



Hope Foundation's
Finolex Academy of Management and Technology
P60, P60-1, MIDC, Mirjole, Ratnagiri, Maharashtra, Pin 415639
Information Technology Department

Big Data Analytics

Date : 25th March 2020

Assignment II- Open Book Test

YEAR/SEM: BE VIII

Name: Jawaad A Rahim Kazi

Roll No: 28

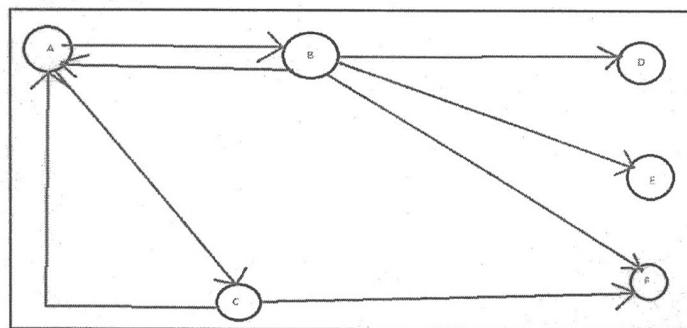
Part I -Answer Following Questions.

1. What is a Data Stream ? Explain with example. (R)
2. What is Data Stream Mining ? Give four examples of Data Stream Mining. (R)
3. What is a Collaborative filtering system in view of Recommendation Systems? (R,A)
4. What is CQL , where it is used ?(R,A)
5. What are Hash Functions? Give their application and one example (R)
6. Define (a) Tendrils and (b) Tubes (c) SCC (d) Simrank (e) Pagerank (R)
7. Where do we see recommendation? List any Four. (R)
8. Draw Block diagram of typical DSMS. (R)
9. What is Random Surfer Concept? (R)
10. Define (a) Hubs (b) Authority pages (R)
11. What is Link Farm ?(R)
12. What is the use of Recommendation Systems? (R,A)
13. What is HITS algorithm? Explain differences between HITS and Pagerank. (R,A)

Part II - Answer Following Questions, Five (05) Marks per Question:

1. Explain Bloom's Filter with example. A bloom filter with $m=1000$ cells is used to store information about $n=100$ items, using $k=4$ hash functions. Calculate the false positive probability of this instance. Will the performance improve by increasing the number of hash functions from 4 to 5. Explain your answer. (A,AN,E)
2. Consider a Web Graph as given Below: (A,AN,E)

Assume that Pagerank values for any page 'm' at iteration 0 is $PR(m)=1$ and teleportation factor for iteration is $\beta =0.85$. Perform the PageRank algorithm and determine the rank of every page at iteration 2.



Priyanka Bandagale

Asst Professor, IT

Q.1

Aus

In connection-oriented communication, a data stream is a sequence of digitally encoded coherent signals used to transmit or receive information that is in the process of being transmitted. A data stream is the set of extracted information from a data provider. It contains raw data that was gathered out of users' browser behavior from websites, where a dedicated pixel is placed.

Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches and sensor data.

Q.2

Aus

Data Stream mining is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities.

Example -

- (1) Adaptive malicious code detection.
- (2) On-line malicious URL detection
- (3) Evolving insider threat detection
- (4) Textual stream classification.

Q.3

Aus

Collaborative filtering systems recommend items based on similarity measures between users and/or items. The items recommended to a

user are those preferred by similar users. This sort of recommendation system can use the groundwork laid on similarity search and on clustering. However, these technologies by themselves are not sufficient, and there are some new algorithms that have proven effective for recommendation systems.

Q.4

Ans Contextual Query Language (CQL), previously known as Common Query Language is a formal language for representing queries to information retrieval systems such as search engines, bibliographic catalogs and museum collection information. Based on the semantics of Z39.50, its design objective is that queries be human readable and writable and that the language be intuitive while maintaining the expressiveness of more complex query languages. It is being developed and maintained by the Z39.50 Maintenance Agency, part of the Library of Congress.

Q.5

Ans A hash function is any function that can be used to map data of arbitrary size to fixed size values. The values returned by a hash function are called hash values, hash codes, digests or simply hashes. The values are used to index a fixed-size table called a hash table.



