ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Data Mining with Big Data Image De-Duplication in Online Social Network Website

Hashmi.S.Taslim, Ganesh Dhanokar

M.E, Dept. of CSE, G.H.Raisoni Engg, College, Jalgaon, Maharashtra, India.

M.E, Dept. of CSE, G.H.Raisoni Engg College, Jalgaon, Maharashtra, India.

ABSTRACT: Big data is a collection of data sets which are large and complex. One fundamental issue in today Online Social Networks is large data storage. Some 618 million people use Face book every day. According to Data Center Knowledge's Rich Miller report Face book stores more than 240 billion photos, with some 350 million new photos uploaded a day. It's a huge amount of data and the additionally most of the time different user post same photo or image on their wall. This creates image duplications on storage server. The proposed work an efficient way of storing and De-Duplication of images on server of On-line Social Networking using Map Reduce technique. In this approach server will maintain only one copy of image on server and provides access to all users who have uploaded it. Map Reduce is simple and parallel computing techniques normally used for analyzing the huge data. Traditional De-duplication schemes works if and only if the second image having the same underlying bits as first. This restricts the performance of many applications as exact images need to be there if want to succeed. In many practical applications where the storage restriction is present, users uploads the modified images varying with the quality or resolution. Experimental results demonstrate in a real dataset, the proposed approach not only effectively saves storage space, but also significantly improves the retrieval precision of duplicate images. In addition, the selection of the images can meet the requirements of people's perception**.**

KEYWORDS: Duplicate image identification, De-Duplication, Data partitioning, Map Reduce, Pearson Correlation, Performance evaluation and Optimization;

I. INTRODUCTION

"Big Data" is data that becomes large enough that it cannot be processed using conventional methods. The term Big Data concerns with the huge volume, complex and rapidly growing data sets with multiple, independent sources .Due to fast development of networking ,data storage and data collection capacity the concept of big data is now rapidly expanding in all science and engineering domains including biological, physical and biomedical sciences. Social networking sites, mobile phones And as storage capacity continues to enlarge, today's "big" is certainly tomorrow's "medium" and "small."[2] The best meaningful definition of "big data" is when the size of the data itself becomes part of the problem. One fundamental issue in today Online Social Networks is large data storage. Some 618 million people use Face book every day. According to Data Centre Knowledge's Rich Miller report Face book stores more than 240 billion photos, with some 350 million new photos uploaded a day. It's a huge amount of data and the additionally most of the time different user post same photo or image on their wall. This creates image duplications on server. Hence it becomes slightly impossible to handle such huge amount of data. Once Hadoop comes to the practice it overcomes the said drawbacks. Hadoop is an open source Map Reduce platform used for the parallel computing of the data. Map-Reduce is one of the simple, best and parallel computing techniques frequently used for analyzing the large amount of data. The Map Reduce algorithm contains two important tasks, i.e. Map and Reduce. Map takes a set of data and converts it into another set of data, where each individual element is broken down into tuples <key, value> pair. In map reduce technique users can have a <key, value> pair that can generate group of intermediate key and value. Also reduce function is created which makes use of all these same intermediate keys. Main side of the technique is the programs written using this model is automatically paralyzed which increase the speed of the execution. However, the De-duplication is a well-known technique of reducing the size of data storage by preventing the storage of identical files. A traditional De-Duplication system works if and only if second image having the same underlying bits as first image. This restricts the performance of many applications as exact image need to be there if want to succeed. In many

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

existing applications where the storage restriction is present, many users upload the modified images varying with the quality or resolution. There are many Systems are already existed and still in this area continues working is going with aim of eliminating the redundant copies of images and significantly improve storage utilization. The basic idea of this project comes from the fact that storage servers are big platform to store and to retrieve the data in huge amount. Where there is greater possibility of duplication of the data can be happen by cause of this there will be huge storage space is used unnecessarily. This Results in slow processing of the system. Many systems are existed to identify the duplicate images but eventually all are directly proportional to the number of the images. So there is an urge of a system is required to reduce the duplicate identification time. So this paper presenting an idea of duplicate image identification process using Pearson correlation based on Map/Reduce technique.

