RETRIEVAL OF TEXTUAL AND NON-TEXTUAL INFORMATION IN CLOUD

Anbarasi M S[1], Divya R[2], Buvaneswari P[3], Illakkiya M[4]

[1]*Assistant Professor, Department of Information Technology, Pondicherry Engineering College, Puducherry, India*
[2]*Student, Department of Information Technology, Pondicherry Engineering College, Puducherry, India*
[3]*Student, Department of Information Technology, Pondicherry Engineering College, Puducherry, India*
[4]*Student, Department of Information Technology, Pondicherry Engineering College, Puducherry, India*

Abstract

With the advent of the Internet there is an exponential growth in multimedia content in various databases, which has a major issue in effective access and retrieval of both textual and non-textual resources. To resolve this problem information is stored in the cloud environment. From Internet the large and complex data cannot be stored and processed using traditional data processing applications. The idea in this proposed method involves parsing of the web page for the extraction of textual data and images. Textual retrieval is done through keyword extraction whereas feature extraction technique is done for the image retrieval. K-means algorithm is used to perform clustering. Based on ranking both the textual data and non-textual data are retrieved together.

Keywords— *Retrieval, Feature Extraction, K- means Clustering, Cloud*

----------***----------

1. INTRODUCTION

The World Wide Web is a very large distributed digital information space. Images are a major source of content on the Internet. The development of technology such as digital cameras and mobile telephones equipped with such devices generates huge amounts of non-textual information, such as images. The ability to search and retrieve information from the Web efficiently and effectively is an emerging technology in information retrieval. The existing system retrieves too many documents, of which only a small chunk are relevant to the user query.

The idea of the proposed work is to retrieve relevant textual and non textual information together for the user query in cloud, so it is entitled as Retrieval of Textual and Non-Textual Information in Cloud (RTNIC). For example, if the user needs to retrieve information for the query animal then this retrieval process provides the relevant non textual information (i.e. images) of animals along with its textual description about the image. The main aim is to obtain the correspondences between the image and its associated text for the easy understanding. Cloud [3] is used to enhance data management and storage of huge amount of information since the growth of information to be stored increases day by day.

Cloud is used since the information being accessed from a centralized storage [3], and does not need any user to be in a specific place to access it. By this way information is delivered and resources are retrieved by web-based system, rather than a direct connection to a server. The main advantages of using cloud are:

- It is highly elastic.
- Everything is provided as service.
- Less power consumed on hardware and software.
- High availability and scalability.
- No data loss.

Thus the information retrieval in the cloud is a popular research area. It is shown in figure 1.In recent years the web document analysis has been done to effective filtering of useful information from them. The multimedia documents consist of different components (texts, images, sounds, videos).But we concentrate on image and text. The main aim is to obtain the correspondences between the image and its associated text for the retrieval accuracy. The research in the web page processing is focuses only on the textual analysis of the segments around the non-textual information.

2. RELATED WORK

Current research in image retrieval uses both the textual and visual features for the retrieval of the image. Ontology based retrieval of image is used to reduce the semantic gap. It focuses on the semantic content which relates to the user's intent. This paper also concentrates on the tag refinement. The visual words which are very sparse to match is overcome. The textual cluster and the visual cluster are mapped to retrieve the image. [10]

The second approach extracts the text and images from the webpage and stores them in the different databases. The image analysis [8] which includes segmentation followed by Scale Invariant feature Transform (SIFT) feature extraction. The segmented images reduces the number of SIFT points which increases the matching of images from the database and the query. The Support Vector Machine (SVM) classifier is used to classify the images to various classes. The textual and the images are subjected to the semantic inclusion.

3. ARCHITECTURE OF THE PROPOSED SYSTEM

The large and complex dataset of texts and images cannot be processed using the traditional database applications. The proposed work is the retrieval of the relevant textual and non-textual information in the cloud.

