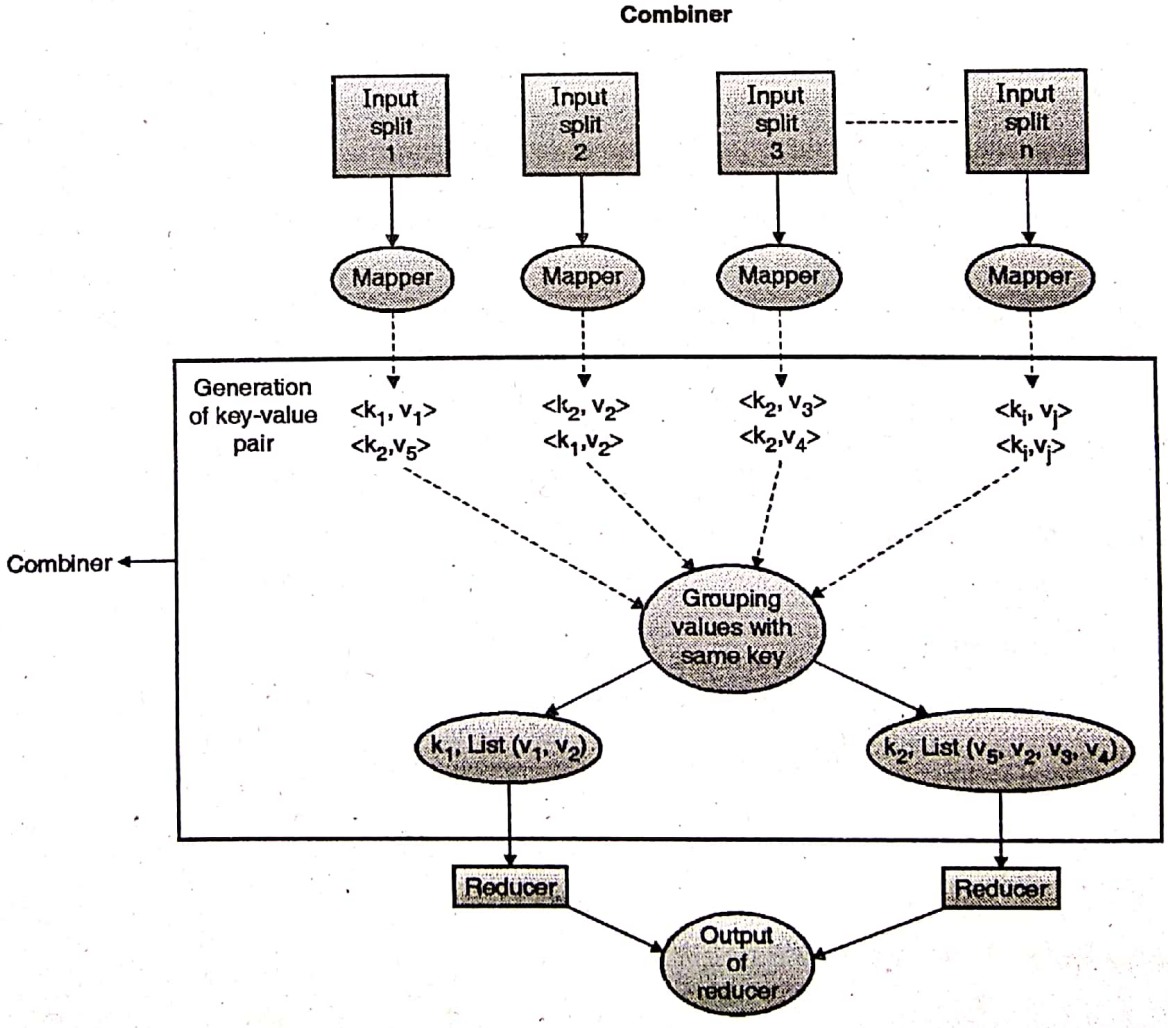


Fig. 8.2.3 : Combiner working mechanism

8.3 Link Spam

Q. What is Link Spam? Explain in Detail.

- When page rank question arises then the giants in information technology such as google will develop some solution to it. Additionally there are some security related issues such as "The spams".
- But we do have destructive minded people in society who will always try to affect the system by performing some malicious activities. Hence for page ranking calculations "spammers" are came into existence.
- Spammers have introduce the tools and techniques through which for any given page its page rank can be increase by a selected multiple such intentional rise in the value of page rank is known as a "link spam".
- For link spam spammers introduce the web pages itself for link spamming.

8.3.1 Spam Farm Architecture

Q. Explain Spam Farm Architecture in detail.

- The malicious web pages introduced by the spammers is known as spam farm. Fig. 8.3.1 shows the basic architecture of spam farm.

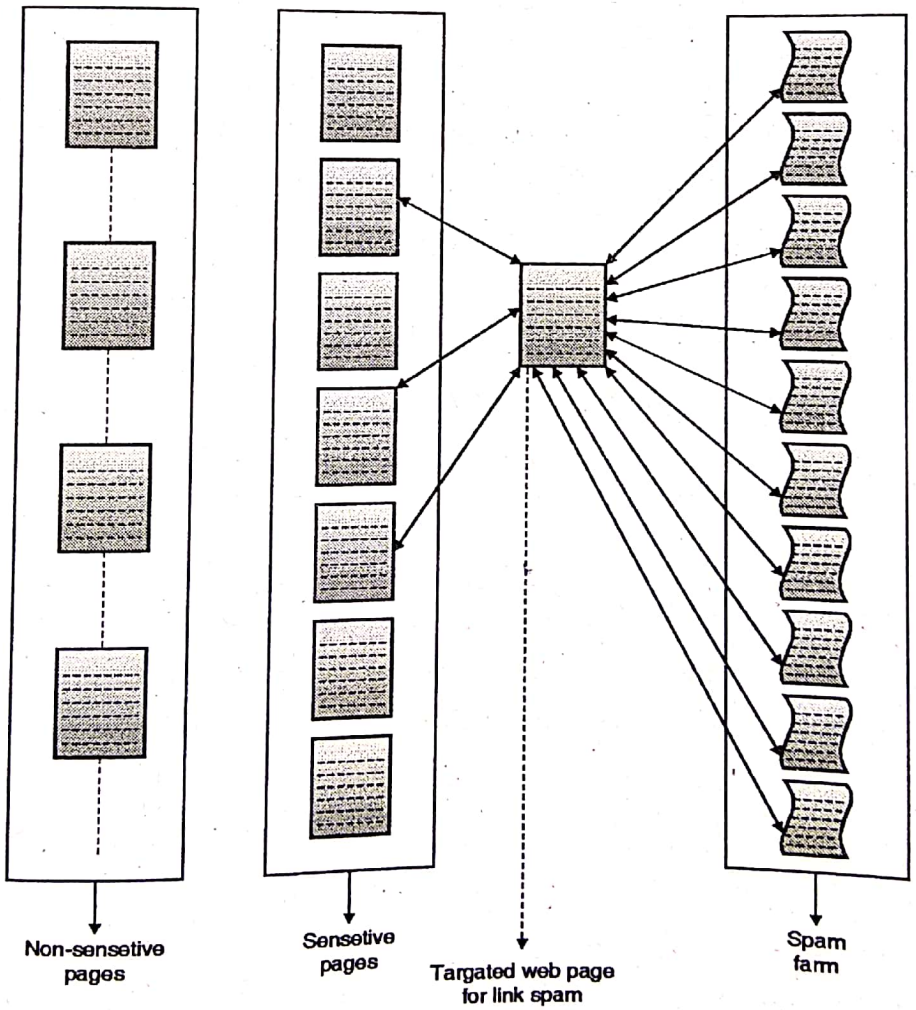


Fig. 8.3.1 : Basic architecture of spam farm

- If we consider the spammers perspective then Fig. 8.3.1 can be divided into 3 basic blocks

1.	Non-sensitive	2.	Sensitive
3.	Spam farm		

1. Non-sensitive

- These are the pages which are generally not accessible to the spammer for any spamming activity. As these pages are not accessible to spammers so, they will not affect by any activity performed by the spammer.
- Most of web pages in a given web structure will fall in this category.

2. Sensitive

- These are the web pages which are generally accessible to the spammers for any spamming related activity. As these pages are accessible to the spammers so, they will get affected easily by spamming activity performed by the spammer.
- The effect of spamming on these pages is generally indirect as these pages are not manipulated by the spammers directly.