A. *MAPREDUCE TECHNIQUE*

Map/Reduce is one of the best, simple and parallel computing techniques normally used for analyzing the huge data. The main motto of Map/Reduce technique is to hide the way by which partitioning takes place and thus to focus on the technique of data processing. In map reduce technique users can have a key/value pair that can generate set of intermediate key and value. Also reduce function is created which makes use of all these same intermediate keys. Number of huge real world tasks can be effectively done using this technique. Main side of the technique is the programs written using this model is automatically paralyzed which increase the speed of the execution. Map/Reduce makes use of Google File System (GFS) as a base layer of storage from which data can be take and store. GFS comes under chunk based data partitioning where the level of fault tolerance is reduced by using replication and data partitioning algorithms. Apache Hadoop is an open source framework of Map/Reduce. Exactly like Map/Reduce, Hadoop consist of two implementation layers.

1) HDFS (Hadoop DFS)
2) HMF (Hadoop Map/Reduce)

HDFS layer comprises of data storage facility and HMF consists of data processing techniques. Fig.1 shows the Map/Reduce Technique.

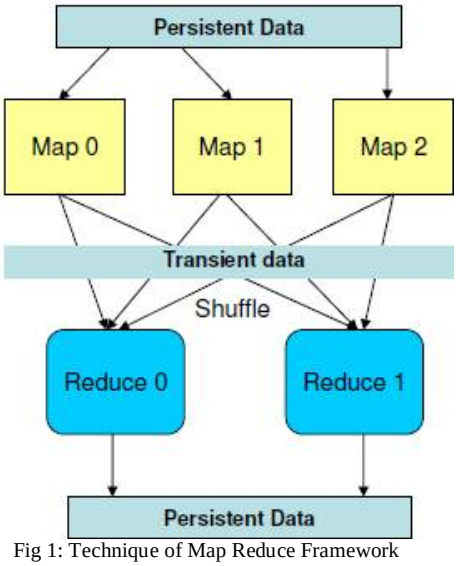

Fig 1: Technique of Map Reduce Framework

B. DATA PARTITIONING

Map/Reduce is an algorithm used in Artificial Intelligence as functional programming. Solve the problems to analyze huge volumes of data set in distributed computing environment. Map/Reduce programming platform is implemented in the Apache Hadoop project that develops open-source software for reliable, scalable, economical, efficient and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

distributed computing. Users normally stores data rows in labelled tables. A data row has a category of key and a number of columns. The table stored lightly, so that rows in the same table can have different columns. Map/Reduce does not guarantee to increase the performance even though we add more nodes because there is a problem for distributing, aggregating, and reducing the large data set among nodes against computing powers of additional machines. The Map Reduce algorithm is for efficiently computing pair wise document similarity in large document collections. In addition to offering specific benefits for a number of real-world tasks, that could be useful for a broad range of text analysis problems. [3]
Partitioning techniques are emerged as a one of the best techniques to rid out of the problem time required for the execution of the system.Fig.2 shows Parallel stream processing architecture overview.