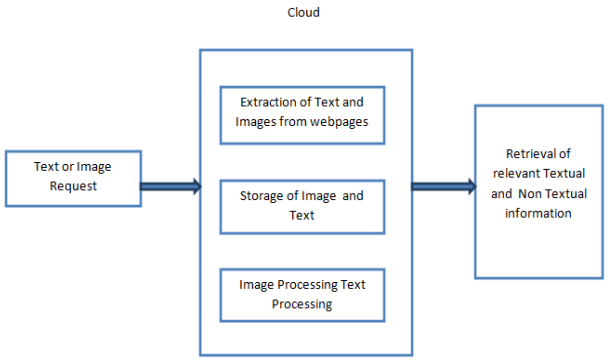

Fig1: RTNIC High Level System Architecture

The proposed approach not only concentrates on the image retrieval alone but also retrieves the relevant textual information. The existing retrieval processes are either retrieving image alone or text alone. The content based retrieval methods are used in the existing processes where it is difficult to match the image contents with the existing images in the database [5]. The system which is retrieval from the cloud concentrates on the retrieval methods that are specific to the cloud environment.

3.1. Preprocessing Stage

The proposed system consists of two phases. The first phase is the preprocessing stage. The second phase is the Common Retrieval phase. The preprocessing stage involves the collection of text and image from the webpage and the storage in the database. The second phase involves the ranking algorithm for the relevant retrieval from the database. The preprocessing phase includes three modules. They are Parser module, Image processing module and the Text processing module. The following section contains the descriptions of the modules and its functions.

3.1.1. Html Parsing Module

The HTML parsing module includes the conversion of the HTML page into a DOM tree [1,6]. The DOM tree based web page segmentation algorithm is used to segment web pages into sections. Each section containing the text and the images are extracted. The tags such as <TD>, <TR>, <TABLE>, <HR> are used to separate the different content passages [9]

3.1.2. Image Processing module

This module explains the image feature extraction process.

3.1.2.1 Feature Extraction Process

The Color histogram is the feature to be extracted from the images. They are trivial and popular to compute. The color histogram method extracts three histograms of the RGB colors. It computes the occurrences of each color. When computing is completed it is normalized because they are collected from different sites.

3.1.2.2 Image Clustering

The image clustering is done using the k-means algorithm. This is an unsupervised clustering process. The k clusters are produced using this k-means clustering.

K-means is a partitioning algorithm which provides k clusters where k is fixed as a priori. K-means algorithm treats each observation in data as a object in the space. The objects within each cluster are closer to each other. The centroid is chosen from the first k points. Then every point in the dataset is assigned to the nearest centroid. Many iteration are done in which each dataset is assigned to the nearest centroid.

3.1.3 Text Processing

Structured Text is extracted from the WebPages where it is subjected to the stop words removal, stemming and finally the keyword extraction.

3.1.3.1 Stop Words Removal

There is a need for the removal of non-informative words in the sentences. There is a pre-defined frequency list where the commonly occurring non-informative words are stored. It is used to eliminate the non informative words.

3.1.3.2 Stemming

Stemming is changing the words to its basic form. For example the walking, walks can be converted to walk.

3.1.3.3 Keywords Extraction

A set of meaningful keywords is extracted to be tagged. The extracted sentence and its associated keywords are stored in the database.

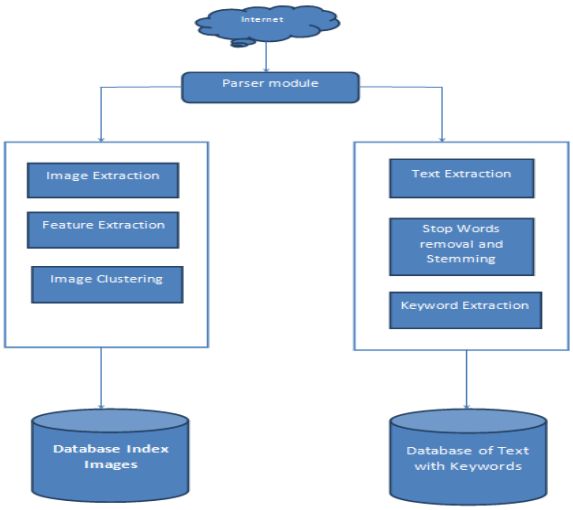

Fig2: Preprocessing of Data.