3. Spam farm

The spam farm is the collection of malicious web pages which are used to increase the number of links pointed to and coming out from a given web page so, ultimately the page rank of a given page i.e. target web page will increase dramatically. There are other category web pages which supports the spamming activity by aggregating page ranking i.e. a part of term (1 - β).

8.3.2 Spam Farm Analysis

Q. Explain the Spam Farm Analysis in detail.

- In spamming activity we do page ranking calculation by considering sensitive, non sensitive and actual malicious spam farm pages.
- The basic page ranking calculation is done and by alternate method 'β' term is calculated which is also known as Taxation method.
- The term 'β' will depicts the fact that, how a part of page rank is segregated among the successor nodes for the next iteration. Actually β is the constant term ranges between 0.8 to 0.9 generally (0.85).
- We know that, there are some web pages who supports the spamming activity.
- So a page rank of one of such supporting page can be calculated with the help of following formula

$$Pr(s) = \beta \cdot Y / m + (1 - \beta) / n$$

Where,

Pr (s) → It represent the page rank of randomly selected supporting web page.

β → It is the constant range from 0.8 to 0.9.

Y → It represent page rank of target web page say 't'.

m → It represent number of web pages supporting spamming activity.

n → it represent total number of web pages in a given web structure.

The term β, Y is totally related to the target web page 't' calculated page rank will get segregated on all m's.

Hence a page rank represented by 'Y' for a given target web page 't' will composed of 3 incoming types :

(i) The number of links which are pointed to the target web page from the non-sensitive web pages represented by variable 'x'.

(ii) Page rank of supporting page with β multiples.

$\beta \cdot (Pr(s)$

Where, $Pr(d) = \beta \cdot Y / m\,t\,(1-\beta)/n$

∴ We can conclude that the page rank 'Y' of target web page 't' will be in the form of

$$Y = x + \beta \cdot m\left(\frac{\beta y}{m} + \frac{1-\beta}{n}\right)$$

$$= x + \beta^2 y + \beta \cdot (1-\beta) \times \frac{m}{n}$$

Here we can introduce a constant Q,

$$\therefore Y = \frac{x}{1-\beta^2} + Q \cdot \frac{m}{n}$$

$$Q = \frac{\beta(1-\beta)}{(1-\beta)^2}$$

$$= \frac{\beta}{(1+\beta)}$$

8.3.3 Dealing with Link Spam

- As we have seen the effect of link spam on the fundamental primary things related to page rank system.
- Link spam will disturb the page rank system completely hence to deal with the link spamming the different search engine thought of different solution, which will help in minimizing the effect of link spam.
- Basically there are two ways to deal with link spamming they are as follows :

(i) A traditional approach where the search engine algorithm will have top-view of whole scenario and eventually algorithm win find such link spams and remove them from the indexing structure.

But as soon as algorithm deletes the spammer web page the spammer will find the alternate way to do the link spamming.

(ii) A modern way to deal with the situation is, modify the procedure for calculation of page rank with reference to below grade link spams.

We will have two alternate procedures to do so :

(i) Trust ranking
(ii) Spam mass

(i) Trust ranking

- In trust ranking the system is going to trust on some of the web pages by assuming that those web pages are not the part of spam farm.

- Such set of web pages is termed as "topic".
- Consider a spam farm page want to increase a page rank of trusted web page. So, spam page can have a link to trusted page but that trusted page will not established a link to spam page.

(ii) Spam mass

- In spam as technique, the algorithm of page ranking will calculate the page rank for every web page also the part of the page rank (affected part) whose contributor is spam page will be analysed. This analysis is done with the help of comparison between normal page rank and page ranking obtained through trust ranking mechanism.
- This comparison can be achieved through following formula :

$$Pr\,(Sm) = \frac{Pr - Prtr}{Pr}$$

Where, $P_r(s_m)$ = page ranking by spam mass technique

P_r = page ranking by traditional method

P_rt_r = page ranking by trust ranking method

If $P_r(s_m) < 0$ i.e. negative

Or

$Pr\,(s_m) > 0$ but < 1 i.e. not close to 1 then that page is not a spam page else it is a spam page.

8.4 Hubs and Authorities

MU - May 17

Q. Explain Hubs and Authorities with neat diagram. (May 17, 5 Marks)

Q. What is Hubs and Authorities? Explain its significance.

- The hubs and authorities is an extension to the concept of page raking. Hubs and authorities will add more preciseness to the existing page rank mechanism.
- The ordinary, traditional page rank algorithm will calculate the page rank for all the web pages available in a given web structure. But user doesn't want to examine or view all of these web pages. He/she just want first 20 to 50 pages in an average case.
- Hence, the idea of hubs-and authorities will came into existence to have efficiency and reduce work load calculating page rank.
- In hubs and authorities page rank will be calculated for only those web pages who will fetch in resultant set of web pages for a given search query.
- It is also known as, hyperlink induced topic search abbreviated as HITS.
- The traditional page rank calculations have a single view for a given web page. But hubs and authorities algorithm will have two different shades of views for a given web page.

1. Some web page has importance as they will present signification information of given topic so these web pages are known as the authorities.
2. Some web pages has importance because they gives us the information of any randomly selected topic as well as they will direct us to other web pages to collect more information about the same. Such web pages known as hubs.

8.4.1 Formalizing Hubs and Authority

- As stated in earlier section hubs and authorities these are the two shades with which a web page can be viewed.
- So, we can allot 2 types of scores for a given web page.

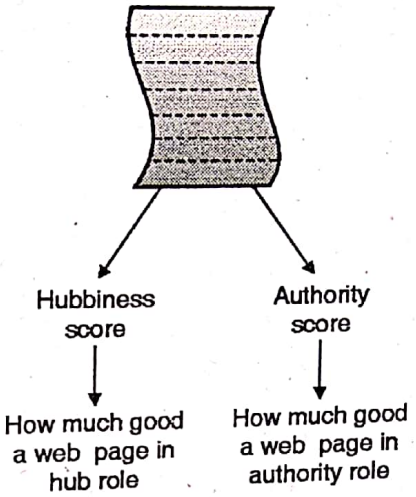


Fig. 8.4.1

h → represents hubbiness score

a → represents authority score

- j^{th} component of a web pages 'h' will give measure of Hubbiness of j^{th} page.
- j^{th} component of a web page 'a' will give measure of authority of j^{th} page.
- To have the notion of 'h' and 'a' consider link matrix 'LM' for web pages in a given web.
- Any element of LM can be represent as LMij

$\therefore$ LMij = 1 if a link is established from i^{th} page to j^{th} page.

LMij = 0 otherwise

The transpose of LM will change the result :

$\therefore$ LM_{ij}^{T} represent transpose of LMij

$\therefore$ LM_{ij}^{T} = 1 if a link is established from j^{th} page to i^{th} page

$\therefore$ LM_{ij}^{T} = 0 otherwise

- More that $LM^{T} = M$ where is a original transition matrix which maintain the record of No. of incoming and outgoing links.
- The difference between LM^{T} and M is,

$LM^{T} = 1$

and $M = \frac{1}{\text{No. of links going outside}}$ for that web page forgiven column.

Ex. 8.4.1 : Let the adjacency matrix for a graph of four vertices (n1 to n4) be as follows :

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Calculate the authority and hub scores for this graph using the HITS algorithm with k = 6, and identify the best authority and hub nodes.

MU - May 16. 10 Marks

Soln. :

$$\text{Given Matrix is} - \quad A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Apply Transpose operation,

$$A^T = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Now, Consider initial Hub score as 1

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{(1^2+1^2+2^2+4^2)}} \\ \frac{1}{\sqrt{(1^2+1^2+2^2+4^2)}} \\ \frac{2}{\sqrt{(1^2+1^2+2^2+4^2)}} \\ \frac{4}{\sqrt{(1^2+1^2+2^2+4^2)}} \end{bmatrix}$$

$$= \begin{bmatrix} 0.2132 \\ 0.2132 \\ 0.4264 \\ 0.8528 \end{bmatrix}$$

Iterate over K.

Review Questions

Q. 1 What is Page Rank ? Explain the Inverted Index ?

Q. 2 What is Page Rank? Explain Importance of Page Rank?

Q. 3 What are Links in Page Ranking? Explain in Detail.

Q. 4 What are links in Page Ranking? Explain Back Links and Forward Links with suitable example?

Q. 5 Explain the Structure of Web in the context of Link Analysis?

Q. 6 Explain Structure of Web? What is the significance of In-Components and Out Components?

Q.7 Discuss in detail Structure of web? Explain the Dead ends

Q.8 Explain Structure of Web? Explain Spider trap in detail.

