Volume 5, Issue 11, November 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey on Retrieval of Textual and Non-Textual Information in Cloud

[1]Rashika Dandel, [2]Monika Gutal, [3]Bhagyashree Dudhal, [4]Poonam Jadhav, [5]Sonali Kadam
[1, 2, 3, 4] Student, [5] Asst. Prof.
[1, 2, 3, 4, 5] Dept. of Computer Engineerin, S.B.P.C.O.E. Indapur, Pune, Maharashtra, India

Abstract— Internet is the exponential growth in the multimedia content. Multimedia consists of different components like texts, images, videos and sounds etc. There are many problems to access and retrieval of both textual and non-textual information so we have to need solve this problem by using cloud to store the information in large amount. In this method extraction of textual information and non-textual information by using parsing the web page. Textual information retrieval is done by using keyword extraction whereas image retrieval is done through feature extraction. For performing clustering K-means algorithm is used. By using ranking phase both textual and non-textual information are retrieved together.

Keywords— web images retrieval, data mining, k-means algorithm, cloud, feature extraction.

I. INTRODUCTION

Internet is the very large open source search engine. Number of images, texts and videos are available on internet. So finding and accessing text and image is difficult task so some additional process is needed for collection of text and image. There are various search engine are available in open source which include Google search, Lycos, Altra Vista Photo Finder [1]. Various technologies are developed such as mobile, camera which generates large amount of non-textual information such as images. In this paper the existing system retrieves so many documents when user sends query to web. In the proposed system we can retrieve the information of texts, images and videos together. In this system we have to design a cloud so we have to give the retrieval of textual and non-textual information in cloud (RTNIC) [1]. For example – if the user needs to retrieve the information of the given query like birds then this retrieval process provides textual and non-textual information.

Cloud means" storing and accessing data our internet instead of computer devices ".For internet cloud is just metaphor. Today cloud is very popular research area. In multimedia document there is text, images, sound and videos but we consider only image and text in this paper [3]. The main aim is to obtain related between the images, associated text retrieves.

Advantages of cloud:

- Cloud is more elastic.
- In cloud everything is provided by service.
- In cloud less power consumed by using hardware and software.
- No information is loss.
- More scalability and availability.

In this system DOM (Document Object Model) the tree are used to separation of text and images. Those separation can store database in separate. We investigate the retrieval of various combinations of text and image. This system is done by using keyword extraction. The various techniques are used to image retrieval such as K-means algorithm. This paper are try to reduce semantic gap problem and the retrieval precision such that textual context and visual features for retrieval. In this paper the approach of proposed system is not only pure combination between image and textual feature but also that enhancing the retrieval accuracy.

This paper is organize is as follows, in section 2 there are related word and section 3 present in detail

II. LITERATURE SURVEY

First approach of this paper is for image retrieval the current research uses both the textual and visual features for retrieval of the image. To reduce the semantic gap ontology based images are used. The tag refinement is also considered in this paper. There are various clusters available from this cluster we have selected textual cluster and the visual cluster. These two clusters are mapped to retrieve the images [2].

The second approach is extract the text and images from web and this data is stored in different database. The Scale Invariant Feature Transform (SIFT) is used for the image analysis which includes segmentation. The segmented images reduces the SIFT points. In this system the Support Vector Machine (SVM) is used to classify the images to various classes. In this way the textual and images means non-textual are subjected to the semantic inclusion[4].

The traditional data base application is used to large and complex dataset of text and images cannot process. RTNIC system is work of retrieval the relevant textual and non-textual information in a cloud shown in fig

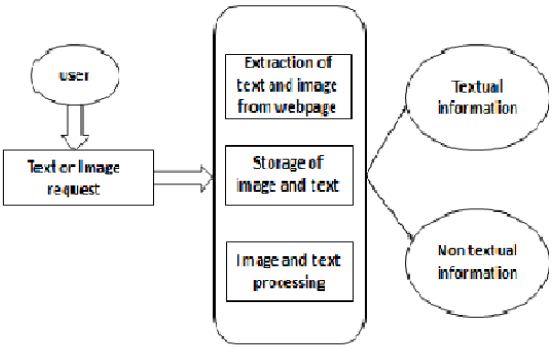


Fig- System Architecture

The proposed approach not only concentrates on the images and the related textual information and videos. The existing retrieval processes are image, text, videos separated. The system which is retrieval information from the cloud [4]. The retrieval information in specific cloud area.