3.2 Ranking Phase

A(r) is any non-increasing function and r(z(k)) [5] is the rank of image z(k). We care only about the ranks of the top K images; we can define A(r) as:

$$A(r) = \max (K + 1 - r; 0)$$

Thus the lower (top) ranked images are assigned higher weights and since A(r) = 0 for r > K, only the top K images of the ranking are considered. Now the top ranked images are mapped with the associated text that is stored in the text database by using the same tag that is used to store the image.

3.3 Retrieval Phase

In the preprocessing stage, the text and the image are stored and indexed. In this retrieval phase the images and texts are retrieved based on the user's query.

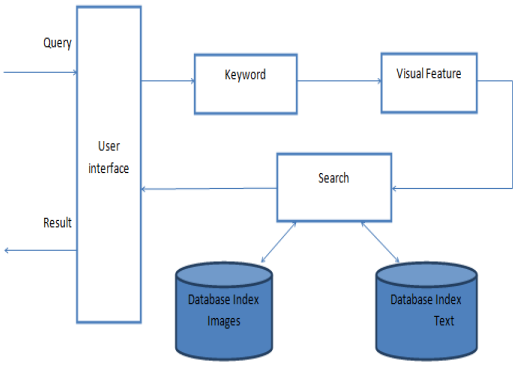

Fig 3: The Retrieval phase

The visual features are computed for the images in the database in the preprocessing stage and the mean of the visual features is matched to the textual term.

The visual feature used here is color histogram of the red, blue and green values. Thus for a single image there are many feature values computed.When the query is given, the keyword is obtained and it is matched with the visual feature of the image in the database. If the values are matched with the database, the images and the corresponding text are retrieved [7].

3.3.1 Text and Image Query

The result will be based on the aggregation of the scores of the image and text retrieval [2]. The aggregation operator can be of different values. The different aggregation operator used can be of different behavior [4]. If it is maximum, we get the highly relevant text and image. If it is minimum, then it can be either the best image or the best text and not both. Another very common approach is the aggregate the results using a mean.

3.3.1.1 The Output of the RTNIC System for Combined Text and Image Retrieval

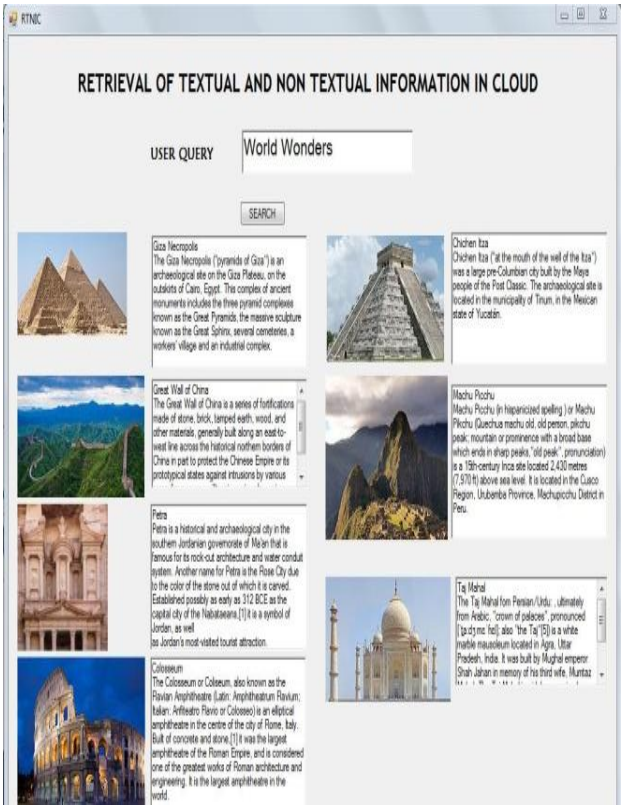

Fig 4: The Output of RTNIC combined image and text retrieval

The User Interface provides the option for the user to the give the textual query [7]. When the query is given the related textual and non-textual data is retrieved. The top k ranked pairs of text and image are retrieved. When the query World Wonders is given the related text and images are retrieved.