Q.9 Explain the role of Page ranking in search engine?

Q.10 Explain the different modification suggested in efficient computation of Web pages.

Q.11 What is thrashing? How it affect the Page ranking Mechanism?

Q.12 Explain Iterating Page Rank Process with MapReduce? Also comment on use of Combiners.

Q.13 What is Link Spam? Explain in Detail.

Q.14 Explain Spam Farm Architecture in detail.

Q.15 What is Spam Farm explain with neat diagram? Also comment on Non-Sensetive, sensitive and spam farm.

Q.16 Explain the Spam Farm Analysis in detail.

Q.17 What is Link Spam? How to deal with Link Spam with Trust Ranking and Spam Mass.

Q.18 What is Hubs and Authorities? Explain its Significance.

□□□

CHAPTER 9

Module - 6

Recommendation Systems

Syllabus

A Model for Recommendation Systems, Content-Based Recommendations, Collaborative Filtering

9.1 Recommendation System

MU - Dec. 17

Q. What are recommendation systems? Clearly explain two applications for recommendation system. **(Dec. 17, 10 Marks)**

Q. What is recommendation system?

- It is vast widely used now-a-days. It is likely a subclass of information filtering system. It is used to give recommendations for books, games, news, movies, music, research articles, social tags etc.
- It is also useful for experts, financial services, life insurance, and social media like Twitter etc.
- Collaborative filtering and content-based filtering are the two approach used by recommendation system.
- Collaborative filtering uses user's past behaviour and apply some predication about user may like and accordingly post data.
- Content based filtering uses user's similar properties of data preferred by user.
- By using collaborative filtering and content based filtering a combine approach is developed i.e. Hybrid recommendation system.

9.1.1 The Utility Matrix

Q. Explain utility matrix with example.

- A recommendation system prefers the preference of a utility matrix. Users and item's these are entities used by recommendation system.
- Users have preference to data and these preferences must be observed.
- Every data itself is part of utility matrix as it belongs to some item category.

Example : A table representing users rating of apps on a scale 1 to 5, with 5 as highest rating Blank represents that user not replied on scale A1, A2 and A3 for Android 1, 2 and 3 i1, i2, i3 for iOS 1, 2, 3 users A, B and C gives rating.

	A1	A2	A3	i1	i2	i3
A	3		4		5	4
B		5				
C	3		4	4		4

Fig. 9.1.1 : A utility matrix representing ratings of apps on a scale 1to 5

- A typical users rating are very minute fraction of real scenario if we consider actual number of application of Android and iOS platform and number of users.
- It is observed in table for some apps there is less number of responses.
- The goal behind utility matrix is to make some predictions for blank spaces, these prediction are useful in recommendation system.
- As 'A' user gives rating 5 to i2 app so we have to take in account parameters of app i2 like its GUI, memory consumption, usability, music/effects if applicable etc.
- Similarly 'B' user gives rating 5 to A2 app so we have to take similar parameter in consideration. By judging both apps i2, A2 features and all we can put predication what can be further recommended to user A and B.
- From user "c" response though there is no use of full rating anywhere still it can be judge and predicted what kind feature based app user 'c' should be recommended.

9.1.2 Applications of Recommendation Systems

- Amazon. Com
- CDNOW. Com
- Quikr.com
- olx.com
- Drugstore.com
- eBay.com
- Moviefinder.com
- Reel.com and so many online good seller/buyer, trading website uses recommendation system.
- Product recommendation, Movie Recommendation, News Articles etc. are likely to be consolidate in a single place applications.

9.1.3 Taxonomy for Application Recommendation System

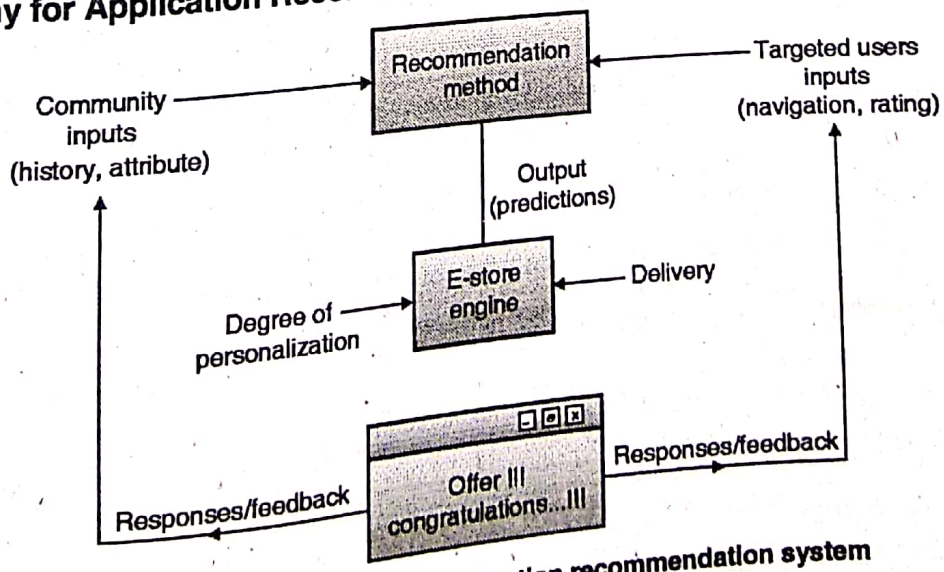


Fig. 9.1.2 : A taxonomy for application recommendation system

9.2 Content Based Recommendation

MU - May 19

Q. How recommendation is done based on properties of product? Explain with suitable example. (May 19, 10 Marks)

Q. Explain item profile of content based recommendation.

It focuses on items and user profiles in form of weighted lists. Profile are helpful to discover properties of items.

9.2.1 Item Profile

- An actor of drama or of movie is considered as an actor set, few viewers prefer drama or movie by their favourite actor(s).
- A set of teachers, some students prefer to be guided by few teacher(s) only.
- The year in which songs album release or made. Few viewers prefer old songs; some prefer to latest songs only, users sorting of songs based on year.
- So many classes are available which provides some data.
- Few domains has common feature for example a college and movie it has students, professors set and actors, directors set respectively. Certain ratio is maintained as many student and few professors in quantity while many actor works under one or two director guidance. Again every college and movie has year wise datasets as movie released in a year by director and actor and college has passing student every year etc.
- Music (song album) and a book has same value feature like songs writer/poet, year of release and author, publication year respectively.

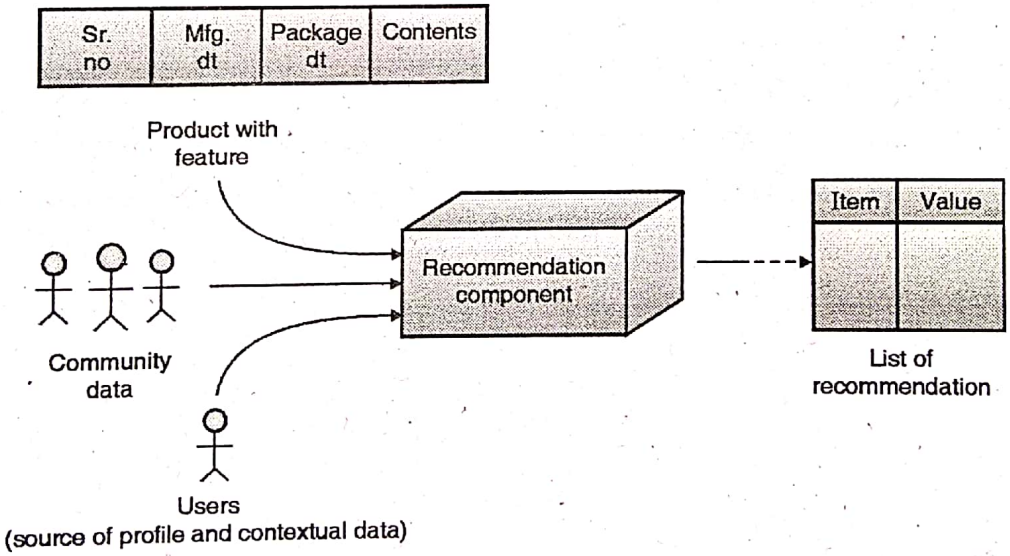


Fig. 9.2.1 : Recommendation system parameters

9.2.2 Discovering Features of Documents

MU - May 17

Q. How would you get the features of the document in a content -based system? (May 17, 3 Marks)

- Document collection and images are the two classes of items.
- We need to extract features from documents and images.

