International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 7 96 - 100

Framework for Map Reducing Technique Using Correlation for Duplicate Image Identification Process

Mr. Deshmukh Amol Sahebrao
Department of Computer engineering of JSCOE
Handewadi Road, Hadapsar,
Pune-411028, India
amoldeshmukh055@gmail.com

Prof. P. D. Lambhate
Department of information technology of JSCOE
Handewadi Road, Hadapsar,
Pune-411028, India
pinu_poonam@yahoo.co.in

Abstract-The duplicate image identification is an image deduplication System which avoids duplicate copies of images from storing in the storage server and reduces Storage space. This technique is used to improve storage utilization by avoiding duplicate images to store in storage server and reduce the time complexity by using Map Reduce technique. With explosive growth of digitization bulk of digital data may uploaded on server every day, deduplication schemes are widely used in backup and recovery System to minimize network and storage overhead by detecting and avoiding redundancy among data. Traditional deduplication schemes work if and only if the second image having the same content as first, so this restricts the performance of many applications as exact images need to be there if want to succeed and these all schemes are suffering from huge time complexity problem to deal with huge amount of data. In this paper, we propose the duplicate image identification system using MapReduce technique which improves the scalability and efficiency of system. Our approach reduce the time required to identify the duplicate image in storage server using MapReducing technique that is been powered with correlation technique.

Keywords*—Duplicate image identification, Deduplication, Data partitioning, MapReduce, Pearson Correlation, Performance evaluation and Optimization;*

*****

I. INTRODUCTION

The amounts of the images being uploaded daily on different web servers are increasing tremendously. By the survey performed on 2010, the 2.5 billion new images are being stored to the Facebook daily. However the system making use of all this data are observed of having some delay as the operational data is of huge amount. 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 MapReduce platform used for the parallel computing of the data. MapReduce 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 deduplication is a well-known technique of reducing the size of data storage by preventing the storage of identical files. A traditional deduplication 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 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 MapReduce technique. In this paper we are presenting cost-effective monitoring solution. For monitoring application, to reduce the time required to identify the duplicate image in storage web server using map reducing technique that is been powered with correlation technique.

II. RELATED WORK

As the large amount of duplicate images are available on the web, web search for a particular images tends to give the number of nearby images also which reduce the performance of the system.[1][3] Elaborates the method of finding the nearby images. To achieve the task to queries which are being popular are taken and the commercial search service to collect the images which are normally analyse as nearby images. As the removing such nearby images from the repository is practically not possible hence the proposed work removes the nearby images from the search answers. By evaluating the technique with many real world queries it had been found that the system gives the promising results compare to the traditional techniques under the same kind. To bring down the idea into reality (DPF, PCA-SIFT, and HBC) algorithms are being used which considerably performs better than the other.

As discussed above existing deduplication systems are operated well if and only if the images to be compared are having same underlying bit codes. But this scheme reduces the usage of applications. So to overcome this [4] presents a novel system of image deduplication which makes usage of high precision duplication approach. This system comprises of five phases as feature extraction, high-dimension indexing, correctness optimization, centroid selection and deduplication evaluation by computing the system on real image datasets it had been noticed that system not only gives the efficient image deduplication technique but also greatly improves the precision of duplicate image retrieval.

[5]Explains HIPI is an image processing library on the map reduce framework. The modeling of library is completed in such a

way that it hides the implementation of complex hadoop map reduce framework and focus more on image as it is the thing about which users worrying a lot. The implementation is done by considering large amount of data, because of this system gives higher throughput in case amount of images exceeds. Map reduce pipeline has arrangement of different formats for accessing the images. The kinds of images that can be used during map reduce steps are filtered by using the culling phase during the mapping phase. Float images, most important part in image processing are obtained by using the encoders and decoders techniques. By adding all these features in the system it provides simplified interface to deal with the images on MapReduce.

