Retrieval of Images & Text using Data Mining Techniques

Kiran P. Khandare[1], Swati V. Kamble[2], Shivani R. Jaiswal[3], Nutan P. Khelkar[4]

[1,2,3,4] *Student, Department of CSE, Des'scoet, Dhamangaon Rly, Maharashtra, India*

---***---

Abstract - *In the domain of Image processing, Image mining is advancement in the field of data mining. Image mining is the extraction of hidden data, association of image data and additional pattern which are quite not clearly visible in image. Data mining refers to the extracting of knowledge /information from a huge database which is stored in further multiple heterogeneous databases. Knowledge/ information is communicating of message through direct or indirect technique. These techniques include neural network, clustering, correlation and association. This writing gives an introductory review on the application fields of data mining which is varied into telecommunication, manufacturing, fraud detection, and marketing and education sector. In this technique we use size, texture and dominant colour factors of an image. The image mining is new branch of data mining, which deals with the analysis of image data. There is several methods for retrieving images from a large dataset. But they have some drawbacks. In this paper using image mining techniques like clustering and associations rules mining for mine the data from image. And also it uses the fusion of multimodal features like visual and textual.*

***Key Words*: Data Mining, Image Mining, Feature Extraction, Image Retrieval, Association, Clustering.**

1. INTRODUCTION

In the real world, huge amount of data are available in education, medical, industry and many other areas. Such data may provide knowledge and information for decision making. For example, you can find out drop out student in any university, sales data in shopping database. Data can be analysed , summarized, understand and meet to challenges.[1] Data mining is a powerful concept for data analysis and process of discovery interesting pattern from the huge amount of data, data stored in various databases such as data warehouse , world wide web , external sources .Interesting pattern that is easy to understand, unknown, valid ,potential useful. Data mining is a type of sorting technique which is actually used to extract hidden patterns from large databases. The goals of data mining are fast retrieval of data or information, knowledge Discovery from the databases, to identify hidden patterns and those patterns which are previously not explored, to reduce the level of complexity, time saving, etc[2].

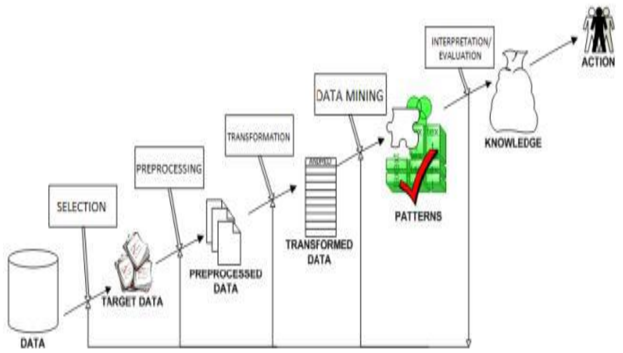

Fig.1. Knowledge Data Mining

- Selection: select data from various resources where operation to be performed.
- Preprocessing: also known as data cleaning in which remove the unwanted data.
- Transformation: transform /consolidate into a new format for processing.
- Data mining: identify the desire result.
- Interpretation / evaluation: interpret

The traditional image retrieval systems are text-based. That means the systems are using the manual annotation of images for image retrieval. But there is some limitations for text-based approach. First one is in the case of image annotation. The large volume of the databases makes this process very difficult. And this annotation is valid for only one language. Second problem arises in the human perception. Individual personal impressions and opinions about an image is different. So it makes limitations to the subjectivity of human perception. And it also make too much responsibility on the ultimate users. The third problem coming with the deeper needs. That means the queries that cannot be described at all. The solution to this problems is CBIR (Content Based Image Retrieval) systems. A single image contain a lot of information's. We can extract these contents as various content features like color, shape, texture etc. In this systems each image will be described by it's own features. The CBIR systems itself take the responsibility of forming the query away from the user. If a user wants to search for sky images, then he can submit an existing sky picture or his own sketch for sky as query. The system will extract image features for this query. It will compare these features with that of other images in a database. Then relevant results will be displayed to the user. In the CBIR systems the visual features like color, shape etc are used. But it make a "semantic gap" problem. But in the proposed system we are fusing the multimodal features. That means it use both visual features and textual features for image retrieval. This concept increase the system efficiency.