The RTNIC system consists of two main phases:

- Preprocessing stage
- Common retrieval phase

III. PROPOSED SYSTEM

Preprocessing Stage

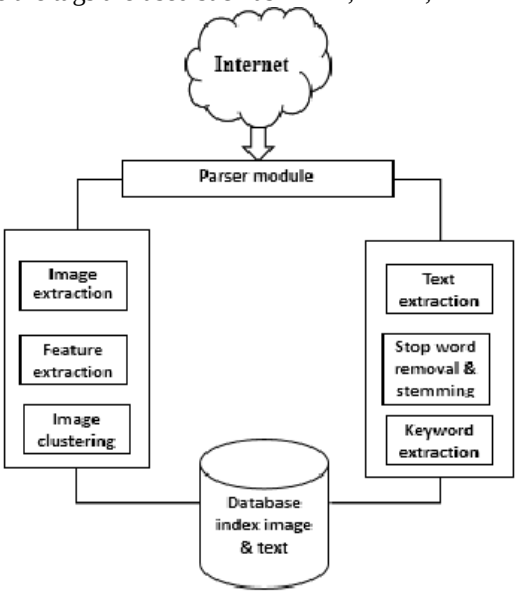
Data preprocessing is an important step in data mining process. In this phase collection of text and image are done from the web page and this data store in the database [5]. In web there are more irrelevant and redundant and noisy data then it is more difficult to extraction of data so we have to use of preprocessing stage it includes feature extraction and selection. Preprocessing phase includes three modules that are:

- Parser module
- Image processing module
- Text processing module

Explanations and its functions of these modules:

2.1.1. Parser module

The parser module is used to conversion of web pages into DOM tree.DOM tree is nothing but segmentation algorithm. This algorithm is used separate the web pages into different section. Each section includes text and the images which are extracted. To separate text and image the tags are used such as <TD>, <TR>, <TABLE>,<HR>



2.1.2. Image Processing Module

Function of this module is related to the image. This module explains the feature extraction process of image and also explains image clustering.

2.1.2.1 Feature Extraction Process

In feature extraction process the color histogram is used because they are popular and it extract feature from the image. In histogram the images is breaking down in various colores components and it extracts three histogram of RGB color such as Red (HR), Green (HG), Blue (HB).

After this process each color is computing to another and finely the histogram of each color is normalized [6].

2.1.2.2 Image Clustering

K-means algorithm is used for performing image clustering and unsupervised clustering process. K-means algorithm provides K-cluster .Where K has fixed as a priori first select K point as the centroid. After choosing centroid the next step is find the nearest ones and consider it.

After that new centroid of the cluster which resulting the previous step are calculated in each the data set is assigned to the nearest new centroid.

Among all partitioning based clustering methods K-means is one of the simplest unsupervised learning methods. There is n number of data objects in K cluster. Where K is the number desired clusters and each cluster. There are centroid the data objects are placed in a cluster which having centroid nearest to that data object to representing a set of n data objects K cluster are found [7].

Method of an algorithm for K-means.

Algorithm

Input:

K-the number of cluster to be partition

n-the number of objects

Output:

A set of k cluster which is based on given similarity function.

Steps:

1) First we can choose the object ‘k is the initial center of cluster
2) Repeat the process

a) Re(assign) the each object to the cluster each cluster

b) For calculate the mean value of the object

3) There are no change until

Highlight:-

It is easy for implementation & undestand to categorical data it is not applicable the total number of cluster in advance there are need to specify k it may be terminate at that local optimum the result and total run time is depends upon the initial partition

2.1.3. Text processing module

Text processing is also called as Natural Language Processing (NLP) module. Large number of structured text is extracted from the [7] webpages by using text processing. The text processing module is subjected to the stop words removal, stemming and the keyword extraction.

2.1.3.1. Stop words removal

In sentences large number non-informative words are present. It causes many problems for searching of query so we have to need removal of these non-informative words. This stop words removal is used to remove these non-informative words. Advantages of this method are for improving accuracy of search results and it is useful for reducing the redundancy of the computation so it is efficient [8].