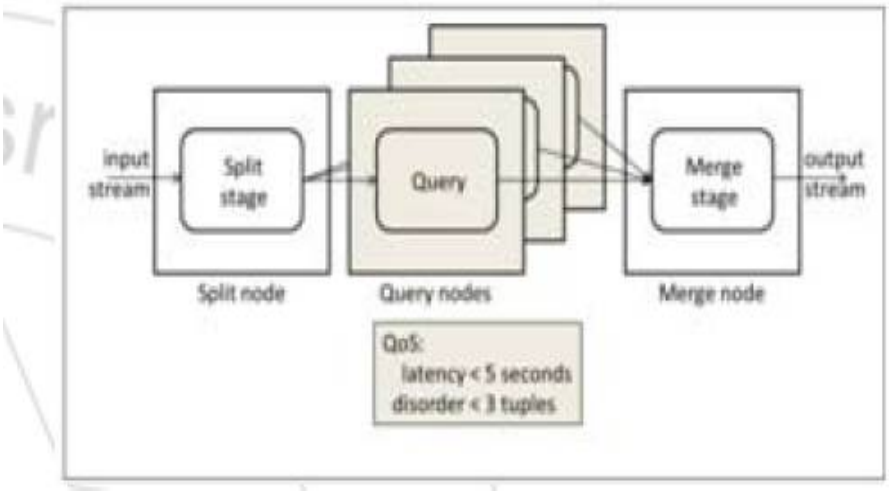

Fig 2: **Parallel stream processing architecture overview.**

In parallel partitioning systems, generally the typical split and then merge pattern is used. Here two nodes are presents, one represent split node and other represents merge node. Split node is used for splitting the things for further execution. The splitting is done in such way that node unbalancing is not happened. The intermediate part on which partitioned data is needed to feed is fixed in advance. After execution of the individual node the result is again gathered at merge node. Merge node combines the all result and then pass it as a output of the system.
The partitioning techniques are normally categorized in two techniques.
1. Batch based partitioning
2. Pane based partitioning.
In batch based partitioning system next to next tasks are grouped to form a batch and thus gives it to the common partitioning system. While in pane based partitioning exactly opposite task of batch based partitioning is done. Here instead of breaking the main task into sub task, the sub tasks are breaking down to the small tasks. Then these tasks are assigned individually to the partitioning.

II. RELATED WORK

1.HIPI an image processing library on Map/Reduce framework. The designing of HIPI an image processing library is done in such way that it hides the implementation of complex Hadoop. Map/Reduce framework and emphasis more on image as it is the thing about which users worrying a lot. The implementation is done by considering huge amount of data, because of this system gives higher throughput in case amount of images exceeds. Map/Reduce pipeline has provision of different formats for accessing the images. The types of images that can be used during Map/Reduce steps are filtered by providing the culling phase during the mapping phase. Float images, most important part in image processing are obtained by using the encoders and decoders phases which presents behind the scene. By adding all these features in the system it gives simplified interface to deal with the images on Map/Reduce[4]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

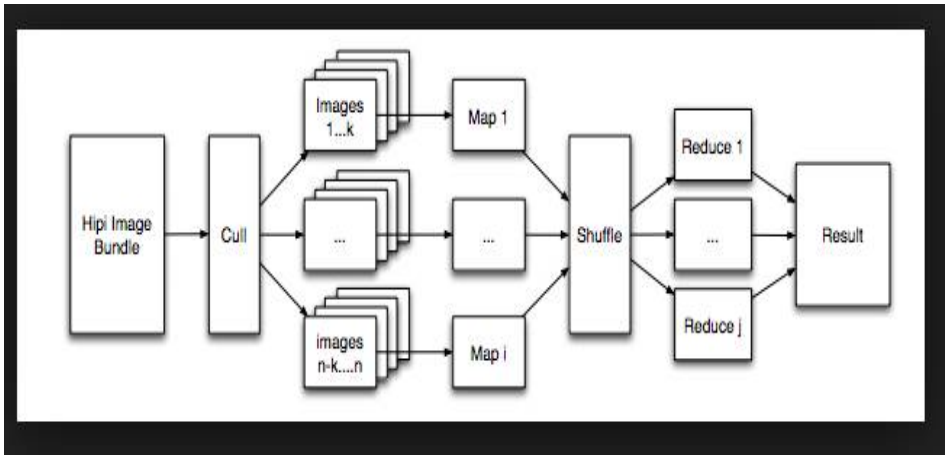

Fig 3: Map/Reduce pipeline architecture

As discussed above traditional De-Duplication systems are performed well if and only if the images to be compared are having same underlying bit codes. But this scenario reduces the usability of applications. So to overcome this presents a novel system of image De-Duplication which makes use of high precision duplication approach. The proposed system comprises of five stages as feature extraction, high-dimension indexing, accuracy optimization, and centroid selection and De-Duplication evaluation by evaluating the system on real datasets it had been observed that system not only gives the efficient image De-Duplication scheme but also greatly improves the precision of duplicate image retrieval.