2. RELATED WORK

Image mining is the process of searching and discovering valuable information and knowledge in large volumes of data. Fig. 1 shows the Typical Image Mining Process. Some of the methods used to gather knowledge are, Image Retrieval, Data Mining, Image Processing and Artificial Intelligence. These methods allow Image Mining to have two different approaches. One is to extract from databases or collections of images and the other is to mine a combination of associated alphanumeric data and collections of images. In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction. When the input data is too large to be processed and it is suspected to be notoriously redundant, then the input data will be transformed into a reduced representation set of features. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. Several features are used in the Image Retrieval system. The popular amongst them are Color features, Texture features and Shape features.

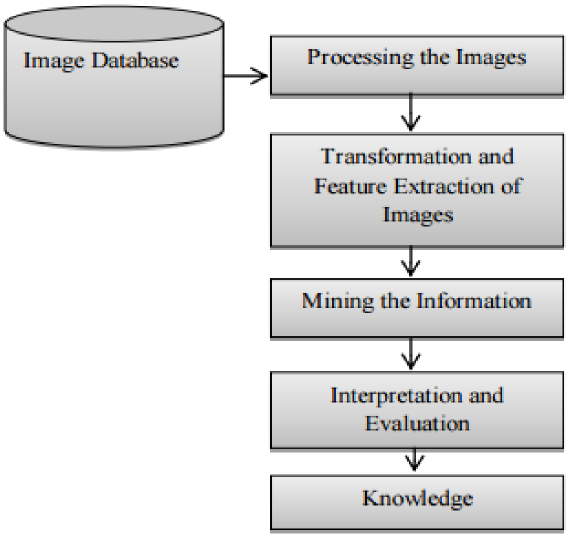
Image Database
Processing the Images
Transformation and Feature Extraction of Images
Mining the Information
Interpretation and Evaluation
Knowledge

Fig:2 Image Mining

Image mining is a new turn of data mining. It concerned with knowledge discovery in image databases. Image mining has two sections , first is mining large collections of images and the second is the combined data mining of large collections of image and associated alphanumeric data[3]. In the case of image-bases, assuming that all the images have been manually indexed or their contents classified may not be feasible. This presents one major problem from the typical data mining approach for numerical data. If the images are labeled with a semantic descriptor, then the mining can be done based on these high level concepts. But if the database contain large volume of images, this will become impossible. An alternative is to rely on automatic/semi-automatic analysis of the image content and to do the mining on the generated descriptors. For example, color, texture, shape and size can be determined automatically. In the image mining process, there is several steps as in the knowledge discovery process. Fig. 2 illustrates the image mining process. We have an image data base that contain a lot of images. First we need to preprocessing the images. Then perform transformation and feature extraction of that image. Mining the information from the extracted features. After that perform interpretation and evaluation of the information. At last we get the knowledge[1].

There are two major issues that will affect the image data mining process. One is the notion of similarity matching and the other is the generality of the application area, that is, the breadth of usefulness of data mining from a practical point of view[3]. Image mining has an important application in the area of medical imagery and patient records. To develop an accurate diagnosis or prognosis both image data like x-rays, SPECT etc. and patient data such as weight, family data etc are examined together to get interesting associations.

Data mining means collecting relevant information from unstructured data. So it is able to help achieve specific objectives. The purpose of a data mining effort is normally either to create a descriptive model or a predictive model. A descriptive model presents, in concise form, the main characteristics of the data set. The purpose of a predictive model is to allow the data miner to predict an unknown (often future) value of a specific variable; the target variable [7]. The goal of predictive and descriptive model can be achieved using a variety of data mining techniques as shown in figure 3[8].

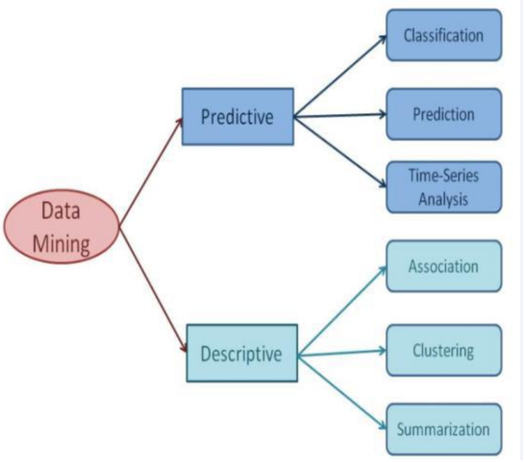

Fig.3. Data Mining Models

Classification: Classification based on categorical (i.e. discrete, unordered).This technique based on the supervised learning (i.e. desired output for a given input is known) .It can be classifying the data based on the training set and values (class label). These goals are achieve using a decision tree, neural network and classification rule (IFThen). for example we can apply the classification rule on the past record of the student who left for university and evaluate them. Using these techniques we can easily identify the performance of the student.

