Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 09, September-2019

Overview of Content based Image Retrieval using Map-Reduce

Tapas Bhadra, Shachi Sonar, Samruddhi Zagade
Department of Information Technology
K.J.Somaiya College of Engineering,
Vidyavihar

Abstract:- **With the exponential increase in the amount of multimedia data, data storage and retrieval has become a big challenge. It has become increasingly difficult to query and retrieve result relevant to user's demand with efficiency and accuracy. Previously, the search of image data was done by using keywords or description of data which failed to produce expected results. The distributed data model of Hadoop, an open-source software, provides us with a solution to this problem by using an image as query input and Map Reduce algorithms for processing. This paper intends to discuss the Content based Image Retrieval using Map-Reduce Algorithm.**

I. INTRODUCTION

With the growing amount of media, especially images, produced through mobile phones, cameras and over the world wide web, the problem of storing and retrieving relevant data has become a real issue. The advances made in the field of digital technology has led to the explosion of the volume of images created.

Earlier, text-based image retrieval systems existed. Metadata in the form of captions and keywords were added to every single image in the database. While this was a viable solution earlier, the exponential growth of digital media has made this an impossible task. Feature extraction is another popular method available for the retrieval of images. This approach though appealing has its cons. Extraction of features can become extremely complicated with large databases. Furthermore, massive computational power is required to run the particular algorithms. This complexity of retrieval of images coupled with the large processing power required for the huge databases presents organizations with a tough challenge. To solve this difficulty regarding image retrieval, Hadoop's Map Reduce functionality plays a vital role. Here, rather than using keywords to retrieve the desired image, certain features of the images like colour, texture, facial features are used. This entire process is known as content based retrieval. Through Hadoop's Map reduce framework's parallel functionality, the processing is sped up considerably. Along with this, through the reduction of features, large databases are stored in a compact manner for retrieval.

II. RELATED WORK

Dewen Zhuang et al[2] has proposed the relevance feedback method to reduce the semantic gap. Image feature dimensionality reduction was performed utilizing the linear discriminant analysis. It diminishes the semantic gap and the storage of image signatures, along with improving the retrieval efficiency and performance. However, the performance is low for a few of the categories. Furthermore, the extraction of visual effect features and measurement of regional similarity still needs to be worked upon. CBIR for JPEG pictures has pulled in numerous individuals' attention and a sequence of algorithms directly reliant on the discrete cosine transform area have been formulated. To exploit these DCT coefficients while considering the color and texture data for the retrieval of JPEG formatted pictures, performing CBIR winds up being productive. Here, decompressing the images and after that, processing in the spatial domain is carried out. The feature vectors are then figured out from a few of the DCT coefficients. This activity is performed in the partially decoded domain. It can incredibly useful in reducing retrieval complexity.

The two methods that we'll be analyzing are given in the next section.

2.1 Text based Image Retrieval

Text-based image retrieval can be done on the basis of the description, keywords as well as text that is available in the image through metadata such as captions or subtitles or any related text. It is the traditional method of searching for an image by describing its most prominent characteristics. Most of the systems that are used for image retrieval take text input, however, other than manual input, there is no text description of an image stored in the database. Metadata of an image can be generated by putting down descriptions and keywords associated with the image. By doing so, one can understand and retrieve the humongous amount of image data while indexing it. Keywords can then be used to search the data on the SE and they form the characteristics of the text on various websites. If text information is stored as metadata for every image, keywords can be used for retrieval of the images easily. The process of storing keyword data has to be done manually for each image and looking at the tremendous amount of image data present on the world wide web it is just impossible for humans to do. Another huge drawback of this technique is that images are highly subjective in nature, making their perceptions unique for everyone.. Therefore, assigning annotations to images becomes wasteful as well as brings about inconsistency.

2.2 Content Based Image Retrieval

CBIR which is Content-Based Image Retrieval is a technique which presents us with the technologies that allow us to organize images depending on their visual features. It presents us with the most visually similar images with respect to the query image through the help of various technologies like Computer Vision. It is frequently known as Query by Image Content which provides a solution for the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 09, September-2019

Image Retrieval problem in large databases, where otherwise querying is a very tedious and time-consuming task. It extracts the most relevant characteristics and uses similarity measures to find related images. It's different from text-based image retrieval where keywords are used to search the images as compared to the image characteristics that are used in this case. Querying, indexing, searching and matching techniques are also used in addition. CBIC uses color percentage, color layout, texture, shape, location, and keywords to provide the most visually similar and sound images with respect to the query. It measures the distance between the features extracted and accordingly displays the image with the minimal amount of distance by the help of histogram similarity. Later, it creates a similarity matrix with the values and selects the value that is the least amongst the others.

