Retrieval of images using data mining techniques

Neethu Joseph.c
Dept.of Comp.Science, Jyothi Eng.College
Thrissur, Kerala, India
neethujosephc@yahoo.co.in

Aswathy Wilson
Dept.of Comp.Science, Jyothi Eng.College
Thrissur, Kerala, India
aswathy@jecc.ac.in

***Abstract*—Data mining is an emerging research area, because of the generation of large volume of data. 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. This system produces a better precise and recalls values.**

Keywords— image mining, clustering, association rules mining, multimodal fusion

I. INTRODUCTION

In this era a large volume of electronic data is created in each seconds. The data may in the form of text, image, audio and video. That means multimedia data plays an important role in the world. Now a days the information's in the form of images takes a vital role. Because the images can easily convey more information than text. So the image retrieval systems are very relevant.

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.

This paper introduce a new image retrieval system, that integrate with the data mining techniques. So this system tie in with image mining. This paper organized as following: the next section gives the theoretical base of image mining. The third section describes the proposed system in detail. The experiment and conclusion are presented in section four and five.

II. IMAGE MINING

The availability of huge amount of data make difficulty for acquiring the knowledge. Data mining is the process of discovering knowledge from a massive amount of data. The data mining is the major step in the process of Knowledge Discovery of Data (KDD). The KDD process perform data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation[4]. We can use data for mining from different data bases, data warehouses and transactional data. Data mining tasks are mainly classified into two classes. That are predictive and descriptive. Predictive means it perform induction on the current data in order to make predictions. And descriptive means it characterize properties of the data in a target data set.

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. 1 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].

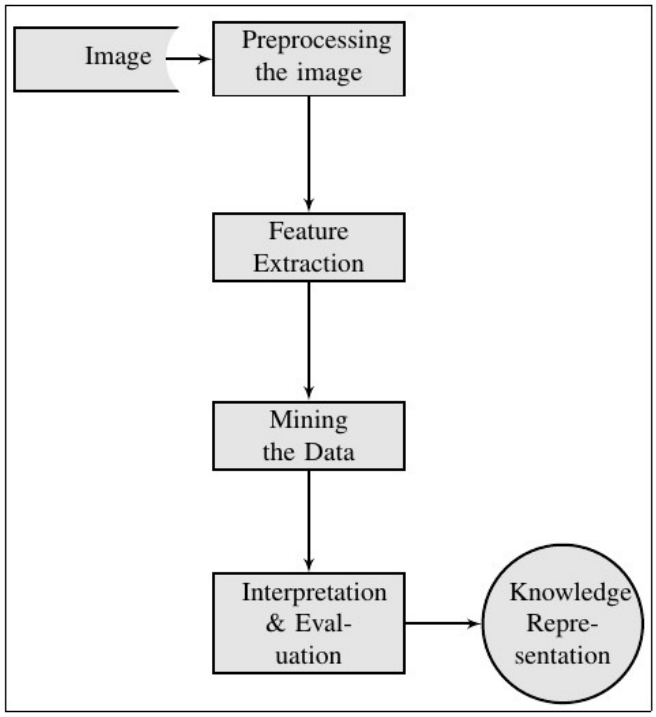

Fig:1 Image Mining

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.

III. METHODOLOGY

This section discuss about the proposed system architecture and the methods to mining the data.