Applications -

- ① Construct a message authentication code (MAC)
- ② Digital signature
- ③ Make commitments but reveal message later.
- ④ Timestamping
- ⑤ Key updating: key is hashed at specific intervals resulting in new key.

The example of a hash function is a book call number. Each book in the library has a unique call number. A call number is like an address: it tells us where the book is located in the library.

Q.6

Ans.

a. Tendrils - Tendrils which are of two types. Some tendrils consist of pages reachable from the in-component but not able to reach the in-component. The other tendrils can reach the out-component, but are not reachable from the out-component.

b. Tubes - Tubes, which are pages reachable from the in-component and able to reach the out-component, but unable to reach the SCC or be reached from the SCC.

c. SCC - It stands for strongly connected Component. A directed graph is strongly connected if there is a path between all pairs of vertices. A strongly connected component (SCC) of a directed graph is a maximal strongly connected subgraph.

d. SimeRank - It is an approach to analyze social

network graphs. Its purpose is to measure the similarity between nodes of the same type and it does so by seeing where random walkers on the graph wind up when starting at a particular node.

e. Page Rank - Page Rank is a function that assigns a real number to each page in the web (or at least to that portion of the web that has been crawled and its links discovered). The intent is that the higher the page rank of a page, the more "important" it is.

Q. 7

Uses of Recommendation -

- ① News Recommendations: Offering news articles to on-line news paper readers, based on prediction of reader interests.
- ② Product Recommendations: Offering customers of an on-line retailer suggestions about what they might like to buy, based on their past history of purchases and/or product searches.
- ③ Movie Recommendations: Netflix offers its customers recommendations of movies they might like. These recommendations are based on ratings provided by users.
- ④ Book Recommendations: Amazon Kindle offers its customers recommendations of books they might like. These recommendations are based on ratings provided by the reader.