III. HADOOP AND MAP REDUCE

Hadoop is an open source software used for the data handling of large databases. It uses the distributed data framework for the storage as well as processing of data. The data types that can be handled by Hadoop can either be structured, unstructured or semi-structured. Hadoop used HDFS which is Hadoop Distributed File System for storage of data. For the computation of such large data, The MapReduce framework consists of two phases, Map and Reduce. In the map phase, data that is stored by splitting at various locations is provided as input to a function which then produces the key value pairs. These sets of key value pairs then form the input for the next phase and over each pair, a user-defined function is then executed in order to produce a set of intermediate key value pairs. In the reduced phase, the aforementioned intermediate key value pairs form the input and groups are determined as per the key, while values are consolidated according to the reduced algorithm that is to be provided by the user. HDFS is one of the key aspects of Hadoop. In HDFS, the storage data is split and stored as datanode and the metadata pertaining to this data is stored as namenode. Namenode is the master which stores all information regarding metadata of the data and Datanode is the slave which stores the actual data. Further, data kept in Hadoop is reliable as multiple copies of it are stored for security and backup.

IV. SYSTEM ARCHITECTURE

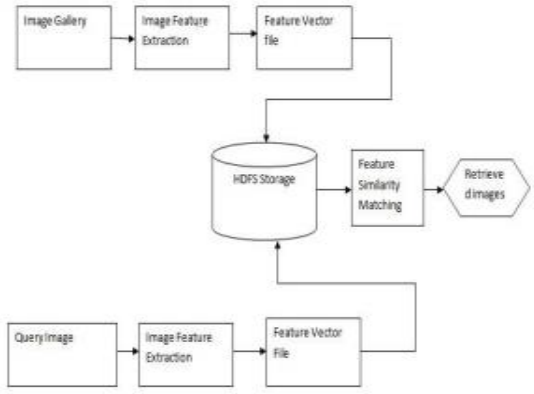


Fig-1: System Architecture

The admin or the host of the system will upload the image and all the features related to the image are extracted such as the color, shape and texture. These features are stored in the feature vector with the help of Hadoop MapReduce. When the user provides a query, the features of the query are extracted and they are compared with the feature vector in the database and the one with the minimum distance is selected for the user.

VI. METHODOLOGY

Content Based Image Retrieval usually consists of two steps, the extraction of features as well as their mapping. While feature extraction is concerned with the accuracy achieved during retrieval, the feature matching is concerned with the efficiency and speed. As the features are found in high dimensions, so is the searching. There are numerous approaches to scan high dimensional space, for example, linear scanning, tree searching, vector quantization, and hashing. Between the mentioned techniques, hashing comes out to be the most effortless approach to limit time complexity as O(1) while devising a fuzzy search strategy. Indeed, even feature matching is upgraded by means of hashing; because of the immense volume of data for CBIR, the time complexity nevertheless persists to be extravagantly high. Primarily, in the age of the explosive expansion of digital data, the stand-alone techniques for CBIR start getting increasingly hard to maintain due to the dissatisfaction from the load of capacity and processing.

Amongst every one of the available modules, 'Feature Extraction' and 'Feature Matching' are the most tedious and dilatory.

The process is given by:

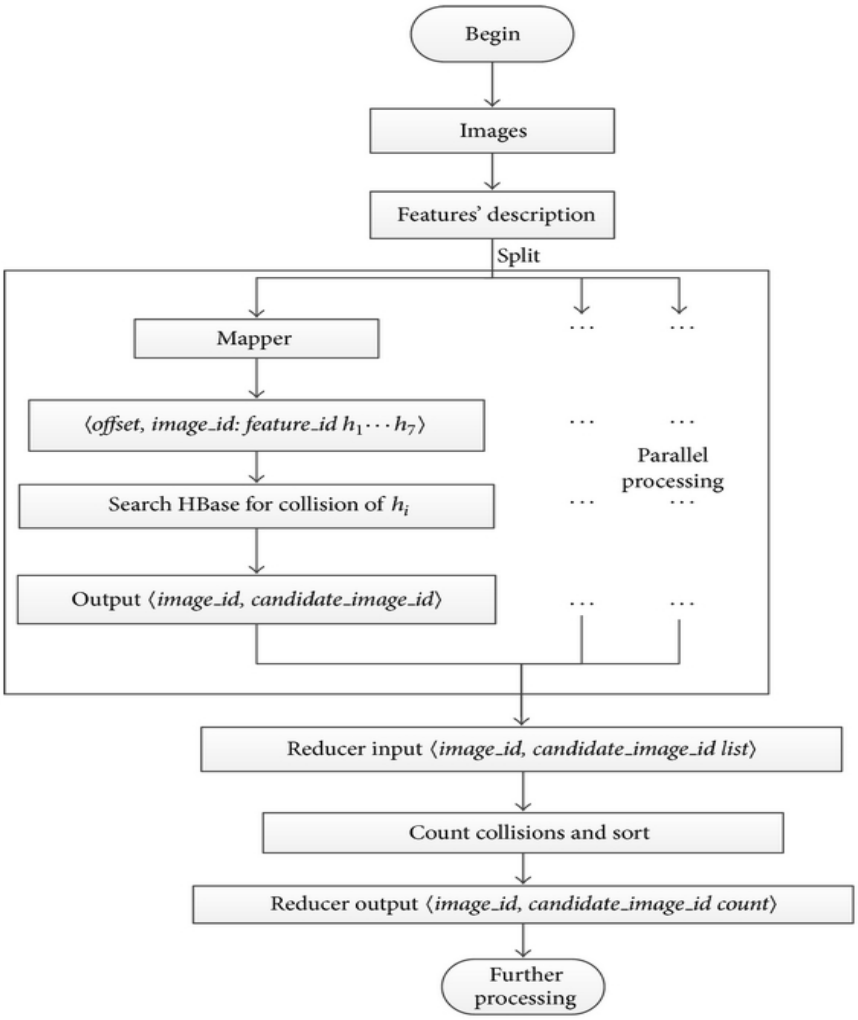
1. Image pre-processing: This includes image scaling as well as converting the image to grayscale
2. Feature Extraction: Here, multiple vectors are captured in 64 dimensions.
3. Hashing: The outputs from steps 1 & 2 are hashed with each other to produce 7 hash codes.
4. Feature Matching: Every hash code is then attributed to their individual hash table after which the features obtained are matched using MapReduce. Further, the results are saved.