[6]Explain the deduplication framework based on HDFS by designing the techniques such as RFD-HDFS and FD-HDFS. RFD-HDFS is best suit for the applications which are related with the finance where there are no chances of flaw whereas FD-HDFS can be used in applications which takes few amounts of errors. The experimental evaluation shows that space allocated by duplicate data is reduced incredibly and the performance of the uploaded files are affected by the integrated schemes. In order to shows the performance of different searching and sorting task on the system having different configurations a useful survey is explained by the [9]. To bring down this idea into existence hadoop and map reduce technique for distributed data processing is used. Here the machine learning complicated classes are separated within the map reduce framework to improve the utilization 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 drawbacks for the complex operations over large database.

[12] Presents a flexible data partitioning techniques for the purpose of data streams processing. Estimated schemes used for the dividing of data are failed to achieve high degree of scalability which reduces the performance of the system and so increase time and cost complexity of the system. So to overcome the problem [12] finds a good technique. Here to different partitioning techniques are proposed i.e. partitioning based on batch and pane based partitioning. Out of the number of techniques explained above, pane based partitioning gives good result. To show the experimental performance the system is tested across the linear load benchmark. Also it gives fewer loads on the load which is dividing the data. The current work does not trouble about the fault tolerance of the system. This issue is kept as a future work by the writers.

As the world is facing the problem of managing the large amount of data, numbers of techniques are proposing to get eliminating of these. Recent survey conducted by IBM shows that near about 2.5 quintillion bytes of data are being daily produced. This data consist of many formats like images, videos, social media site opinions, sensor data, transactional data etc. it is impractical to deal with this data. Hence from last decade MapReduce is emerging as promising framework to deal with this large amount of data. [13]As the main benefit of the MapReduce is to give scalable applications, it has been used in many levels from academics to the industry. Authors try to explain the complete theories behind the map reduce.

III. EXISTING SYSTEM

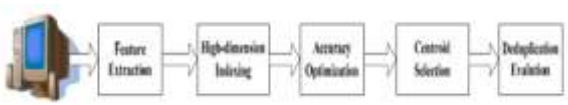

Fig.1. Images Deduplication System Architecture

In this traditional deduplication system Corel datasets are used which contains a large number of similar image groups. This dataset is an input of traditional system. The work is separated into five stages as feature extraction, high-dimension indexing, accuracy optimization, centroid selection and deduplication evaluation by evaluating the system on real datasets. It had been observed that system identify the duplicate images but eventually this system is directly proportional to the number of the images. So there is an urge of a system is required to reduce the duplicate identification time.

IV. PROPOSED METHODOLOGY

In this paper the focus is on the possible strategies of duplicate image identification. The basic idea of this proposed system comes from the fact that storage servers are big platform to store and to retrieve the data in huge capacity. Where there is greater possibility of duplication of the data can be happen due to 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 the main purpose of this proposed system is to reduce the time required to identify the duplicate image in storage server using map reducing technique that is been powered with correlation technique.

4.1 System Overview

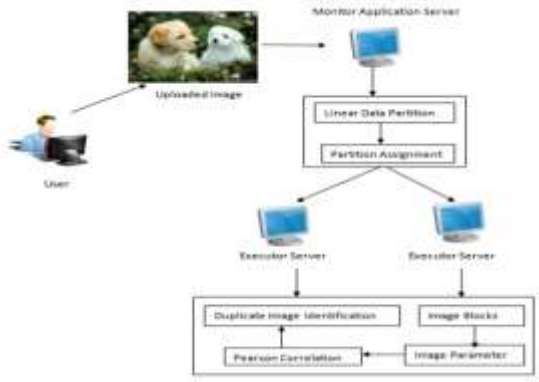

Fig.2. Proposed System Architecture

Here we describe our Framework for Map reducing technique using correlation for duplicate image identification process in the above figure with the following steps:

Step 1: Linear Data Partitioning

Here 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

Here 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

$f(BV) = V_s \bmod B;$

Where, BV is block vector.