2.1.3.2. Stemming

Stemming is the important method in text processing module. Stemming is used to changing the word into its basic form. Stemming is also nothing but removing suffix and prefix of given words. For example ‘playing’, ‘plays’ can be converted to ‘play’.

2.1.3.3. Keyword Extraction

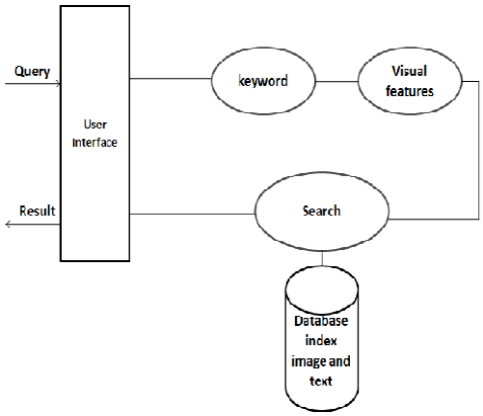
It is the set of meaningful keywords this keywords. This keywords extracted to be tagged. After this the extracted sentence and its associated keywords are store in database.

IV. TEXT AND NON-TEXT RETRIEVAL

There are two phases are present in text and non-text retrieval.

4.1. Retrieval Phase

In this retrieval phase the text and images are retrieved by user’s query. The preprocessing stage indexed is used for storing text and image.



4.1.1. Text and Image Query

The user is going to send the query to showing output. The result is based on the aggregation of the scores of the text and image retrieval. The aggregation is nothing but the operator which having different values. It is uses for different behaviour [7].

4.2. Ranking Phase

We care about the ranks of the top K images. Also we can define A(r) as:

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

Since A(r) =0 for r>K, only the top K images of the ranking are considered.

V. DESIGN OF SYSTEM

1. Design the Basic Architecture of Data Mining
2. User are Send the query request to Database
3 Apply Preprocessing Phase, Ranking Phase on the Data Mining Architecture.
4. Finally, output of RTNIC System for Combined Text and Image Retrieval [8]

5.1 Camparasion

	TEXT	NON-TEXT
SYMBOLIC	Letters, numbers, characters making up the alphabet	Raw data: time samples, pixels, transforms coefficients, etc.
LEXICAL	Words and all theirs variations about rot forms	Threshold events clusters, classes
SYNTACTIC	Grammatical rules, phrases and sentence structure	Probabilistic or kinematical co irrelations, physical constraints over space, time or other relevant dimensions
SEMENTIC	Meaning, perspective, understanding, decisions regarding beliefs or actions	Situational assessment, indications and warnings, predictions, understanding, decisions regarding beliefs or actions

VI. CONCLUSION

The propose this work concentrates in the giving information of the textual and non-textual information which are relevant. Retrieving Text and image from existing system is not sufficient the aim of proposed system is combining text retrieval & image retrieval together. The cloud is useful for security purpose and storage large datasets. The future work will be retrieval of video from database.

REFERENCES

[1] M.L. Kherfi, D. Ziou, A. Bernardi, “Image Retrieval from the World Wide Web: Issues, Techniques, and Systems”, ACM Computing Surveys Vol.36, No. 1, 2004, pp. 35–67.

[2] Y. Alemu, J. Koh, and M. Ikram, “Image Retrieval in Multimedia Databases: A Survey” Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 2009.

[3] A. BalaSubramanium, “Information Retrieval Techniques for non textual media”.

[4] Cong Wang, “Toward Secure and Dependable Storage Services in Cloud Computing”, IEEE Transactions on Services computing, 2012.

[5] 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.

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

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

[8] Z. Gong, Q. Liu, "Improving Keyword Based Web Image Search with Visual Feature Distribution and Term Expansion", Journal Knowledge and Information Systems Vol. 21, No.1, 2009.

[9] H. Wang, S. Liu, L.T. Chia, "Does Ontology Help in Image Retrieval? A Comparison between Keyword, Text Ontology and Multi-Modality Ontology Approaches", Proceedings of the 14th annual ACM International Conference on Multimedia, 2006, pp. 23-27.

[10] Y. Yang, Z. Huang, H. T. Shen, X. Zhou," Mining Multi-Tag Association for Image Tagging", Journal World Wide Web Archive, Vol. 14, No.2, 2011.