A. 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.

B. 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).

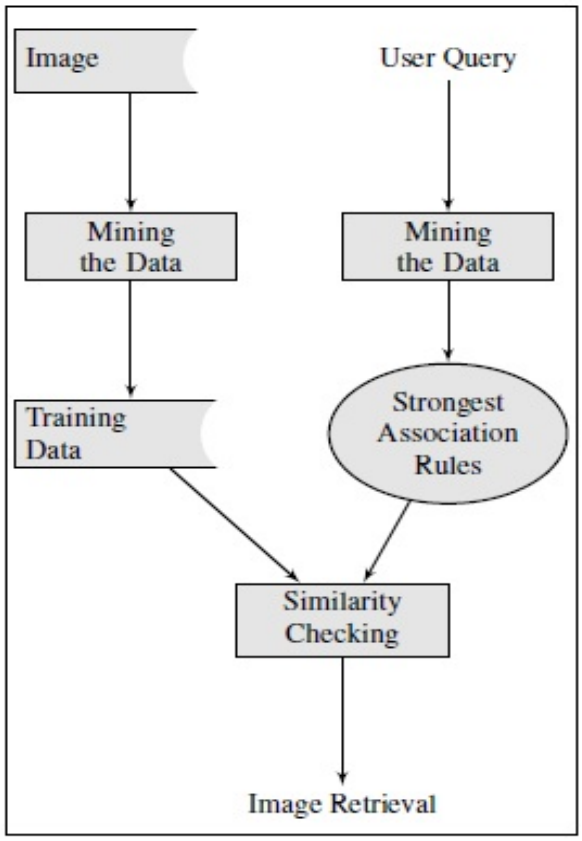

Fig:2 Architecture of image retrieval system

2. Fusion of multimodal features

Image search engines use text-based methods for image retrieval. Most of the existing content- based image retrieval use the visual features for retrieving purposes. The fusion of the multimodal features has a potential to improve the retrieval performance[5]. The fusion of textual features and visual features can be done in different levels like earley fusion, late fusion, and trans media fusion.

3. Clustering

Clustering is a data mining task to group similar data together into clusters. For example identify customers with similar buying habits. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. There is several algorithms for clustering.

In this system K-means clustering algorithm is performed [9]. It's a type of centroid-base clustering. k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This approach minimizes the overall within-cluster dispersion by iterative reallocation of cluster members.

K clusters are formed by assigning each data point to its closest cluster mean. The algorithm uses the Eucledian distance. Virtual means for each cluster are calculated by using all data points contained in a cluster. The second and third step is iterated until a predefined number of iteration is reached or the consistence of the clusters does not change anymore. The runtime for the algorithm is O(n). K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. Clustering will be more advantage for reducing the searching time of images in the database[2].

4. Generating association rules

Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository.

For simple example find all items which are frequently purchased with milk. Using association rules mining algorithm generate association rules from the textual clusters and visual clusters.

a. Constructing transaction database

We are interested to find out the association between text feature clusters and visual feature clusters. The transaction database contain large number of transactions. Each transaction contains a text cluster and visual cluster. If the cardinality of the common images set was not zero, the clusters combined at the same transaction.

b. Calculate support and confidence

The support and confidence are two interestingness measures in data mining. The rule $X \rightarrow Y$ holds with support s if s% of transactions in domain D contain X U Y. Rules that have a s greater than a user- specified support is said to have minimum support.

The rule $X \rightarrow Y$ holds with confidence c if c% of the transactions in D contain X also contain Y. Rules that have a c greater than a user-specified confidence is said to have minimum confidence

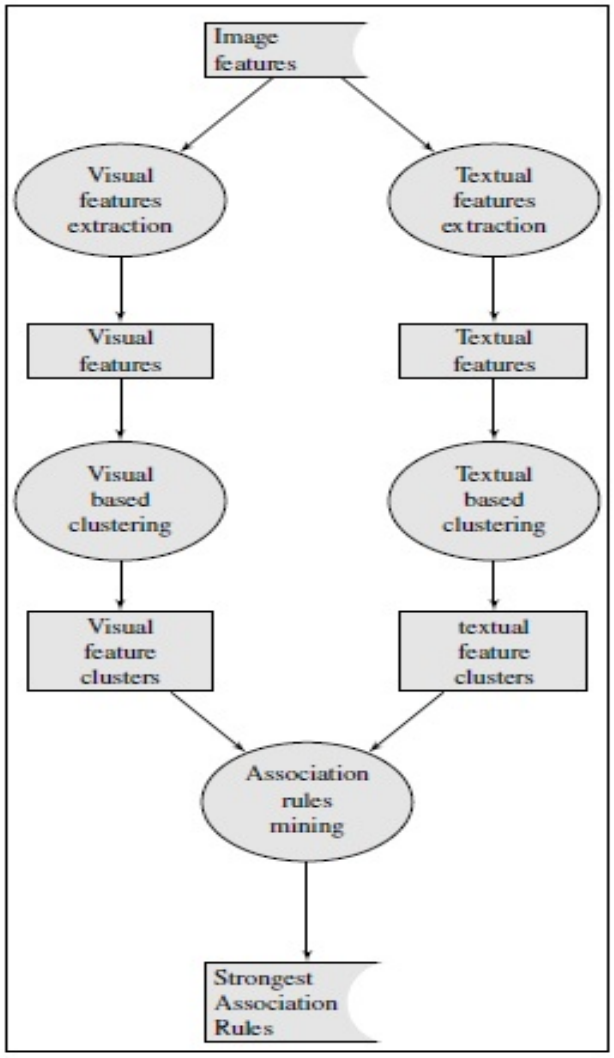

Fig:3 Mining the data

.

c. Mining frequent item set

We need to identify the frequent item sets from the transactions. The support and confidence is used to discover all frequent patterns of the association between textual feature clusters and visual feature clusters. The frequent item sets are used later to generate strong ARs.