Step 3: Image Parameter With Entropy Matching

Entropy Matching: Entropy matching step contains mainly 2 steps as described below

3.1 Mean and Standard Deviation calculation

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

which represents the some brightness of the image in that pixel. If mean of an image is high then it means that the image is bright and if mean is low then it means that the image is dark. The standard deviation in image is also calculated by using the mean and each pixel values. The standard deviation reveals something about the contrast of image in particular blocks. If standard deviation is high then it shows the high contrast of image in a particular block. If standard deviation is low then it will show the low contrast in image of a particular block.

The mean and standard deviation can be calculated using the following equations respectively.

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} x_{ji} \quad (1)$$

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2} \quad (2)$$

This can be represent with the following algorithm

Algorithm 1: Mean and Standard Deviation

Input: Image File

Output: Mean and Standard deviation Values.

0: Start

1: Get Image path.

2: Get Height and width of the Image (L*W).

3: Declare MR=0, MG=0, MB=0

4: FOR x=0 to width.

5: FOR y=0 to Height.

6: Get a Pixel at (x, y) as signed integer.

7: Convert pixel integer value to Hexadecimal to get R, G, and B.

8: MR=MR+R , MG=MG+G, MB=MB+B

9: End Inner FOR

10: End Outer FOR

11:MR=MR/(L*W),MG=MG/(L*W),MB=MB/(L*W)

12: Declare VR=0, VG=0,VB=0

13: FOR x=0 to width

14: FOR y=0 to Height

15: Get a Pixel at (x, y) as signed integer

16: Convert pixel integer value to Hexadecimal to get R, G, and B.

17: VR= VR+ (R-MR)* (R-MR)

18: VG= VG+ (G-MG)* (G-MG)

19: VB= VB+ (B-MB)* (B-MB)

20: End Inner FOR

21: End Outer FOR

22:VR=VR/(L*W) , VG=VG/(L*W) , VB=VB/(L*W)

23:SR= SQRT(VR) , SG=SQRT(VG) , SB=SQRT(VB)

24: Stop

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, that means the minimum difference of the entropies must be within 5.

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

$$r = \frac{\sum x_i y_i - \frac{\sum x_i y_i}{N}}{\sqrt{\left(\sum {x^2}_i - \frac{(\sum x_i)^2}{N}\right)}\sqrt{\left(\sum {y^2}_i - \frac{(\sum y_i)^2}{N}\right)}} \quad (3)$$

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

Algorithm 2: Pearson Correlation

Input: Two parameter matrix of N rows and 2 columns and Let matrix be M

Output: Pearson factor (i.e. in between 0 to 1)

0: Start

1: Calculate sum of square of column1 as SS1

2: Calculate sum of square of column2 as SS2

3: Calculate Square of mean of column 1 as m1

4: Calculate Square of mean of column 2 as m2

5. Calculate square root of SS1-m1 as SQ1

6: Calculate square root of SS2-m2 as SQ2

7: Calculate denominator as DR as SQ1 * SQ2

8: Calculate sum of column 1 as sum1

9: Calculate sum of column 2 as sum2

10: Calculate product of sum1 and sum2 as TP

11: Calculate Mean product as MP as TP/ N

12: Calculate sum of product of all rows as PS

13: Calculate nominator as NR as MP*PS

14: Calcualte pearson coefficient as NR/DR

15: Return Pearson Coefficient.

16: Stop

B. Mathematical Model:

The whole proposed system is expressed mathematically in the below model.

1. S= { } be as system for Map reducing for Duplicate Image Identification

2. Identify Input as U={ U_1,U_2, U_3…..U_n};

Where U_n= Uploaded Image

3. Identify D as Output i.e. Duplicate Image Identification

S= {U_n, D};

4. Identify Process P

S= {U_n, D, P};

P= {D_p, I_p, P_c, C_c};

Where, D_p= Data Partition

I_p =Image Parameter

P_c= Pearson Correlation

C_c= Correlation Comparison

5. S = {U_n, D_p, I_p, P_c, C_c, D};
The union of all subset of S Gives the final proposed system.

1) Image blocks can be formed by the following equation:

$$f(BV) = V_s mod B;$$

Where, BV is block vector.

