Efficient Technique for Web Image Mining

Praveen Kumar
Department of CSE
National Institute of Technology Patna
Patna, India

Md. T. U. Haider
Department of CSE
National Institute of Technology Patna
Patna, India

ABSTRACT

Web image mining is a growing area in present environment. It defines for the use of web image mining techniques on web to find the hidden information which is present in the image as text. In this paper a literature survey has been proposed on web image mining. Web image mining is a technique of searching ,retrieving and accessing the data from an image, There are two type of web image mining techniques i.e. Text based web image mining and image based web image mining. The objective of this paper is to present tools and technique which are used in past and current evaluation. We will show a summarize report for overall development in web image mining.

Keywords

Web image mining, Accountability, image retrieval, data mining.

1. INTRODUCTION

World Wide Web is allowed user to access image. There are a millions of images on web. Accessing this image on the basis of its internal properties, require some efficient technique. Web image mining handles with the extraction of inbuilt knowledge, relationship between image data or other properties which not stored with the images. The features of an image may be colour, Texture, edge and shape etc. The mining data may vary from structured to unstructured [1].

Searching knowledge from image database is the main objective of image mining. Since today digital images has popular especially on the www and many other applications. Users in many fields like medical image are using the chances got by the efficient to use and manipulate web stored images. The difficulties of web image mining are become wide. Web images are ease environment for storing and explaining physical, temporal and spatial part of important information contained in different domain.

Text based image mining is the type of computer vision to the web image mining problem of finding for digital images in a large collection of database. Previous techniques for image mining have complex and time consuming. Traditional methods of image indexing require associating it with a number or with a keyword or any description which make complex. Image web mining is the technique of determining and searching knowledge or important information in images which is located on the web. For image mining the concerned areas are database systems, machine learning, soft computing, Artificial Intelligence and Pattern matching or recognition [10]. A very high resolution image contains many features and these features can be used to classify the image. Segmentation technique is used for dividing an image into multiple parts. Main techniques use in segmentation are Thresholding, colour based segmentation such as K mean clustering, watershed segmentation or texture filter etc. [2]. These approach works on the basis of colour, grey value or texture property. In first approach, we divide an image on the basis of abrupt changes in intensity. Edge is a good example of it. In second approach, the image is divided on the basis of regions. Thresholding and histogram equalization is a good example of regions based segmentation. After segmentation classification operation is performed. Both unsupervised and supervised technique can be used for knowledge searching.

2. FEATURE EXTRACTION OF IMAGE

In image web mining, feature of image is used for finding a piece of information. Text in image is an important feature of all innate property that shows visual property and each property have a feature of homogeneity. It has vital information about the image pixels. It also shows the importance of the image and its surrounding environment. In other words, it has a feature that shows contrast, directionality, roughness, energy and so on. The main steps for extracting features are:

(a) Calculating local image feature (Finding local image property): to calculate this, the features of neighbourhoods within an image which is called block are extracted and then classified.

(b) Finding image feature using image grid: The size of the neighbourhood within an image is calculated. This measure is called block. Now, find the space between two neighbourhoods. [3]. An image mining or extraction system is a software system for Searching, browsing and retrieving images from large collection databases. The feature factor of images can be calculated within a large image collection based on the basis of colour property using mathematical approaches. These approaches are applied and introduced for web mining of images. Images can be categorize using threshold values which is red , green and blue colour combinations for web mining of images, which are experimented and results are verified. This method provides the best solution in large image set compared with total of 8000 images with different resolution [4]. All proposed methods are helpful to perform the better output and based on query required images extract from the data ware housing. So we select the combinations of colours which are standard deviation and mean with median, expecting best result and good performance. Web images are stored in data ware housing for fast performance and Content based image extraction is used to extract require image. Haar wavelet is used for image compression without data losses. [5] Edge and texture features are extracted from the compressed web images. Gabor transforms and Sobel edge detector are used for extracted compressed image. These can also be applied for Image mining on robots. It is a knowledge driven process. Knowledge can be achieved from two techniques one is from data and other is from expert. Web image mining is a technique to process internal image property or knowledge from web image.