Regression: Regression is used to map a data item to a real valued prediction variable [8]. In other words, regression can be adapted for prediction. In the regression techniques target value are known. For example, you can predict the child behaviour based on family history.

Time Series Analysis: Time series analysis is the process of using statistical techniques to model and explain a time-dependent series of data points. Time series forecasting is a method of using a model to generate predictions (forecasts) for future events based on known past events [9]. For example stock market.

Prediction: It is one of a data mining techniques that discover the relationship between independent variables and the relationship between dependent and independent variables [4].Prediction model based on continuous or ordered value.

Clustering: Clustering is a collection of similar data object. Dissimilar object is another cluster. It is way finding similarities between data according to their characteristic. This technique based on the unsupervised learning (i.e. desired output for a given input is not known). For example, image processing, pattern recognition, city planning.

Summarization: Summarization is abstraction of data. It is set of relevant task and gives an overview of data. For example, long distance race can be summarized total minutes, seconds and height. Association Rule: Association is the most popular data mining techniques and fined most frequent item set. Association strives to discover patterns in data which are based upon relationships between items in the same transaction. Because of its nature, association is sometimes referred to as “relation technique”. This method of data mining is utilized within the market based analysis in order to identify a set, or sets of products that consumers often purchase at the same time [6].

Sequence Discovery: Uncovers relationships among data [8]. It is set of object each associated with its own timeline of events. For example, scientific experiment, natural disaster and analysis of DNA sequence.

3. PROPOSED SYSTEM

System Architecture

The image database consists of a number of images with their extracted features. The features are both visual and textual. In the data mining process we use different data mining tasks. First we extract the visual and textual features on an individual basis. Then perform clustering algorithm on the two features separately. As a result we get visual feature clusters and textual feature clusters. Then the association rules mining algorithm is performed on the fusion of these clusters. So we get many association rules. From that based on a criteria we select strongest association rules. Our training data are these strongest association rules. The proposed system receives the input query in the form of images. Because rather than a text query an image is more specific. Then perform the data mining process on the input image. So we get strongest association rules of that input image. Then we perform similarity checking with these data and training data. Then retrieve the most relevant images from the data base. The following Fig: 2 illustrates the proposed system architecture.

Mining the Data

The data mining process is the key function of this image retrieval system. The Fig:3 describes the data mining process and the description of each tasks described below.

1. Feature extraction

Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately. There is different methods for feature extraction, and different features for a single image.

Most of the web based image search engines use the textual metadata for retrieval process. It produce a lot of garbage results. And if we use the visual features only makes a semantic gap problem. So in this system both the visual features and textual features are extracted.

a. Visual features

The Descriptor is the syntactic and semantic definition of the content. The visual descriptors or image descriptors describes the elementary features of the content like shape, color, texture etc. This features collectively known as general information descriptors. The use of visual features makes the image retrieval process more efficient. The advantages of color features are they should be stable under varying viewing conditions, such as illumination, shading, and highlights. And they should have high discriminative power[11].

b. Textual features

The textual features are an important factor in the case of images. Now a days most of the images are text based. The texts in the images are may be surrounding text or human submitted annotations. In this system we focused on the SURF features of an image The SURF stands for Speeded Up Robust Feature. SURF is a local invariant interest point detector and descriptor. Based on the median value of the descriptor of the images are compared and retrieved. The SURF is a descriptor based on bags of visual words [8]. Visual Words can be represented by small parts of an image which carry some kind of information related to the features (such as the color, shape or texture), or changes occurring in the pixels such as the filtering, low-level feature descriptors. A vocabulary containing 5000 visual words was built using

SURF features from a random sample of the collection and all the images were then indexed with elements of this vocabulary. The feature space is composed of 5000 dimensions .SURF is a local invariant interest point detector and descriptor. First step is to find out an interest point in a location (detector), next the neighbourhood of every interest point is represented by a feature vector (descriptor).