It is the case when we have both the text content and image content in the database.

3.1.3.2 The Output of the RTNIC System for the Image Retrieval

There might be some cases where there may be images but not the related text associated to it. In that case the images alone are retrieved. For example the query "Book" is given as the input. Since the images only are available for this query, they are displayed according to the rank.

Fig 5: The Output of the RTNIC image retrieval

3.1.3.3 The Output of the RTNIC for Text Retrieval

There are exceptions when we have the text data available for the query but no images relating to the query. In that case the RTNIC displays the text associated with the query. For example when the query "Heaven" is given, there is no images associated with it. Hence the textual information alone is retrieved based on the ranking.

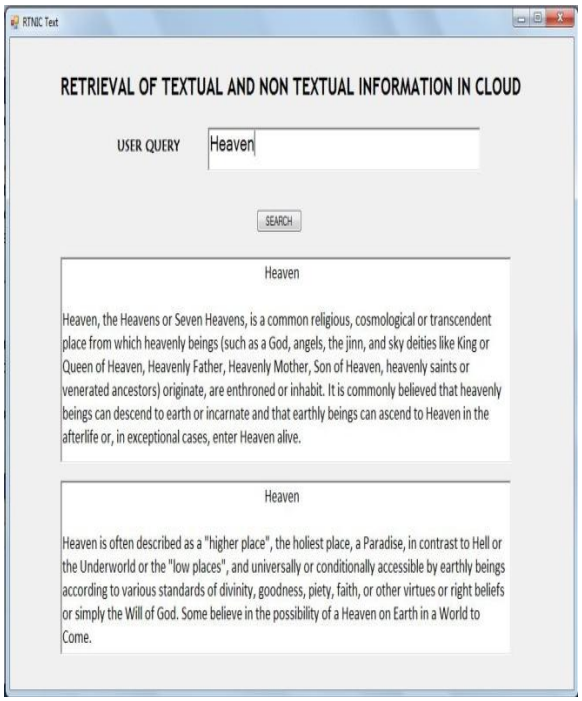

Fig 6: The Output of the RTNIC for the text retrieval

4. EXPERIMENTAL RESULT OF RTNIC

4.1. Performance Evaluation

Then the following two measurements quantify the quality of the search:

Recall = R / M = Number of retrieved images and text that are also relevant / Total number of relevant images and text.

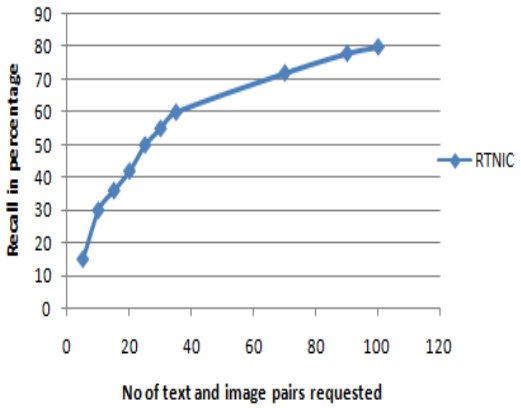

Fig 7: Recall

Precision = R / N = Number of retrieved images and text that are also relevant / Total number of retrieved images. The recall is the answer to the question: How close am I to getting

all good matches? The precision is the answer to the question: How close am I to getting *only* good matches?