Association rule mining is also useful for web image mining. Basically it is a process to find correlation and association among large images. Here confidence and support are considered for rule interestingness. Association rule mining technique is used for image which belongs to this category and its steps are shown in figure1:

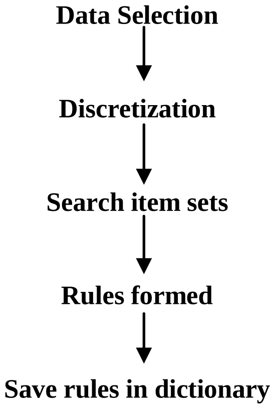


Figure1. Steps for association rule mining

Step1: Data is selected from data base.

Step2: Discretization is applied on numerical attribute of selected image. It divides every features of attribute into fixed interval and converts it into a Boolean value.

Steps3: Search the item sets that is frequently used.

Step4: Finally generate rule for item sets that is frequently used.

Step5: Store these rules into knowledge dictionary.

3. APPROACHES FOR WEB IMAGE MINING

In web image mining, every image that kept in a backend contains its properties which is save by extracted its feature and these features is further compared with the users requirement. Hence, this is a combination of many fields, such as machine learning, pattern matching, filtering and so on. There are four steps used for web image mining which is shown in figure2.

3.1 Collecting image on the basis of pattern matching*:*

A brute force technique can be used for template matching. If an image contains p×q pixels then we find the fixed value of entire image then we compare the entire image and find the one that matches the current template using table look –up or dictionary lookup technique.

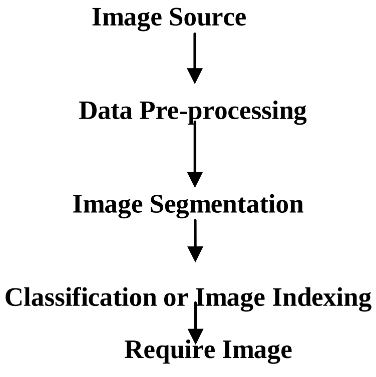


Figure2. Shows a basic steps of web image mining

3.2 Image pre-processing

Since lot of dirty and noisy data occur in large images like image may be unclear. For this, there are many techniques like image thresholding, border tracing and wavelet based segmentation, and etc. is used for pre-processing of image[9]. The feature vector of image is calculated. Source image are generally not changed. So first a virtual database for storing these images is created.

3.3 Compute feature vector

Color and edge is important feature because different color and different edge represent different images. We compute histogram for color feature by equalizing the colors within an image and calculating the total number of pixels for every image. This is not local feature of image. We can neglect the features of image translation, image rotation and scaling [6].

3.4 Extracting information from feature vector

Classification technique is used for extracting information from feature vector. Classification may be supervised or may be unsupervised which divide the image into several parts. The objective of image classification during web image mining is to collect a lot of information in which user have interest. Experimentally, supervised classification is more suitable for web image mining than unsupervised classification. Generally clustering of image is performed in starting of the mining processing. When image is being clustered then expert system is required to evaluate the image of every cluster [7]. The classification techniques called Decision classifier tree can be used to classification purpose. In this method, testing and training is performed on image set. In training steps, image feature is extracted and the image is classified on the basis of its features.

In Supervised method, decision tree is used for extracting low level features from images. When selecting the property, we are considered the minor difference between image pixels and text pixels. Text based areas in images have either simple straight lines or sharp edges or geometric shapes. Fixed size and type font is used. Hence the pixels is distributed in image is homogenous where there is a text. However, text area in images has fixed boundaries and edges are distributed randomly. The text areas in these types of images have more contrast with their backend. Lastly, a decision tree is used for breaking complex decision making steps which became easy to compute and fast feature extraction [8].

The process followed in the classification are shown in figure3.