- There many kinds of document. Let say news articles in a newspaper. There are many articles in a newspaper but user reads very few of them. A recommendation system suggest for articles to a user supposed to be interested to read.
- Similarly there are so many websites and web pages, blogs could be recommended to a group of interested users.
- It will be more friendly to users if we can classify blogs accordingly to the topics.
- Unfortunately, document classes cannot provide available information features.
- A substitute is used for word identification on which characterize topic of document.
- We need to sort document by removing repeatedly used common words i.e. elimination of stop words. The words remain after elimination of stop words are proceed further to count their TF i.e. term frequency.
- IDF i.e. inverse Document frequency of each word in a document is observed and calculated.
- The high scoring words are the group of character which characterize document.
- Let, n words found in document with highest TF and IDF values of count. Then those n words are fixed to all similar documents.
- Words whose TF, IDF values of count founds threshold becomes a part of feature set.
- Such few words set represent a document. To know similarity between any two documents, there distance need to be measured between sets it can be done by either

(a) Jaccard Distance (b) Cosine Distance

9.2.3 Obtaining Item Features from Tags

- In case of published book database many features are available like Title, ISBN, Edition, Printing, Price, Compositor, Editor, Copyright etc., but user (reader) of book mainly concern with title and price tag mostly.
- We can get number of features of same items by tag items by entering phrase or search value range for price.
- By keeping tag option available, users can search at item on feature of tag value like price range for book, colour shade for cloth etc in online shopping.
- One problem in tag is to create tags and such enough tag awareness at user level.

9.2.4 Representing Item Profile

Q. How item profile is represented ?

- Some features are numerical for instance we might take the rating for mobile application based on Android OS platform to be a feature.
- This rating is a real number like 1, 2, 3, 4 and 5 stars by some users of it.
- By keeping options of only one component like 1 star or 5 stars will not make a sense to get possible average rating of
- By keeping 5 options in rating a good possible average for rating can be observed without lose of structure implicit in an application.
- By keeping 5 options in rating a good possible average for rating can be observed without lose of structure implicit in numbers.
- Numerical value based rating provides a single component of vector representing to items.
- Boolean value and other integer valued or real-valued parameter becomes components of vectors. Numerical features tell about similarity of items.

9.2.5 User Profiles

Q. What is a user profile in content based recommendation?

- Vectors are useful to describe items and user's preferences. Users and items relation can be plotted with the help of utility matrix.

  Example : Consider similar case like before but utility matrix has some nonblank entries that are rating in 1-5 range. Consider, user U gives responses with average rate of 3 there are three applications (Android OS based games) got ratings of 3, 4 and 5. Then user profile of U, the component for application will have value i.e. rated average of 3-3, 4-3 and 5-3 i.e. value of 1

- On other hand, user v gives average rating 4. So user v responses to application are 3, 5 and 2.
- The user profile for v has in the component for application, the average of 3-4, 5-4 and 2-4, i.e. value – 2/3.

9.2.6 Recommending Items to Users based on Content

- Between user's vector and item's vector cosine distance can be computed with help of profile vectors for users and items both.
- It is helpful to estimate degree to which user will prefer as an item (i.e. prediction for recommendation).
- If user's and response (like 1 to 5 scale for mobile apps) vectors cosine angle is large positive fraction. It means angle is close to 0 and hence there is very small consine distance between vectors.
- If user's and responses vector cosine angle is large negative fraction. It means angle is close to degree of 180 which is a maximum possible cosine distance.
- Cosine similarity function = for measuring cosine angle between two vector

$$\text{Cos}\,\theta = \frac{V_1 . V_2}{\| V_1 \| V_2 \|}$$

- In vector space model

$$V_d = [W_1 d, W_2 d, \ldots WNd]^T$$

Where,

$W_{t,d}$: TF* IDF weight of term 't' in 'd' document

TF : Term Frequency

IDF : inverse document frequency

9.2.7 Classification Algorithm

- It is used to know user's interest. By applying some function to new item we get some probability which user may like.
- Numeric values also help to know about degree of interest with some particular item.
- Few of techniques are listed as follows :

  (1) Decision Tree and Rule Induction

  (2) Nearest Neighbour Method

(3) Euclidean Distance Metric

(4) Cosine Similarity Function

Some other classification algorithm are :

(1) Relevance feedback and Rocchio's algorithm

(2) Linear classification

(3) Probabilistic methods

(4) Naive Bayes.

9.3 Collaborative Filtering

MU - Dec. 16, May 17

Q. Explain with example collaborative and content based filtering in a recommendation system. (Dec. 16, 10 Marks)

Q. Explain Collaborative Filtering based recommendation System. How it is different from content based recommendation systems ? (May 17, 10 Marks)

- Recommendation system in collaborative filtering becoming interesting as few domains are used move by research scholar and academician like human-computer interaction, information retrieval system and machine learning.
- Few famous recommender systems in some popular fields like Ringo-music, Bellcore-video recommender (movies), Jester-jokes etc.
- Collaborative filtering began to use in the early 1990s. Most widely used example of collaborative filtering and recommendation system is Amazon. com.
- To recommend among large set of values to users is very important. Recommendation must be get appreciated by user else effort taken for it were worthless.
- Netflix has 17,000 movies collection while Amozon.com has 4,10,000 title in its collection, so got proper selection of recommendation is necessary.
- Toolbox used for collaborative filtering becomes advanced with help of Bayesian interface, case-based reasoning method, information retrieval.
- Collaborating filtering deals with 'users' and items. A preference given by any user to an item is known as 'rating' and is represented by triplet value set of (User, Item, and Rating).
- Rating triplet of (users, items, rating) is used to create a sparx matrix and it is referred as rating matrix.
- 'Predict task' and 'recommend task' are used for evaluation and use of recommendation system.

Table 9.3.1 : Sample rating matrix con 5 star scales to apps

Apps(items) / Users	Whatsapp	Hangout	Telegram	Viber
User A	4	2	3	3
User B	3	3	5	3
User C	3	2	4	2

- Predict task tells about preference may given by a user or what user's likely preference to an item?
- Recommend task helpful to design n-items list for user's need. These n-items are not on basis of prediction preference because criteria to create recommendation may be different.

9.3.1 Measuring Similarity

Q. What is measuring similarity in collaborative filtering?

- Among values of utility matrix it is really a big question to measure similarity of items of users.

Table 9.3.2 : Utility matrix

Users \ Apps(items)	Whatsapp	Hangout	Telegram	Viber	Skype	Hike
User A	4			5	1	
User B	5	5	4			5
User C				2	4	
User D		3				

- Above utility matrix data is quite insufficient to put reliable conclusion. By considering values from A and C, they rated two apps in common but their ratings are diametrically very opposite.

9.3.2 Jaccard Distance

- In this sets of items rated are considered while values in matrix are ignored.

$$d_J(A,B) = 1 - J(A,B)$$
$$= \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

- Alternatively Jacard distance can be given by,

$$A \Delta B = (A \cup B) - (A \cap B)$$

- For example, user A and User B have an intersection of size 1 and a union size is of 5, its Jaccord similarity be 1/5 and Jacard distance be 4/5.
- User A and User C have Jaccard similarity 2/4 and Jaccard distance is same i.e. ½ .
- So, comparatively A and C are closer than A and B.
- User A and User C has very less matching choice of apps but user A and user B both rated nearly similar to one app i.e. Whatsapp.

9.3.3 Cosine Distance

- If user doesn't give any rating from 1 to 5 to any app then it is considered as a 0 (zero)

- Cosine angle between User A and User B is,

$$\frac{4 \times 5}{\sqrt{4^2+5^2+1^2}\sqrt{5^2+5^2+4^2}} = 0.380$$

- Cosine angle between user A and user C is,

$$\frac{5 \times 2 + 1 \times 4}{\sqrt{4^2+5^2+1^2}\sqrt{2^2+4^2+5^2}} = 0.322$$

- A is closer to B compare to C as longer cosine value implies a smaller angle.

9.3.4 Rounding the Data

- Rounding data by assigning one value to higher rating and assign NULL to lower values.
- For example, in our utility matrix few apps having ratings like 3, 4 and 5 will consider it as "1" and those having ratings like 2 and 1 will consider it as unrated keep it NULL.
- This approach will gives correct conclusion as Jaccard distance between A and B is 3/4 and A and C its 1, since C appears further from A compared to B.
- This approach will give correct conclusion and it can be verified by applying cosine distance that matrix will be.