- Foo, Jun Jie, et al , "Detection of near-duplicate images for web search."The proposed system comprises of five stages as feature extraction, high-dimension indexing, accuracy optimization, and centroid selection and De-Duplication evaluation by evaluating the system on real datasets it had been observed that system not only gives the efficient image De-Duplication scheme but also greatly improves the precision of duplicate image retrieval .Gives a detailed survey on the various De-Duplication strategies being used. Various issues presents in De-Duplication schemes such as bandwidth, high throughput, computational overhead, De-Duplication efficiency, usability of read and write operations.
- As the huge amount of duplicate images are available on web, web search for a particular images tends to give the number of nearby images also which degrades the performance of the system. [15] Different De-Duplication schemes such as application based source De-Duplication scheme, a semantic attribute based source De-Duplication, GPU based source data De-Duplication ,Hadoop based data duplication, hash level based De-Duplication, and causality based De-Duplication are explained.
- “Dean, Jeffrey, and Sanjay Ghemawat"Map-Reduce: simplified data processing on large clusters.", [12].Presents a search theory which shows how the Map/ Reduce technique is used at different works of Google. Authors state the reason behind this. First is, it is simple to use. Even the programmer with less knowledge of parallel and distributed systems can use it effectively. It presents the work scenario in abstract way by hiding the details of load balancing, fault tolerance and parallelization. And the second is huge real word scenarios are effectively expressed using this. E.g. Map/Reduce is effectively used at Google in web search for storing, sorting and data mining. Finally authors conclude that the map reduce can be effectively used for the keeping data without its loss. In order to shows the performance of different searching and sorting task on the system having different configurations. To bring down this idea into reality Hadoop and map-reduce technique for distributed data processing technique is used. Here the machine learning problems classes are distinguished within the map reduce framework to improve the implementation of Hadoop. At the final part of the system they makes statement that the map reduce technique is a best option for the simple operations but still it has many flaws for the complex operations over large database. [13].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

II. PROPOSED ALGORITHM

A. *DESIGN CONSIDERATIONS:*

- Input file is loaded in to the HDFS.
- File is being split in to the block size of 64MB.
- The mapping function is performed on the basis of <k1, v1> pairs.
- Then it is stored on the HBase tables.
- *Pearson co-relation for Reducer*

B. *DESCRIPTION OF THE PROPOSED ALGORITHM:*

Aim of the proposed algorithm is to avoid the duplicate image on storage server.The proposed algorithm is consists of four main steps.

Step 1: Linear Data Partitioning

In this step the total number of images are been taken and they are been divided among the number of working executor servers.

Step 2: Image Blocks

In this step image of size M x N is loaded in a singular vector V of size S, and then block size is allocated as B, Then image blocks can be formed by the following equation

$$\mathbf{f(BV) = Vs \bmod B};$$

Where, BV is block vector

Step 3: Image Parameter With Entropy Matching

Entropy Matching contain two steps:

3.1 Mean and Standard Deviation: Mean and Standard Deviation calculate quality and resolution. Each image is consists of a

range of some pixels values. These values in each image can be used to calculate the mean of image.

3.2 Entropy Calculation

E = entropy (I) returns E, a scalar value representing the entropy of an image I. Entropy is a statistical measure of Randomness that can be used to characterize the texture of the input image. Entropy is defined as E= log (δ). Where, δ is the Standard deviation of the image. Once query image entropy is calculated then the entropy of the dataset image is also been Calculated and then both the entropies are matching with the basic threshold.

Step 4: Duplicate identification by Pearson correlation:

Here in this step for every pair of columns of data from the image blocks is checked for the correlation using Pearson correlation and the image which is having highest correlation blocks is considered as duplicate and then it will eliminate from the storage server.