5. Generate strong association rules

Our association rules containing visual clusters at the left hand side and textual clusters at the right hand side. The association rules which is equal to or greater than minimum confidence and support consider as strong association rules.

6. Similarity checking

To retrieve images from the data set, we give an image as an input query. Then we search for the images in the training data set that have the similar strong association rules. That means search and retrieve the images which have similar strong association rules of the input image. Then retrieve most relevant images.

IV EXPERIMENTS

A. Dataset And Tools

The experiment is done at the test bed of ImageCLEF 2011 Wikipedia collection. It consists of 50 topics and 54,534 Wikipedia images along with their user-provided annotations in three different languages . The data set contain two files. One is a .xml file which contain image ID, name and description. Other file includes the feature descriptor values [10].This dataset was our choice because it is a typical example for Web images and it is available for public use with its ground truth.

The tool used to perform this experiment is Matlab. Matlab is basically a high level language. It has many specialized toolboxes for making things easier for us. A very large database of built-in-algorithms for image processing and computer vision applications are encoded with it. More over it support data mining techniques like clustering, association rules etc.

B. Experimental Setup

The experiment is repeated for many times. After analyzing the results set minimum support and minimum confidence as 2% and 70% respectively. The input query is given as images.

C. Experimental Results

The precision and recall are the two measures used to evaluate the performance. Precision is the fraction of number of relevant images retrieved to the total number of images retrieved. Recall is the fraction of number of relevant images retrieved to the total number of relevant images in the database.

Comparing to other image retrieval methods, this system produces a good result. The performance measures are higher than the existing methods. The Fig: 4 illustrates the precision - recall curve of the image retrieval system.

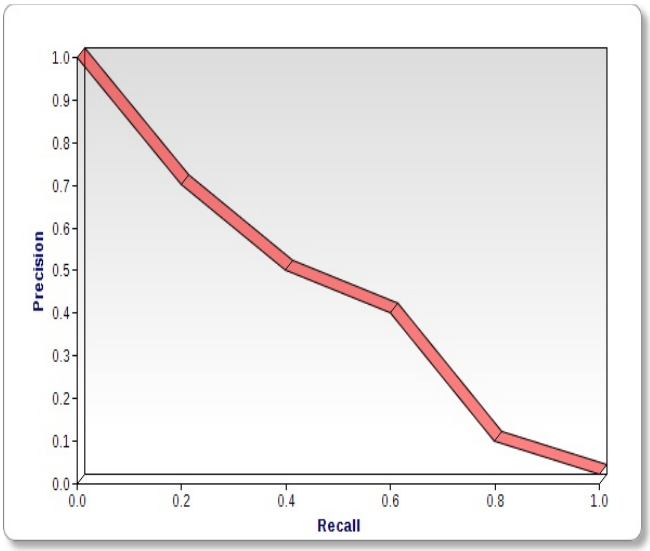

Fig:4 Precision - Recall graph

V. CONCLUSION

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. The proposed image retrieval system functioning effectively.

ACKNOWLEDGMENT

The authors would like to gratefully and sincerely thank the anonymous reviewers and advisors for their constructive comments.

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

[3] Carlos Ordonez, Edward Omiecinski,"Image Mining:A new approach for Data Mining."

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

[5] 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 2014

[6] Pradeep K. Atrey • M. Anwar Hossain, " Multimodal fusion for multimedia analysis: a survey", Springer-Verlag 2010

[7] Xin Zhou, Adrien Depeursinge, " Information Fusion for Combining Visual and Textual Image Retrieval", 2010 International Conference on Pattern Recognition.

[8] Herbert Bay ,Tinne Tuytelaars, and Luc Van Gool,(2008) , "SURF: Speeded Up Robust Features", ECCV 2006 conference in Graz.

[9] Parul M.Jain, Dr. A. D. Gawande, Prof. L. K. Gautam (2013)," Image Mining for Image Retrieval Using Hierarchical K-Means Algorithm", International Journal of Research in Computer Engineering and Electronics.

[10] T. Tsikrika, A. Popescu, J. Kludas, (2011), "Overview of the Wikipedia Image Retrieval Task at ImageCLEF 2011", In: Working Notes of CLEF 2011, Amsterdam, The Netherlands.

[11] Theo Gevers, Joost Van De, Harro Stokman, "Color Image Processing: Emerging Applications"

[12] Ruhan He , Naixue Xiong , Laurence T. Yang , Jong Hyuk Park, " Using Multi-Modal Semantic Association Rules to fuse keywords and visual features automatically for Web image retrieval", www.elsevier.com/locate/inffus