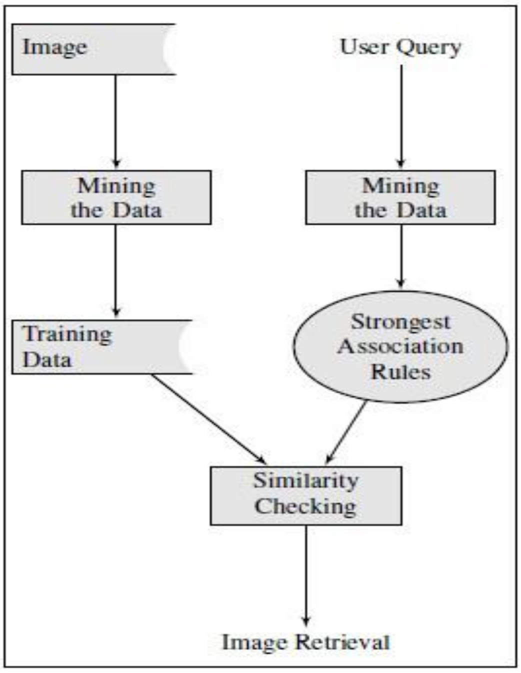

Fig.4 Architecture of image retrieval system

4. CONCLUSIONS

In this paper, we define and solve the problem of data loss. The main objective of the image mining is to remove the data loss and extracting the meaningful information to the human expected needs. This method use both textual features and visual features to create clusters and generate association rules. The method gives the ability to retrieve images that are semantically related by using the extracted visual features of the query image and by exploring the related association rules from the mining.

5. REFERENCES

[1] J. Priya , Dr. R. Manicka Chezian , " A Survey on Image Mining Techniques for Image Retrieval ", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 7, July 2013

[2] A.Kannan, Dr.V.Mohan, Dr.N.Anbazhagan, " Image Clustering and Retrieval using Image Mining Techniques", 2010 IEEE International Conference on Computational Intelligence and Computing Research Carlos Ordonez, Edward Omiecinski,"Image Mining:A new approach for Data Mining."

[3] Jiawei Han , Micheline Kamber , Jian Pei, "Data Mining; Concepts and Techniques", Reference text, Third edition.

[4] Raniah A. Alghamdi,Mounira Taileb,Mohammad Ameen," A New Multimodal Fusion Method Based on Association Rules Mining for Image Retrieval", 17th IEEE Mediterranean Electrotechnical Conference, Beirut, Lebanon, 13-16 April 2014Janani M and Dr. Manicka Chezian. R, "A Survey On Content Based Image Retrieval System", International Journal of Advanced Research in Computer Engineering & Technology, Volume 1, Issue 5, pp 266, July 2012.

[5] Aboli W. Hole Prabhakar L. Ramteke, "Design and Implementation of Content Based Image Retrieval Using Data Mining and Image Processing Techniques" International Journal of Advance Research in Computer Science and Management Studies Volume 3, Issue 3, March 2015 pg. 219-224

[6] Anil K. Jain and Aditya Vailaya, "Image Retrieval using color and shape", In Second Asian Conference on Computer Vision, pp 5-8. 1995.

[7] Harini. D. N. D and Dr. Lalitha Bhaskari. D, "Image Mining Issues and Methods Related to Image Retrieval System", International Journal of Advanced Research in Computer Science, Volume 2, No. 4, 2011.

[8] Hiremath. P. S and Jagadeesh Pujari, "Content Based Image Retrieval based on Color, Texture and Shape features using Image and its complement", International Journal of Computer Science and Security, Volume (1) : Issue (4).

[9] Brown, Ross A., Pham, Binh L., and De Vel, Olivier Y, "Design of a Digital Forensics Image Mining System", in Knowledge Based Intelligent Information and Engineering Systems, pp 395-404, Springer Berlin Heidelberg, 2005.

[10] Rajshree S. Dubey, Niket Bhargava and Rajnish Choubey, "Image Mining using Content Based Image Retrieval System", International Journal on Computer Science and Engineering, Vol. 02, No. 07, 2353-2356, 2010.

[11] Aura Conci, Everest Mathias M. M. Castro, "Image mining by Color Content", In Proceedings of 2001 ACM International Conference on Software Engineering and Knowledge Engineering (SEKE), Buenos Aires, Argentina Jun 13-15, 2001.

[12] Er. Rimmy Chuchra "Use of Data Mining Techniques for the Evaluation of Student Performance: A Case Study" International Journal of Computer Science and Management Research Vol. 01, Issue 03 October 2012.