Q.8

Ans

Ad-hoc
Queries

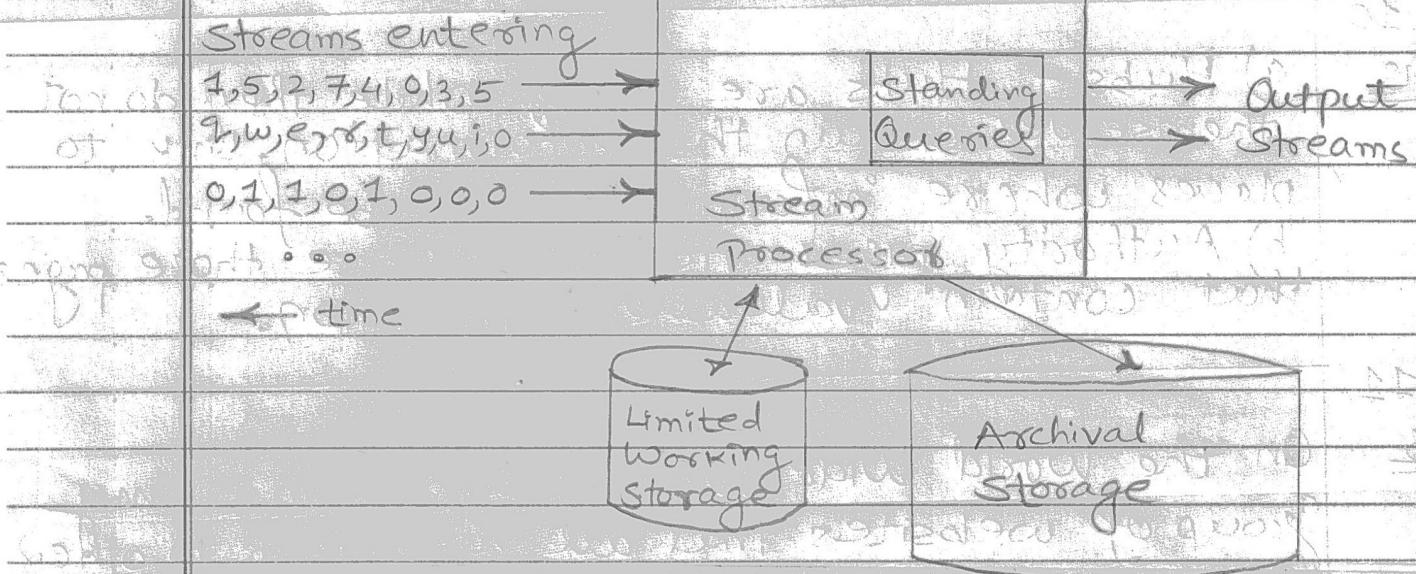


Fig: Block diagram of typical DSMS

Q.9

Ans

The random surfer model provides a basis for calculating the PageRank algorithm. The model represents the behavior of internet users and provides a probability of a random user visiting a webpage. A surfer moves through the internet in two ways. They may enter a URL or use a bookmark to go directly to a webpage. Or they may follow a series of successive links until they again access a new page. In a random surfer model it is assumed that the link which is clicked next is selected at random. The content does not matter. Moreover it is assumed that clicking another link is not an infinite chain,

but that a random user will lose interest in following those links at a certain point and visit a new website instead.

Q. 10

Ans

a) Hubs - Hubs are pages which do not themselves contain the information, link to places where information can be found.

b) Authority pages - Authorities are those pages that contain valuable information.

Q. 11

Ans

On the World Wide Web, a link farm is any group of websites that all hyperlink to other sites in the group for the purpose of increasing SEO rankings. In graph theoretic terms, a link farm is a clique. Although some link farms can be created by hand, most are created through automated programs and services. Eg: Blog Nw.

Q. 12

Ans

A recommendation system is a subclass of information filtering systems that ~~take~~ seeks to predict the "rating" or "preference" a user would give to an item. They are primarily used in commercial applications.

Recommendation systems are utilized in a variety of areas and are most commonly recognized as playlist generators for video and music services like Netflix, YouTube and Spotify, product recommenders for services such as Amazon or content recommenders for social media platforms such as Facebook.



and Twitter. These systems can operate using a single input like music, or multiple inputs within and across platforms like news, books and search queries. There are also popular recommendation systems for specific topics like restaurants and online dating. Recommendation systems have also been developed to explore research articles and experts, collaborators and financial services.

Q.13

Ans

Hyperlink Induced Topic Search (HITS) algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg. This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search. HITS use hubs and authorities to define a recursive relationship between webpages.

Parameters	HITS	Page Rank
Scoring Criteria	Good authorities are pointed to by good hubs and good hubs point to good authorities.	Webpage is important if it pointed to by other important pages.
Number of scores	Dual rankings @ One of scores with the most authoritative documents related to the query. ⑥ Otherwise with the most "hubby" documents.	Page Rank score is query independent

Parameters	HITS	Page Rank
Query independence	HITS score is calculated after getting the neighborhood graph according to the query.	Page Rank score is query independent.
Resilience	Susceptible to spamming since addition of pages slightly affects the ranking.	Since PageRank is able to isolate spam, it is resilient to spamming.

Ans 14

A Bloom filter is a space-efficient probabilistic data structure that is used to test whether an element is a member of a set. An empty bloom filter is a bit array of m bits, all set to zero. We need K number of hash functions to calculate the hashes for a given input. When we want to add an item in the filter, the bits at K indices $h_1(x), h_2(x), \dots, h_K(x)$ are set, where indices are calculated using hash functions. Example - Suppose we want to enter "jawwad" in the filter, we are using 3 hash functions and a bit array of length 10, all set to 0 initially.

0	0	0	0	0	0	0	0	0	0
1	2	3	4	5	6	7	8	9	10

First, we'll calculate the hashes as following -

$$h_1("jawwad") \% 10 = 1$$

$$h_2("jawwad") \% 10 = 4$$

$$h_3("jawwad") \% 10 = 7$$



Now, we will set the bits at indices 1,4 and 7 to 1.
jawwad

1	0	0	1	0	0	1	0	0	0
1	2	3	4	5	6	7	8	9	10

Again, we want to enter "Kazi", similarly
we'll calculate hashes

$$h_1("Kazi") \% 10 = 3$$

$$h_2("Kazi") \% 10 = 5$$

$$h_3("Kazi") \% 10 = 4$$

set the bits at 3,5 and 4 to 1.