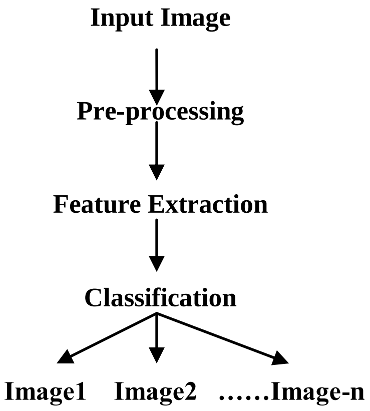
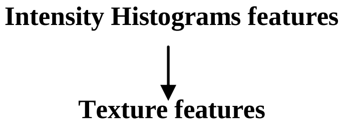


Fig.:3 Steps of image classification

For training set, images which are in JPEG format are chosen because JPEG can be resized for low resolution. For extracting the text feature, we transform the image pixels into binary value which shows better the contrast value. Medical images have less contrast compared to other images because text is organized in different background. Figure4 shows the steps used for extracting feature from images:



Intensity Histograms features

Texture features

Fig.4: Steps for feature extraction

Digital images contain two important features that differentiate the different images. The text colour and its background are uniform with respect to the other images. Also there is more text alphabets present in an image with respect to the background pixels. We can also differentiate with the help of intensity of the pixels. Hence histogram features are calculated for classification purpose. Intensity histogram quantizes the gray scale values into the range 0-255 for using 256 bit histograms and then we normalized the histograms by dividing the value by total sum of the histogram. We obtained confusion matrix for the classification using training data set.

We use fuzzy set for labelling the steps for web image. The main work of fuzzy set is to calculate membership value of image which is requested by user. For labelling an image, we first collect image in a form of set and apply fuzzy rule. The selected image set has value 1 where rest image have value 0[10].

Support Vector machines are used in web image mining for pattern matching. We use support vector machine for classification because SVM increase the size of classes and search the best and relevant hyper-plane. It divides the web image into two classes and then we use fully rule for labelling an image.

4. CONCLUSION

Web image mining needs collaboration among different domains such as computer vision, image processing, image retrieval, data mining, machine learning, database and artificial intelligence. In this paper the area of image mining has been reviewed by proving the overall steps involve in web image mining. There are several methods for segmentation and classification which can use to improve or enhance the web image mining. Low level features such as energy, contrast, mean and skewness can also be used for extracting image. Those images can be considered where there is scope to combine caption text and a scene within an image. Web image mining is useful for extracting information from web images. The available search engines usually give a large number of pages with useless images in response of user requirement whereas user always wants best useful image. The weighed page rank and page Rank technique focus on the importance of link with respect to the text of the image.

5.REFERENCES

[1] Deepak Kolippakkam, Huan Liu, et al. “Feature Extraction for Image Mining”.

[2] Goldman, S.,Zhou, Y. “Enhancing supervised learning with unlabelled data.” , In: procceding OF the 17th international conference.

[3] Asanobu K . Data Mining for Typhoon Image Proc. on MDM. KDD2001

[4] Ramadass Sudhir,” A survey on image mining techniques: theory and applications”, Computer Engineering and Intelligent Systems ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol 2, No.6, 2011

[5] Agrawal, R., Imielinski, T., Swami, A. N.: “Mining association rules between sets of items in large datasets”, In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 207-216 (1993).

[6] Agrawal, R., Srikant, R.: “Fast algorithms for mining association rules”, In: Proceedings of the 20th international conference. Very Large Data Bases, pp. 487-499 (1994).

[7] R. F. Cromp and W. J. Campbell.,”Data mining of multidimensional remotely sensed images”, International Conference on Information and Knowledge Management (CIKM), 1993.

[8] S. Chitrakala , P.Shamini and d.manjula ,”Multi class enhanced image mining of heterogeneous texture image using muliple image features”, International Conference Patiala 2009.

[9] Russ, John C. (1998), *The Image Processing handbook.* Boca Raton, FL: CRC Press, 3rd Edition.

[10] Monika Sahu, Madhup Shrivastava, M. A. Rizvi, “Image mining: a new approach for data mining based on texture”, 3rd International Conference on Computer and Communication Technology.

[11] J. Li, W. Wu, T. Wang and Y. Zhang, “One step beyond histograms: Image representation using Markov stationaryfeatures,” Computer Vision and Pattern Recognition, June 2008.

[12] L. Page et al. “The PageRank citation ranking: bring order to the web,” technical report, Stanford Digital Library Technologies, 1999-0120, Jan. 1998.

[13] J. Liu, Q. Liu, J. Wang, H. Lu, S. Ma, “Web image mining based on modeling concept-sensitive salient regions,” Int’l Conf. on Multimedia .