Pearson Correlation can be represent as below

$$r = \frac{\sum x_i y_i - \frac{\sum x_i y_i}{N}}{\sqrt{\left(\sum {x^2}_i - \frac{(\sum x_i)^2}{N}\right)}\sqrt{\left(\sum {y^2}_i - \frac{(\sum y_i)^2}{N}\right)}}$$

Where,

N = Number of pairs of data

$\sum xy$ = Sum of the product of paired data

$\sum x$ = Sum of x data

$\sum y$ = Sum of y data

$\sum x^2$ = Sum of squared x data

$\sum y^2$ =Sum of squared y data

This can be represent with the below algorithm

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

III. PSEUDO CODE

INPUT: Two parameter matrix of N rows and Two columns and let matrix be M
OUTPUT: Pearson factor r (i.e in between 0 to 1)

Step 1: Calculate sum of square of column1 as SS1
Step 2: Calculate sum of square of column2 as SS2
Step 3: Calculate Square of mean of column 1 as m1
Step 4: Calculate Square of mean of column 2 as m2
Step 5: Calculate square root of SS1-m1 as SQ1
Step 6: Calculate square root of SS2-m2 as SQ2
Step 7: Calculate denominator as DR as SQ1 * SQ2
Step 8: Calculate sum of column 1 as sum1
Step9: Calculate sum of column 2 as sum2
Step10: Calculate product of sum1 and sum2 as TP
Step11: Calculate Mean product as MP as TP/ N
Step12: Calculate sum of product of all rows as PS
Step13: Calculate nominator as NR as MP*PS
Step14: Calculate Pearson co-efficient as NR / DR
Step15: Return Pearson Coefficient.
Step16: Stop

IV. RESULTS PERFORMANCE

1.System output:

- Create Account

First user creates his profile by entering his/ her personal data like first name, password, Full name, Gender, Address, and Email id etc. and user enters his username and password to login into the account for uploading and downloading images on/from the web server. Create account from sign up. To access application, user has to create his account with website; its more like an sign up form .User account is created from users name and password. Validations are done on password field if in case user enters wrong password, it asks for incorrect password again. After valid password user account will get created along with security keys. User will be notified with User ID to login with. This User ID will create his account i.e. folder in HDFS. Information about login i.e. user id and passwords are stored in HBase table which is also a part of Hadoop. It is a database which allows creating table and storing data in it. It is scalable database which can expand at run time. After successful creation of account, User now will go to Log In screen to log in. Fig 4 shows the Create Account Screen. Fig 5 shows login screen.

Fig 4 User create the Signup

Fig 5 Login screen

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

After logging in, user will see a screen having user profile .User profile having two links. One is picture and second is profile picture. When user want post the picture to different user .then click the picture link. When user want upload the picture him/her Profile then click the profile picture.

- Upload Process

When user selects upload images link on screen, he is directed to a page where he can specify the folder path of total images. All the images present in the folder gets uploaded on HDFS . At time of each upload, Map-Reduce process will run. It splits an image into 16 small images, and calculates feature vector of every small image i.e colour moment of each split image. This feature vector is nothing but a colour moment of an image. In upload image screen user see the single copy of images who have upload the picture. It contains images of different formats with variable sizes. Images shown are having formats like .jpg, .png, .bmp, .gif and .wbmp. Following is set of images uploaded to system. Fig 6 shows Upload images in storage server.

Fig 6 shows Upload images in storage server

- Result Analysis

Fig 7 shows the result Time performance Graph.Identification of the duplicate images for different number of images through different number of runs.