Table 9.3.3 : Utility matrix values (ratings) 3, 4, 5 are replace by 1 and 2 and 1 value (ratings) are kept unrated (NULL)

Apps(items) / Users	WhatsApp	Hangout	Telegram	Viber	Skype	Hike
User A	1			1		
User B	1	1	1			
User C					1	1
User D		1				

9.3.5 Normalizing Rating

Low rating get convert into negative while high rating get converted into positive as it is subtracted from average rating, this is known as Rating Normalization.

9.4 Pros and Cons in Recommendation System

Q. Write any two pros and cons for Collaborative filtering and content-based filtering.

9.4.1 Collaborative Filtering

Pros

i) No knowledge engineering efforts needed.

ii) Serendipity in results.

iii) Continuous learning for market process.

Cons

(i) Rating feedback is required.

(ii) New items and users faces to cold start.

9.4.2 Content-based Filtering

Pros

(i) No. community requirement.

(ii) Items can be compared among themselves.

Cons

(i) Need of content description.

(ii) New users face cold start.

Review Questions

Q. 1 What is recommendation system?

Q. 2 Enlist application of recommendation system and taxonomy for application recommendation system.

Q. 3 Explain utility matrix with example.

Q. 4 Explain item profile of content based recommendation.

Q. 5 How item profile is represented ?

Q. 6 What is a user profile in content based recommendation ?

Q. 7 Explain in collaborative filtering.

Q. 8 What is measuring similarity in collaborative filtering ?

Q. 9 What is Jaccard distance and cosine distance in collaborative filtering ?

Q. 10 Explain rounding the data and normalized rating.

Q. 11 Write any two pros and cons for Collaborative filtering and content-based filtering.

□□□

CHAPTER 10

Module - 6

Mining Social Network Graph

Syllabus

Social Networks as Graphs, Clustering of Social-Network Graphs, Direct Discovery of Communities in a social graph

10.1 Introduction

Q. What is sociogram and Barbasi-Albert algorithm ?

- Social network idea came into theory and research in 1980s by Ferdinand Tonnis and Emile Durkhiem. Social network is bind with domain like social links, social group.
- Major work started in 1930s in various areas like mathematics, anthropology, psychology etc. I.L. Moreno provides foundation for social network as provided a Moreno's sociogram which represent social links related with a person.
- Moreno's sociogram example : Name the girl with whom you would like to go to industrial visit tour.

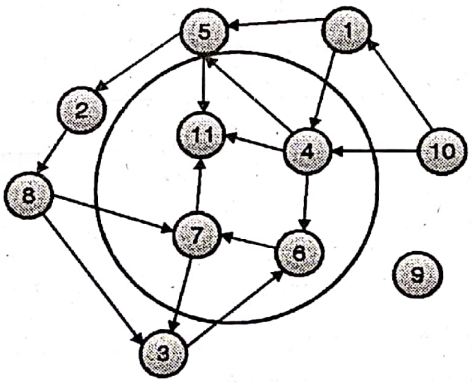


Fig. 10.1.1 : Moreno's sociogram with industrial visit tour

- Sociogram gives interpersonal relationship among members participated in group. Sociogram present choice in number.

$$C = \frac{\text{Number of mutual choices}}{\text{Number of possible mutual choices in the group}}$$

- The Barabasi-Albert (BA) model provides a network which initially connected with M_o nodes of network.

$$P_i = \frac{k_i}{\sum_j k_j}$$

Where, K_i - Degree of node i

j - All per existing node

- The new nodes gives preference to get attach with heavily linked nodes.

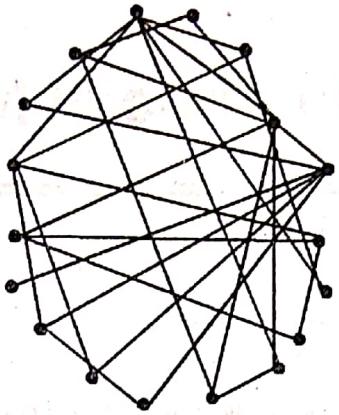


Fig. 10.1.2 : Barabasi algorithm model shows steps of growth of network ($M_0 = M = 2$)

- BA model is used to generate random scale free network. Scale-free network used in most of popular domain like the internet, World Wide Web, citation network and few social networks.
- Social network deals with large-scale data. After analyzing large data a huge set of information can be achieved.
- Linkedin, Facebook are vast widely used and very popular examples for social network. As we can find friends over the network with 1st, 2nd, 3rd connection or mutual friends (i.e. friends of friend) in Linkedin and Facebook respectively.
- Google+ is one of social network which gives link nodes in groups categories like Friends, Family, Acquaintances following Featured on Google+ etc.
- Social network is huge platform to analyze data and obtain information. Further will see efficient algorithm to discover different graphs properties.

10.2 Social Network as Graphs

MU - Dec. 16

Q. Write a note on Social Network Graphs. (Dec. 16, 10 Marks)

Q. How can a social network treated as graph?

- In general a graph is collection of set of edges (e) and set of vertices (V). If there is an edge exists between any two nodes of graph then that node relates with each other.
- Graphs are categories by many parameters like ordered pairs of nodes, unordered pairs of nodes.
- Some edge has direction, weight. Relationship among graph is explained with help of an adjacency matrix.
- Small network can be easily managed to construct a graph, it is quite impossible with huge/wide network.
- Summary statistics and performance metrics are useful for design of graph for a large network.
- Network and graphs can be elaborate with the help of few parameters like diameter i.e. largest distance between any two nodes, centrality degree distribution.
- Social website like Facebook uses undirected social graph for friends while directed graph used in social website like Twitter, Google+ (plus). Twitter gives connection like 1st , 2nd , 3rd and Google classify linked connection in friends, family, Acquaintances, Following etc.

10.2.1 Parameters Used in Graph (Social Network)

Q. Explain the graph parameters listed below :
(i) Degree (ii) Geodesic distance (iii) Density.

Q. How degree, closeness, between's centrality is measured?

Every node is distinct in a network and it is part of graph by set of links. Some general parameter consider for any social network as graph are :

(a) **Degree :** Number of adjacent nodes (considering both out degree and in-degree). Degree of node n_i denoted by $d(n_i)$.

(b) **Geodesic Distance :** Actual distance between two node n_i and n_j, expressed by d (i, j).

(c) **Density :** It gives correctness of a graph, it is useful to count closeness of network.

(d) **Centrality :** It tells about degree centrality i.e. nodes appearance in the centre of network centrality has types.

Example :

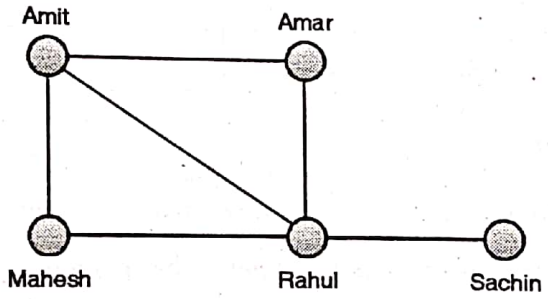


Fig. 10.2.1

Degree of each nodes are as follows :

Nodes	Degree
Amit	3
Amar	2
Mahesh	2
Rahul	4
Sachin	1

Density of undirected graph is 0.6.

Geodesic Distances between two nodes is as follows :

	Amit	Amar	Mahesh	Rahul	Sachin
Amit	–	1	1	1	2
Amar	1	–	2	1	2
Mahesh	1	2	–	1	2
Rahul	1	1	1	–	1
Sachin	2	2	2	1	–

Degree of centrality

$$C'_D(n_i) = \frac{d(n_i)}{(g-1)}$$

Closeness centrality

$$C'_C(n_i) = C'_C(n_i)(g-1)$$

Between's centrality

$$C_B(n_i) = C_B(n_i) / \left[\frac{(g-1)(g-2)}{2}\right]$$

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) / g_{jk}$$

g_{jk} = The number of geodesics connecting jk

$g_{jk}(n_i)$ = The number that actor i is on.

10.2.2 Varieties of Social Network

Q. What is social network? Explain any one type in details.

10.2.2(A) Collaborative Network

- A network where each node has some value and as it gets connected with another node its values get changed.
- A tennis player has some records on his name in single. There are some other records on his name associated with another player name in doubles.
- A node may have different values depending on its connection with neighbouring node.
- Several kinds of data are available having two or more common networks.

10.2.2(B) Email Network

- When a node represents an Email account it is a single node. Every node of an e-mail is in link with at least one e-mail account (i.e. sender mail ID and receiver mail ID).
- Sometimes email are send from one side and sometime e-mail are send from both side in such scenario edges are supposed weak and strong respectively.