6.1 Feature Matching

Potential features need to be extracted and matched. That is exactly the purpose of this module. All the candidate matching features are searched for and the similar images are chosen according to their match count. For ease, we're matching two images based on their hash code. This step takes up the maximum amount of time and is implemented with the help of Hadoops MapReduce framework.

If we take {K, V} to be the Key-Value pair in MapReduce, at that point, the work process of a query in parallel can be depicted as follows:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 09, September-2019



The practicality of this parallelization depends on two factors. Firstly, the divisions or splits made in the data have to be completely independent of each other. Meaning, that no interrelation must exist between the two entities. The input to the mapper, as well as the features' description, must follow the same format. Every line would contain the information about a particular feature's hashcode along with its unique image ID which are both independent. Secondly, we assume that all the outputs from the mappers would be aggregated by the MapReduce's reducers. As the number of reducers is limited to one, the results from the parallel processing will be collected and counted to a single reducer. The types of features that are matched are:

1. *Color*

Color histograms of the images in the database as well as the queried image are created. A distance measure using the similarity matrix of colors is computed by checking the proportion of pixels within the specified threshold value.

2. *Texture*

In this technique, the spatial definitions and the visual patterns of the images is taken into account. Depending upon the number of textures available, sets are prepared. The textural characteristics are represented in statistical approaches and structural approaches[4].

3. *Shape*

Shape could be thought of as the surface definition of a particular image. It has certain contours and outlines. Through shapes, you can distinguish a region from its surroundings. Fourier transform and Moment Invariants are few of the available shape descriptors.

VII. CONCLUSION

The constantly growing nature of digital images poses a challenge for organizations to carry out image retrieval for their problems. Methods like text based retrieval fail to keep up with this growth of media. Content Based Image Retrieval is a strong candidate as a solution to this problem of image retrieval. Hadoops ability of parallel processing not only speeds up the process, but the MapReduce framework helps with one of the most challenging aspects of Image Retrieval. By the usage of various algorithms like KMeans, Fuzzy C means Clustering, Convolutional Neural Networks, as well as Support Vector Machine, a comprehensive and complete system for Image Retrieval can be created.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 09, September-2019

REFERENCES:

[1] KUSUMA.B, 2MEGHA.P.ARAKERI. Survey On Content Based Image Retrieval Using Map reduce Over Hadoop 5 SURVEY ON CONTENT BASED IMAGE RETRIEVAL USING MAPREDUCE OVER HADOOP

[2] Gao Li-chun and Xu Ye-qiang. Image retrieval based on relevance feedback using blocks weighted dominant colors in MPEG-7. Journal of Computer Applications.vol.31(6), pp.1549-1551, 2011.

[3] Chunhao Gu Yang Gao. A Content-Based Image Retrieval System Based on Hadoop and Lucene

[4] Techniques of Content Based Image Retrieval: A Review Sheetal A. Wadhai 1 , Seema S. Kawathekar 2

[5] A New Approach for Large-Scale Scene Image Retrieval Based on Improved Parallel -Means Algorithm in MapReduce Environment Jianfang Cao,1 Min Wang,2 Hao Shi,2 Guohua Hu,1 and Yun Tian1