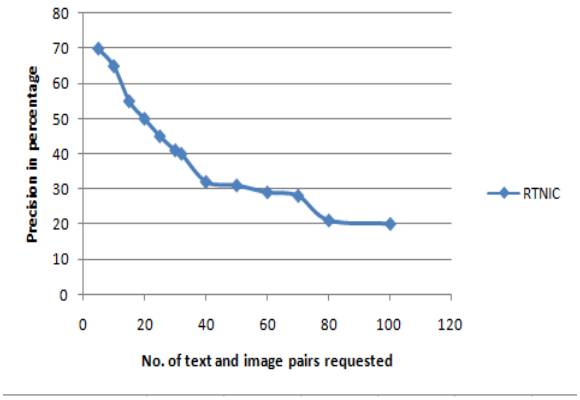

Fig 8: Precision

5. CONCLUSIONS AND FUTURE WORK

The information in the Internet must be archived, maintained and effectively managed for the retrieval. The proposed work concentrates in the retrieval of the textual and non-textual information which are relevant. Thus the proposed work is aimed at complementing text retrieval and image retrieval each other. The cloud environment enhances the data security and storage issues in large dataset. The future work will be the video and text retrieval Effective retrieval of video from the databases and efficient indexing of videos.

REFERENCES

[1]. Alaa Riad, Hamdy Elminir and Sameh Abd-Elghany, "Web Image Retrieval Search Engine based on Semantically Shared Annotation", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 3, March 2012

[2]. A. BalaSubramanium, "Information Retrieval Techniques for non textual media".

[3]. Cong Wang, **"**Toward Secure and Dependable Storage Services in Cloud Computing", IEEE Transactions on Services computing, 2012.

[4]. L. P. Florence, "Image and Text Mining Based on Contextual Exploration from Multiple Points of View," Twenty-Fourth International FLAIRS Conference, 2011, Palm Beach, Florida, 18-20 May.

[5]. N. Haque. "Image Ranking for Multimedia Retrieval*"*. Ph.D. thesis, School of Computer Science
and Information Technology, Royal Melbourne Institute of Technology, 2003.

[6]. Martina Zachariasova, Robert Hudec, Miroslav Benco, and Patrik Kamencay, "Automatic Extraction of Non-Textual Information in Web Document and Their Classification",IEEE 2012

[7]. Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle,Gert R.G. Lanckriet1, Roger Levy, NunoVasconcelos , "A New Approach to Cross-Modal Multimedia Retrieval".MM'10.

[8]. G.Tryfou and N. Tsapatsoulis," Web Image context extraction based on Semantic representation of web page visual segments", 7th International workshop on Semantic and Social media adaptation and Personalization, 2012.

[9]. M.J. Parag, I. Sam, "Web document text and images extraction using DOM analysis and natural language processing, " In Proceedings of the 9th ACM symposium on Document engineering Doc Eng 09

[10]. Yin-Hsi Kuo, Wen-Huang Cheng, *Member, IEEE*, Hsuan-Tien Lin, *Member, IEEE*, and Winston H.Hsu,"Unsupervised Semantic Feature Discovery for Image Object Retrieval and Tag Refinement", IEEE Transactions on Multimedia, Vol. 14, No. 4, August 2012.

BIOGRAPHIES

Dr. M. S. Anbarasi has completed B.E (Comp Sc & Tech), M.E.(SE) & Ph.d in Data Mining from Anna University CEG Campus, Chennai 25. She has 15 years of teaching experience and 7 years of research experience in the areas Data Mining, Software Engineering and Cloud Computing

Divya R is the final year student, Department of Information Technology in Pondicherry Engineering College, Puducherry, India. Her areas of interest are Big Data and Information Retrieval.

ILLAKKIYA M is the final year student of Department of Information Technology in Pondicherry Engineering College, Puducherry, India. Her areas of interest are Data Mining and Warehousing.

Buvaneswari P is the final year student of Department of Information Technology in Pondicherry Engineering College, Puducherry, India. Her areas of interest are Web mining and Image Retrieval.