10.2.2(C) Telephone Network

- These nodes consist with values like phone numbers which gives it a distinct value.
- As a call is placed between two user nodes get some additional values like time of call period of communication etc.
- In telephone network edge gets weight by number of calls mode by it to other. Network assign edges with the way they contact each other like frequently, rarely, never get connected.

10.3 Clustering of Social Network Graphs

- Cluster gives data into subsets of related or linked objects. Custer coefficient gives degree to which various nodes of a graph tend to duster together.

- Graphs are used to represent data by few clustering algorithms. Clusters can be generated on basis of graph based properties.

10.3.1 Distance Measure for Social-Network Graphs

- Measuring a distance is an essential task for applying clustering technique on any graph. Few graph edge has label, it represents distance measure. Some edge of graph may be unlabeled.
- The distance d(x, y) = 0 if there is an existence or presence of an edge i.e. nodes appear close as there is an edge. The distance d(x, y) = 1 means no edge or nodes appear distant.
- ⊥ and ∞ can be used to represent values for an existing edge.

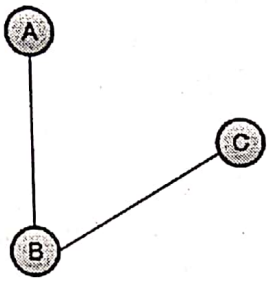


Fig. 10.3.1 : Example for triangle inequality

- '0 and ⊥' or '⊥ and ∞' these are not true 2-valued distance measure. Uses of these values violate triangle inequality when there are edge and nodes combination as shown in Fig. 10.3.1
- In above example, there is edge (A,B) and edge (B, C) but there is no any edge between node A and node C i.e. edge (A.C)
- Above example can be valued by assigning value ⊥ to distance of an existing edge and 1.5 to distance of missing edge

10.3.2 Applying Standard Cluster Method

Q. Explain following clustering algorithm in short :
(a) Hierarchical (b) K-means (c) K-medoid d) Fuzzy C-means.

- Clustering is popular technique to find patters from large dataset domain. It is useful in visualization of data and hypothesis generation.

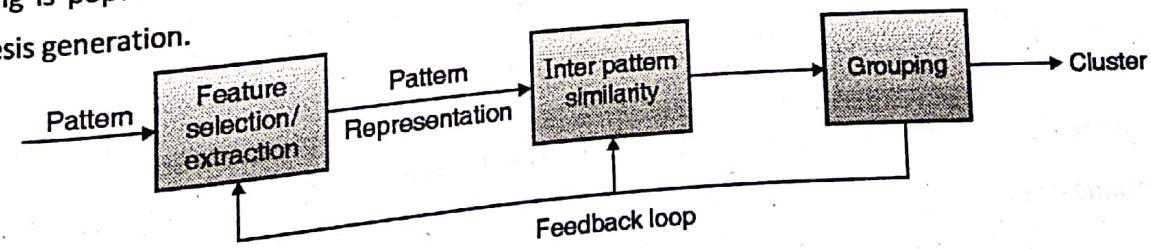


Fig. 10.3.2 : Overview of clustering

- Various clustering algorithms are as follows :

(A) Hierarchical
(B) K-means
(C) K-medoid
(D) Fuzzy C-means

(A) Hierarchical clustering

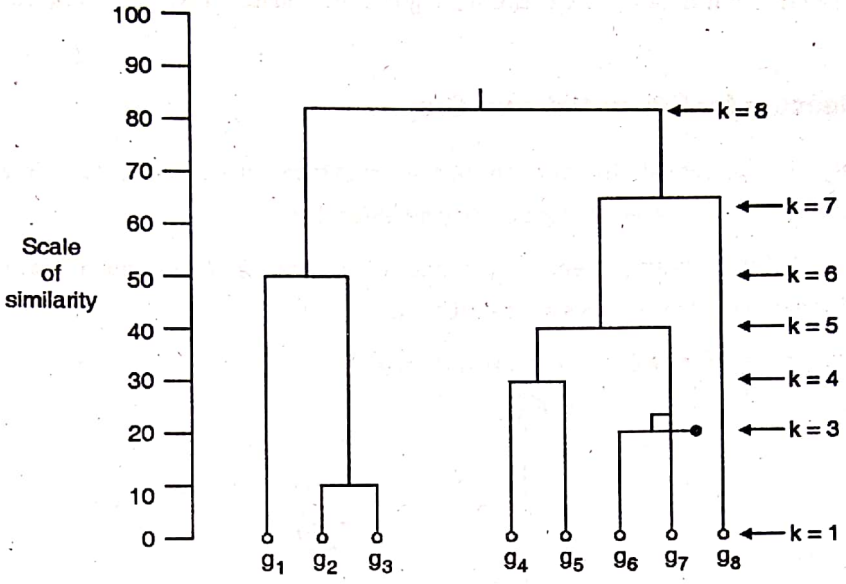


Fig. 10.3.3 : Hierarchical Clustering example

There are two types of hierarchical clustering :

(i) Agglomerative (bottom-up)

(ii) Divise (top-down)

(i) Agglomerative (bottom-up)

It starts with each document assuming as a single cluster, ever almost all documents are belonging to one cluster.

(ii) Divise (top-down)

It start with all document which are part of a single and same cluster. Every node generates a cluster for its own.

(B) K-means clustering

- It is one of the unsupervised clustering algorithm.
- Number of cluster represented by 'k' it is an input to algorithm.
- It is basically an iterative in nature, it work on numerical data and it is easy to for implementation.
- Bayesian Information Criterion (BIC) of Minimum Description Length (MDL) can be used to estimate K ('K' is a user input)
- It is easy to work with any distance measure with K-medoids, K-medoids is general version of K-means algorithm.

(C) K-medoid clustering

- It work with quantitative variable types and numerical.
- In both categorical variables and outliers Euclidean distances do not work in better way.
- It is more intensive.
- Compare to K-means, it is computationally costlier.
- It is applied for categorical data and when data points are not available (i.e. pair wise distances available only).

(D) Fuzzy C-means clustering (FCM)

- It is unsupervised and it always converges.
- It allows one piece of data which is part of two or more clusters.
- It is used frequently in pattern recognition.

10.3.3 Betweenness

- To find communities among social networks some specialized technique are developed as there are few problems with standard clustering methods.
- Betweenness is shortest path available between two nodes. For example an edge (x, y) is betweenness of node a and b such that the edge (x, y) lies on shortest path between a and b.
- a and b are two different communities where edge (x, y) lies somewhere as shortest path between a and b.

10.3.4 The Girvan - Newman Algorithm

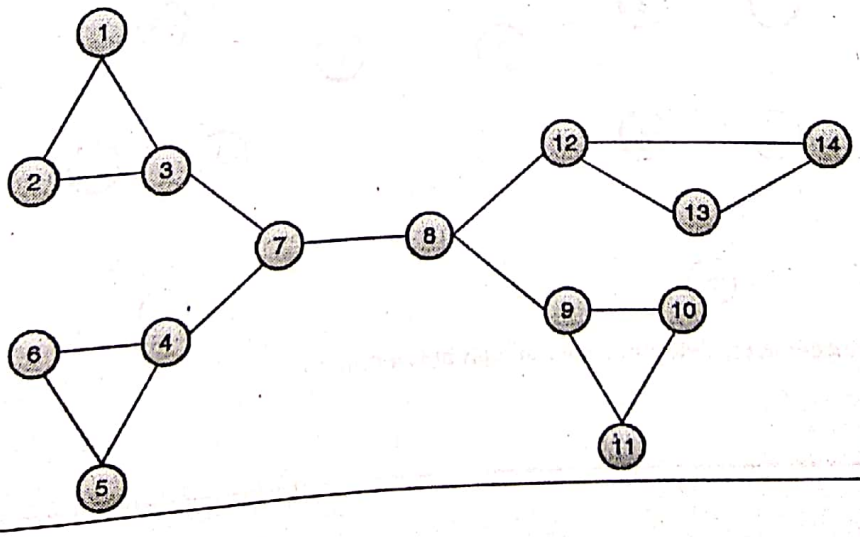
MU - May 19

Q. Explain Girvan-Newman algorithm to mine Social Graphs. (May 19, 10 Marks)

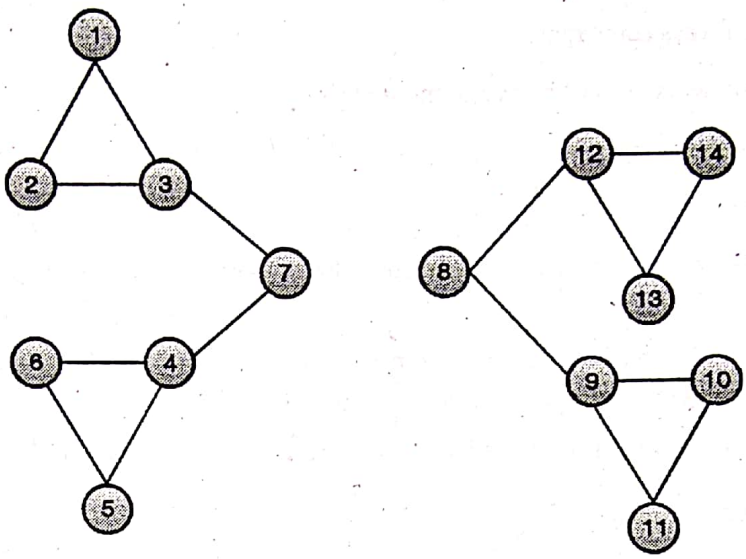
Q. Explain Girvan-Newman algorithm in details with help of suitable example.