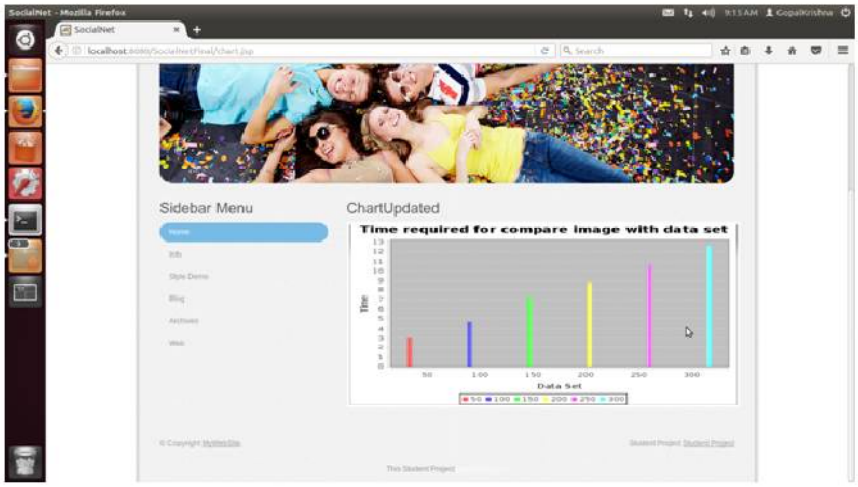

Fig 7Time performance Graph

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

V. CONCLUSION AND FUTURE WORK

The proposed work duplicate image identification system for avoiding duplicate images from storing on the storage server using Map/Reducer technique. Map/Reduce technique is used for fasten duplicate image identification process, and Pearson correlation technique is used for duplicate image identification based on image parameter's. This system reduce the time required to duplicate image identification using map reducing technique that is been powered with correlation technique.

REFERENCES

1.Hashmi .S.Taslim 'Data mining with big data Image De-Duplication in on line social networking website 'International journal of Computer Science and Application vol8 , No.2, pp-15-19, Apr-June 2015
2. I. M. jayasree, \Data mining: Exploring big data using hadoop and map reduce," Vol 4,Special issue, 2013.
3. U. H. Jeffrey. Dean, \Map reduce: Simplified data processing on large cluster," 2003.
4. L. S.chris, L.Liu, \Hipi:ahadoop image processing interface for image-based map-reduce tasks," 2011.
5. M. M. Condro Wibawa, Irwan Bastian, \Document similarity measurement using ferret algorithm and map reduce programming model," Jan-2015.
6 G. N. Nayakam, \Study of similarity coefficients using map/reduce. Programming model, "The Supervisory Committee certifies, 18, 2012.
7. e. a. Sheu, Ruey-Kai, \Design and implementation of file de-duplication framework on hdfs," International Journal of Distributed Sensor Networks2014, 2014.
8. e. a. Dyer, Christopher, \Fast, easy, and cheap: Construction of statistical machine translation models with Map-reduce." Proceedings of The Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008.
9. S. M. Stupar, Aleksandra and R. Schenkel, \Rank reduce processing k-nearest neighbour queries on top of Map/reduce." Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval, 2010.
10. P. G. Manikantan U.V, \A survey on data de-duplication in cloud storage environment," (IJSRET), ISSN 2278 - 0882 ,Volume 4, Issue 4, April 2015.
11. A. F. D. Gillick and J. De Nero, \Map/reduce: Distributed computing for machine learning," Berkley.
12. e. a. Sweeney, Chris, \Hipi: a hadoop image processing interface for image-based map reduce tasks," Chris University of Virginia,, 2011.
13. C. Balkesen and NesimeTatbul, \Scalable data partitioning techniques for parallel sliding window processing over data streams," International Workshop on Data Management for Sensor Networks (DMSN), 2011.
14. E. Mohammed, \Applications of the map reduce programming framework to clinical big data analysis: current landscape and future trends," International Journal of Computer Science and Emerging Technologies (IJCSET) 40, Volume 1 Issue 2, 29 oct,2014.
15. e. a. Foo, Jun Jie, \Detection of near-duplicate images for web search," Proceedings of the 6th ACM international Conference on Image and video retrieval. ACM, 2007.

BIOGRAPHY

Hashmi .S.Taslim is a pursuing master degree in computer science in G.H.Raisoni, Engg college ,Jalgaon, MS, India