2) Entropy is defined as

$$E=\log(\delta);$$

Where, δ is the standard deviation of the image.

V. RESULT DISCUSSION

To show the effectiveness of proposed system some tests are conducted on java based windows machine using apache tomcat as server. To measure the performance of the system we set the bench mark by selecting the data set of 150 images for duplication removal process using Map reducing technique. To determine the performance of the system, we examined how many relevant images are identified as duplicates based on our Map reducing technique approach.

To measure this precision and recall are considering as the best measuring techniques. So precision can be defined as the ratio of the number of relevant images are identified as duplicates to the total number of irrelevant and relevant images are identified as duplicates. It is usually expressed as a percentage. This gives the information about the relative effectiveness of the system.

Whereas Recall is the ratio of the number of relevant images is identified as duplicates to the total numbers of relevant images are not identified as duplicates and it is usually expressed as a percentage. This gives the information about the absolute accuracy of the system. The advantage of having the two for measures like precision and recall is that one of these is more important than the other in many circumstances.

So, $Precision = \frac{A}{A+C} \times 100$

and $Recall = \frac{A}{A+B} \times 100$

Where;

A = The number of relevant images are identified as duplicates,

B = The number of relevant images are not identified as duplicates, and

C = The number of irrelevant images are identified as duplicates.

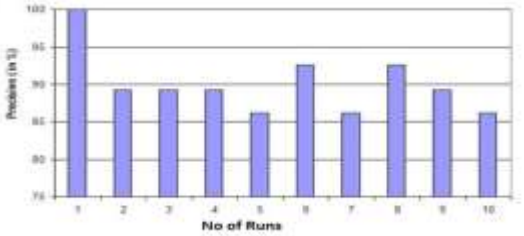

Fig.3. Average precision of the proposed approach

In Fig.3. we observe that the tendency of average precision for the images are identified as duplicates is more than the average of the other Map reducing techniques.

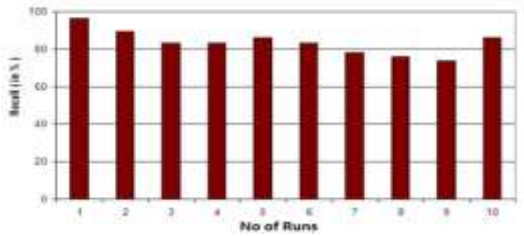

Fig.4. Average Recall of the proposed approach.

In Fig.4. We observe that the tendency of average Recall for the images are identified as duplicates is more than the average of the other Map reducing techniques.So this shows that our proposed system is achieving high accuracy than any other method.

In another experiment of evaluating performance of our system, System is measured for performance time for image deduplication using map reducing technique and without using map reducing technique. And the result obtained is depicted in the below table 1.

Sr No	Time in (MilliSecond) Without Mapreduce	Time in (Milli Seconds) With Mapreduce
1	254	150
2	234	104
3	253	120
4	275	132
5	276	151
6	302	157
7	327	178
8	377	193

Table 1: Time comparison table

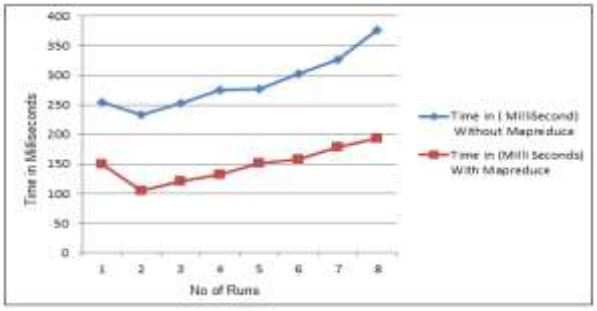

Fig.5. Time performance Graph

The above graph in the Fig. 5. drawn for identification of the duplicate images for different number of images through different number of runs. This plot clearly indicates that time has been drastically reduced for number images for map reducing technique incorporated using two servers with mongoDB NOSQL database.