- It is published in 2002 by Michelle Girvan and mark Newman for :
  - o Community detection.
  - o To measure edge - betweenness among all existing edges.
  - o To remove edge having large valued betweenness.
  - o To option optimized modular function.
- Girvan Newman algorithm checks for edge betweenness centrality and vertex betweenness centrality.
- Vertex betweenness centrality is total number of shortest path that pass through each vertex on the network.
- If any ambiguity found with above (vertex betweenness centrality) then every path is adjusted to equal weight 1/N among all N paths between two vertices.
- Edge betweenness centrality is number of shortest path which pass through given edge.

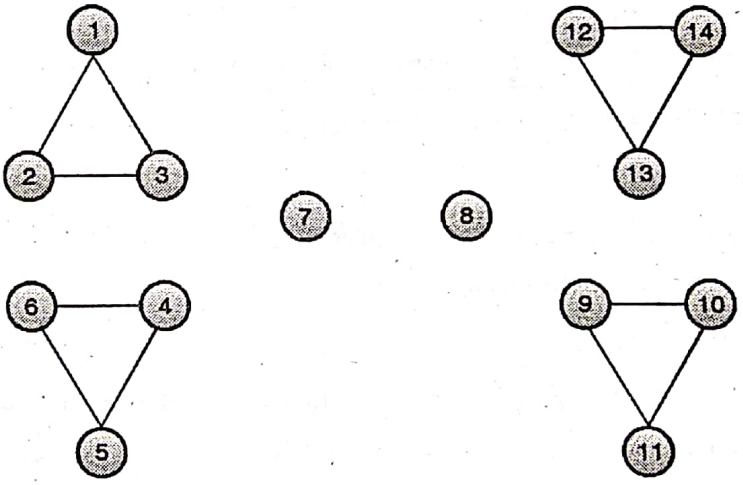
Example : Successive delete edges of high betweenness.



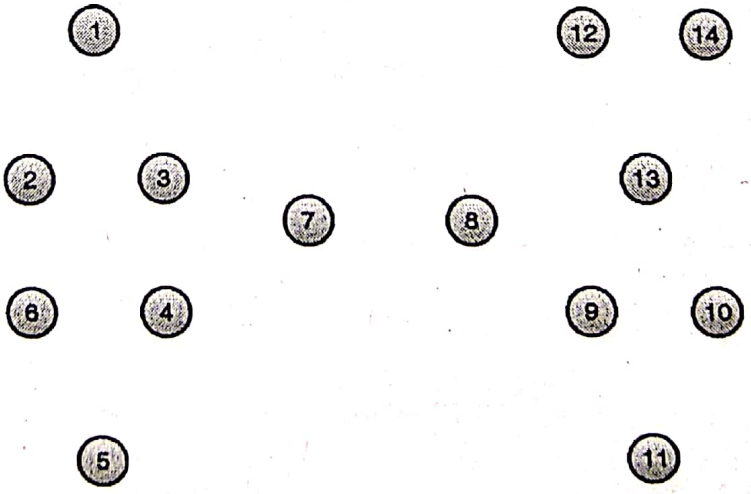
Step 1



Step 2



Step 3



– Standard process to successively deleting edges of high betweenness.

Step 1 : Find edge with highest betweenness of multiple edges of highest betweenness if there is a tie- and remove those edges from graph. It may affect to graph to get separate into multiple components. If so, this is first level of regions in the portioning of graph.

Step 2 : Now, recalculate all betweenness and again remove the edge or edges of highest betweenness. It will break few existing component into smaller, if so, these are regions nested within larger region. Keep repeatation of tasks by recalculating all betweenness and removing the edge or edges having highest betweenness.

10.3.5 Using Betweenness to Find Communities

- It is an approach to find most shortest path within a graph which connect two vertex.
- It is process of systematically removal of edges, edges having highest betweenness are preferred to remove first, process is continued till graph is broken into suitable count of connected components.

Example :

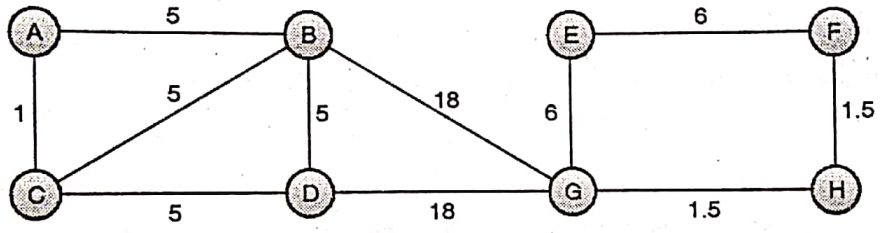


Fig. 10.3.4 : Betweenness score for graph example

- Between C and B there are two path, so edge (A, B), (B, D), (A, C) and (C, D) get credited by half a shortest path.
- Clearly, edge (D, G) and (B, G) has highest betweenness, so it will get removed first, it will generate components namely {A, B, C, D} and {E, F, G, H}.
- By keeping removal with highest betweenness next removal are with score 6 i.e. (E.G) and (E, F). Later, removal with score 5 i.e. (A, B), (B, D) and (C, D) .
- Finally graph remains as :

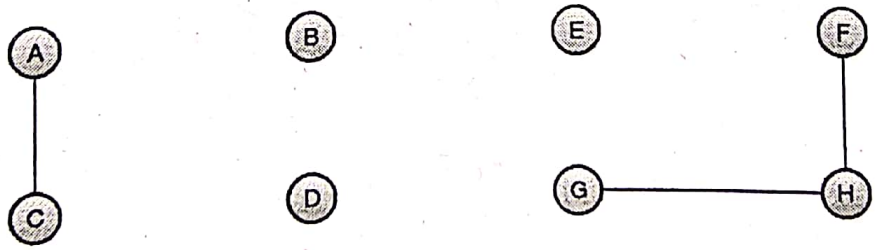


Fig. 10.3.5 : All edges with betweenness 5 and more are removed

- 'Communities' implies that A and C more close to each other than to B and D. In short B and D are "traitor" to community {A, B, C, D} because they have friend G outside the community.
- Similarly G is "traitor" to group {E, F, G, H} and only F, G and H remain connected.

10.4 Direct Discovery of Communities

MU - Dec. 17

Q. Explain any one algorithm for finding communities in a social graph. (Dec. 17, 6 Marks)

Q. Explain communities discovery.

Discovery of communities deals with large number edges search from a graph.

Finding cliques

- Cliques can be defined as a set of nodes having edges between any two of vertices.
- To find clique is quite difficult task. To find largest set of vertices where any two vertices needs to be connected within a graph is known as maximum clique.

10.4.1 Bipirate Graph

- It is graph having vertices which can be partitioned into two disjoint sets suppose set V and set U. Both V and U sets are not necessary of having same size.
- A graph is said to be bipirate if and only if it does not posses a cycle of an odd length.

Example :

Suppose we have 5 engines and 5 mechanics where each mechanic has different skills and can handle different engine by vertices in U. An edge between two vertices to shows that the mechanics has necessary skill to operate the engine which it is linked. By determining maximum matching we can maximize the number of engines being operated by workforce.

10.4.2 Complete Bipirate Graph

- A graph $K_{m, n}$ is said to be complete bipirate graph as its vertex set partitioned into two subsets of m and n vertices respectively.
- Two vertices are connected if they belong to different subsets.