1	0	1	1	1	0	1	0	0	0
1	2	3	4	5	6	7	8	9	10

Now, if we want to check "jawwad" is present
in filter or not. We'll do the same process
but this time in reverse order. We calculate
respective hashes using h_1 , h_2 and h_3 and check
if all these indices are set to 1 in the bit
array. If all the bits are set then we can say
that "jawwad" is probably present. If any
of the bit at these indices are 0 then "jawwad"
is definitely not present.

Let $m=1000$, $n=100$, $k=4$.

We have, probability of False positivity,

$$P = \left(1 - \left[1 - \frac{1}{m}\right]^{kn}\right)^k$$

$$= \left(1 - \left[1 - \frac{1}{1000}\right]^{4 \times 100}\right)^4$$

$$= \left(1 + \left[1 - \frac{1}{1000} \right]^{400} \right)^4$$

$$= (1 - 0.6702)^4$$

$$P = 0.0118$$

If we consider $k=5$, then

We have, probability of false positivity,

$$P = \left(1 - \left[1 - \frac{1}{m} \right]^{k \cdot n} \right)^k$$

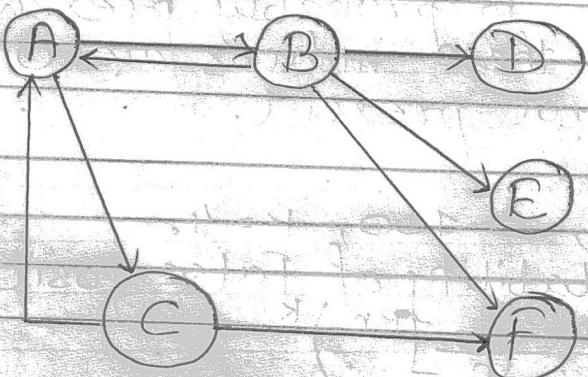
$$= \left(1 - \left[1 - \frac{1}{1000} \right]^{5 \times 100} \right)^5$$

$$= \left(1 - \left[1 - \frac{1}{1000} \right]^{500} \right)^5$$

$$= (1 - 0.6064)^5$$

$$P = 0.0095$$

From above results we can see that, by increasing number of hash functions (k) from 4 to 5, the false positive probability of the instance got reduced to 0.0095 from 0.0118. Hence, we can say that the lesser the no. of hash functions the greater is the false positive probability.





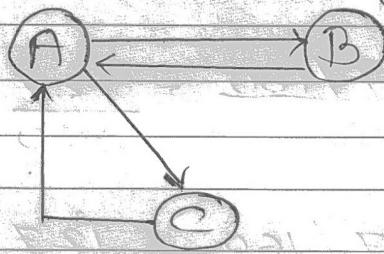
Transition Matrix

$$M = \begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 \end{bmatrix}$$

From the graph it is clear that D, E and F are dead ends.

Before applying PageRank algorithm, remove the dead ends.

Hence, the new graph will be



So, the new transition matrix will be

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

$\beta = 0.85$ and initial PageRank vector

$$v = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Iteration 1:

$$V = \beta M v + (1 - \beta) e/n$$

$$V = \frac{17}{20} \begin{bmatrix} 0 & \frac{1}{2} & 1 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + \frac{3}{20} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 17 \\ 20 \end{bmatrix} + \begin{bmatrix} 2 \\ 1/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \end{bmatrix}$$

$$= \begin{bmatrix} 34/20 \\ 17/40 \\ 17/40 \end{bmatrix} + \begin{bmatrix} 1/20 \\ 1/20 \\ 1/20 \end{bmatrix} = \begin{bmatrix} 35/20 \\ 19/40 \\ 19/40 \end{bmatrix}$$

Iteration 2:

$$V = \beta m v + (1 - \beta) e / n$$

After iteration 2, PageRank of
 page A = $343/400$
 page B = $127/160$
 page C = $127/160$

$$\text{page D} = \frac{\text{PR}(B)}{2} = \frac{127}{160} = \frac{127}{320}$$

$$\text{page E} = \frac{\text{PR}(B)}{2} = \frac{127}{160} = \frac{127}{320}$$

$$\text{page F} = \frac{\text{PR}(B)}{2} + \frac{\text{PR}(B)}{2} = \frac{127}{320} + \frac{127}{320} = \frac{254}{320}$$