VI. CONCLUSION AND FUTURE WORK

A. Conclusion

The work was stimulated for improving storage utilization and reduces the time required to duplicate image identification. However many systems are already existed to identify the duplicate images but eventually all are directly proportional to the number of images. The proposed work duplicate image identification system for avoiding identical images from storing on the storage server and improve storage utilization. MapReduce 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.

B. Future Work

The duplicate image identification system with map reduce framework can be further enhanced with multiple features like Image saving, Searching, authentication and image extraction. We can implement one single tool which handles all these features.

ACKNOWLEDGEMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule University of Pune and concern members of cPGCON2016 conference, organized by, for their constant guidelines and support. We are also thankful to the reviewer for their valuable suggestions. We also thank the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] Jia Xu, Bin Lei, Yu Gu, M Winslett. "Efficient Similarity Join Based on Earth Movers Distance Using MapReduce.", IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 8, August 2015.
[2] Sandeep B. Aher, Poonam D. Lambhate, "Real time monitoring of system counters using Map Reduce Framework for effective analysis", IJCET, Vol.5, No.4 August 2015.
[3] Foo, Jun Jie, et al., "Detection of near-duplicate images for web search", Proceedings of the 6th ACM international conference on Image and video retrieval. ACM, 2007.
[4] Chen, Ming, Shupeng Wang, and Liang Tian, "A high-precision duplicate image deduplication approach", Journal of Computers, 8.11 (2013), 2768-2775..
[5] Sweeney, Chris, et al. "HIPI: a Hadoop image processing interface for image-based mapreduce tasks", Chris. University of Virginia, (2011).
[6] Sheu, Ruey-Kai, et al. "Design and Implementation of File Deduplication Framework on HDFS", International Journal of Distributed Sensor Networks 2014 , (2014).
[7] Neelaveni, P., and M. Vijayalakshmi, "A Survey on Deduplication in cloud storage.", Asian Journal of Information Technology, 19.6 (2014): 320-330.
[8] Dean, Jeffrey, and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters", Communications of the ACM 51.1 (2008): 107-113.
[9] Gillick, Dan, ArloFaria, and John DeNero, "Mapreduce: Distributed computing for machine learning.", Berkley, Dec 18 (2006).
[10] Dyer, Christopher, et al., "Fast, easy, and cheap: Construction of statistical machine translation models with MapReduce", Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008.
[11] Stupar, Aleksandar, Sebastian Michel, and Ralf Schenkel, "RankReduce processing k-nearest neighbor queries on top of MapReduce.", Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval. 2010.
[12] Balkesen, Cagri, and NesimeTatbul, "Scalable data partitioning techniques for parallel sliding window processing over data streams.", International Workshop on Data Management for Sensor Networks (DMSN), 2011.
[13] Sakr, Sherif, Anna Liu, and Ayman G. Fayoumi. "The family of MapReduce and large-scale data processing systems.", ACM Computing Surveys (CSUR) 46.1 (2013):11.

Author Profile:

Amol S. Deshmukh, is currently pursuing M.E (Computer Engineering) from Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India. Savitribai Phule, Pune University, Pune, Maharashtra, India-411007. He received his B.E. (Information Technology) Degree from BIGCE, College of Engg. Solapur, Maharashtra, India - 413255 in 2014. His area of interest is Cloud Computing and Parallel Computing.

Prof.P.D.Lambhate, received her Degree from WIT, Solapur. ME(Comp) from BVCOE Pune, Pursing PhD. In computer Engineering.She is currently working as Professor at Department of Computer and IT, Jayawantrao Sawant College of Engineering, Hadapsar, Pune, India 411028, affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India-411007. Her area of interest is Data mining, search engine.