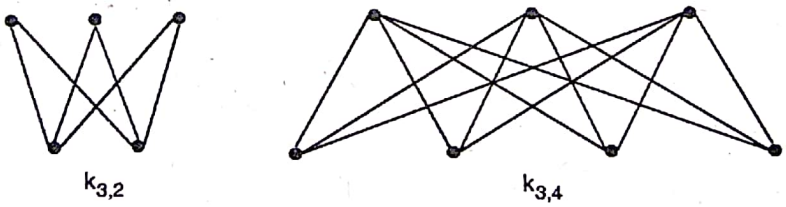


Fig. 10.4.1

10.5 Simrank

Q. What is simrank ? Explain in brief.

- It is one of the approach of analyze a social network graphs.
- Graphs consists of various types of nodes, simrank is useful to calculate the similarity from same type nodes.
- Simrank is useful for random walkers on a social graph while starting with a particular node.
- Simrank needs calculation and it is done at every starting node for limited sizes graph.

10.5.1 Random Walker on Social Network

- Social network graph is mostly undirected and web graph founds directed. Random walker of social graph can meet to any number of neighboring node of it.

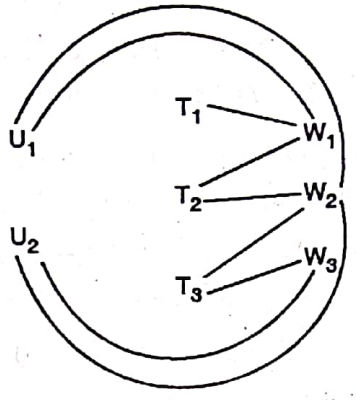


Fig. 10.5.1 : A tripartite graph example for random walker social network

- Suppose, example as shown in Fig. 10.5.1 U-users, T-tags, W-web pages.
- At very first step walker will go for U_1 or W_1. If walker prefer to go W_1 then in next attempt it will visit to T_1 or T_2. If walker prefer to visit U_1 then in next attempt it will meet either T_1, T_2 or T_3.
- T_1 and T_2 both tag placed on webpage W_1 and reachable by user U_1 and U_2.
- By keeping trend of visiting start at any node traversal can visit all node of a_1 network. Irrespective of start node walker can visit all node so it is known as random walker on social network.

10.5.2 Random Walks with Restart

- Random node visiting walker may stop at some node in random.
- To know when a random walker may stop can be calculate with help of probabilities putting in a matrix of transition.
- Suppose, M is transition matrix of graph G entering at row a column b of M is 1/k if node b of graph having degree K and one of the adjacent node is a else entry is 0 (zero).

Example :

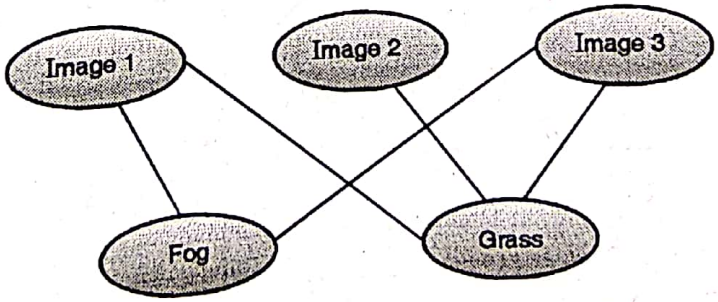


Fig. 10.5.2 : A simple bipartite social graph

- In Fig. 10.5.2 network consists of three images, and two tags "Fog" and "Grass" Image1 and Image 3 has two tags and image 2 has only one tag i.e. "Grass".
- Image 1 and Image 3 are comparably more similar than Image 2 and random walker with restart at Image 1 probably support that intention after applying analysis to it.

- Nodes can be kept in order like Image 1, Image 2, Image 3, fog, grass. The transaction matrix for graph will be like.

$$\begin{bmatrix} 0 & 0 & 0 & 1/2 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 1/2 & 1/3 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/2 & 1 & 1/2 & 0 & 0 \end{bmatrix}$$

- The fifth column of node "Grass" which is connected to each of image node. If therefore it has some degree like 3 then non-zero entries to node "Grass" column must have to be 1/3.
- The image nodes correspond to first three rows and first three columns of transaction matrix so entry 1/3 appears in the first three rows of column 5. Since "fog" node does not have an edge to either itself or "Grass" node.
- Let, β be probability of random walker, so $1-\beta$ is probability the walker will teleport to initial node N. e_N is column vector that has 1 in the row for node N otherwise its 0 (zero).

10.6 Counting Triangles using MapReduce

Q. What is counting triangles using MapReduce? Enlist its application.

- In era of Big data, approx 2.5 quintillion byte data increasing per day. In 2004, Google introduced Map Reduce used in search engine.
- Map Reduce used for processing and to generate large data sets. Map function gives data processed by a pair of key set while Reduce function used to merge those data values.
- Map and reduce function likewise function introduced in Lisp also.

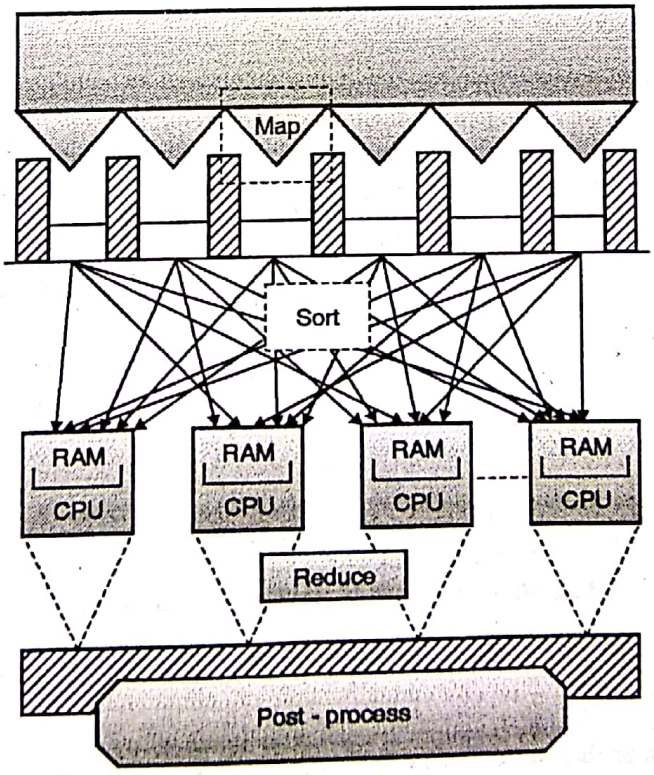


Fig. 10.6.1 : Map-Reduce process general overview

- Map-Reduce process does with help of three stages :
  - Mapping
  - Shuffle
  - Reducing
- Counting of triangle is helpful to know community around any node within social network; it helps to know 'clustering co-efficient'.
- Let, a graph G = (V.E.) a simple undirected and unweighted graph.

$$\text{Let, } n = |v|$$
$$m = |E|$$
$$T(v) = \text{set of neighbors of } v$$
$$= \{ W \in V \mid (v, w) \in E \}$$
$$dv = |T_{(v)}|$$

Cluster co-efficient (cc(v)) for a node where $v \in V$ is,

$$\binom{dv}{2} cc(v) = |\{(u, w) \in E \mid u \in T(v), w \in T(v)\}|$$

Above expression gives cluster coefficient for node v

- Map Reduce is used in page ranking. It is also useful in :
  - (i) Web access log states,
  - (ii) Inverted index construction
  - (iii) Document clustering
  - (iv) Statistical machine translation
  - (v) Machine learning
  - (vi) Web link-graph reversal
  - (vii) Distributed sorting
  - (viii) Distributed pattern based search
  - (ix) Machine translation

Review Questions

Q.1 What is sociogram and Barbasi-Albert algorithm?

Q.2 How can a social network treated as graph ?

Q.3 Explain the graph parameters listed below :
  - (i) Degree
  - (ii) Geosestic distance
  - (iii) Density

Q. 4 How degree, closeness, between's centraliy is measured ?

Q. 5 What is social network ? Explain any one type in details.

Q. 6 Explain following clustering algorithm in short :

(a) Hierarchical

(b) K-means

(c) K-medoid

(d) Fuzzy C-means

Q. 7 Explain Girvan-Newman algorithm in details with help of suitable example.

Q. 8 What is betweenness ? Explain use of betweenness to find communities.

Q. 9 Explain bipirate graph and complete bipirate graph.

Q. 10 What is simrank ? Explain in brief.

Q. 11 What is MapReduce ? Enlist its application ?