Automatic Web Image Categorization by Image Content:A case study with Web Document Images

Dr. Murugappan. S
Department of CSE
Annamalai University
India

Abirami S
Department of CSE
College Of Engineering Guindy
Chennai, India

Mizpha Poorana Selvi S
Department of CSE
College Of Engineering Guindy
Chennai, India

***Abstract*—Crawling in Web images has become a challenging problem today due to its rising popularity. Among the most valuable Web assets, Categorizing images on the Web is quite difficult. This paper proposes a simple and effective method to separate the Document Images from the available web image sources. This system concentrates on Automatic Image categorization process over web images by employing a filtering technique to discriminate the document images, available in WWW. The strength of this technique lies in capturing the Image information by intensity and frequency histograms for discrimination of web document images.**

Keywords-Document Images, Web Crawler, Image Fetcher, Image Categorizer, Image File Filter.

I. INTRODUCTION

With the fast evolvement of Internet technology, the Internet has played important role in our daily life. Websites offer services and information to their users. With the rapid development of Web, gaining useful information from large amount of information is more and more difficult.

As resources in the internet is very large it brings irrelevant information to users. In this case, the requirement for information is also developing and changing all the time. Therefore, the search engine technology is introduced. Owing to, universal search engines can not meet the need of users to get specific information. .

Due to the huge size of the Web, it has become difficult for the users to find their required information. Searching Engines' efficiency determines that, people can quickly gain the information of their interest from large-scale network. Even though lot of tools are available today, to manipulate textual information from web documents, access to web document images is in its infant stage.

In order to access the information present in web document images a suitable mechanism is required to discriminate the textual images from other images available in the Internet. Therefore, this paper attempts to design an Automatic Web image Categorization System (AIC) which filters out the Document images from Non document images. available in www. keyword and it will fetch the web pages

II. RELATED WORK

The first web crawler, Matthew Gray's web crawler was written in the spring of 1993, for supercomputing Application[1]. In 1994, Pinkerton proposed Breadth First crawling, which is the simplest strategy for crawling. Here crawling has been done in the order in which they are encountered. It does not have any knowledge about the topic. So it provides low performance for keyword based searching[2].

In 1998, Best First crawling technique, was proposed. In that system it simply computes the lexical similarity between the topic's keywords and the source page for the link. Cosine similarity was used by this crawler. In 1998, Brim and Page proposed the Page rank model. The page rank of a page represents the probablity that a random surfer from page to page will be on that page at any given time. Here it maintains data structures in addition to frontier to compute page rank.

In 1999, the new topic called Focused Crawling was introduced. Focused crawlers (also known as topical crawlers) selectively collect Web pages relevant to a certain domain, try to predict whether or not a URL is pointing to a relevant Web page before fetching the page, and carefully visit URLs in an optimal order[3][8].

In 2007, a new technique was introduced called super-peer based P2P system for building an incremental topic-specific web crawler. This develops a topic-based repository of web pages that would later be used in the construction of ontologies. Current crawlers suffer from centralized architecture, having single point of failure and heavy load. Super-peer systems strike a balance between the inherent efficiency of centralized search and the autonomity, load balancing and robustness to attacks, provided by distributed search, with heterogeneity of capabilities across peers.

These Focused crawlers have been used in a variety of applications such as digital libraries search engines and competitive intelligence.

Considering the highly visual and graphical nature of the world wide web, the number of image search engines is very limited. In the last few years Several image search engines like Google Image Search, WebSEEk [6], WebSeer [7],have been developed. While earlier image search engines used text

only, WebSEEK and WebSeer use content based information like color, texture, shape etc. for indexing and query.

However, very few tools are currently available for searching for images and videos. This absence is particularly notable [9, 10]. Visual information is published both as embedded in Web documents and as stand-alone objects. The visual information takes the form of images, graphics, bitmaps, animations and videos[11].

Region Based image search Engine, used for image collection, segmentation into regions and region feature extraction from real Web sites and for test images.It is based on Region Extraction. The segmented regions correspond to semantic objects, allowing efficient indexing and retrieval[5].

But most of the users are more interested in the identification of objects and actions depicted by images than in the color, shape, and other visual properties that most Content Based retrieval systems provide[14]. Because object and action information is more easily obtained from captions, caption-based retrieval appears to be the only hope for broadly useful image retrieval[15].

In 2002 Neil C. Rowe, proposed a intelligent agent Web crawler and caption filter, searches the Web to find image captions and the associated image objects. He mainly searches the clues words from the captions of meta data, and other text clues except captions. It uses a broad set of criteria to yield higher recall than competing systems, which generally focus on high precision[4].

In 2008 An Image Categorizer was introduced that categorizes logo and Trade mark images in the Web Images. It uses the image content features to categorizes the Web images[13].

Even though many crawling techniques were available there is no system that filters the document images from non document images. This paper proposed Automatic Document Image Categorizing System that categorizes the Document images and Non Document images, and our experimental results are promising.

The paper is organized as follows. In the following section depicts the architecture of AIC System. Section III describes the process involved in the proposed Automatic Image Categorization System. Experimental results are presented in Section IV. Performance Evaluation was presented in Section V. Finally, conclusions are drawn in Section VI.

III. SYSTEM ARCHITECTURE

From Google API the seed url set are extracted[12]. After Allowability checking, the allowable seed urls has gone through RUI web Crawling process to identify relavant sub urls. Filter modules filters out the visited and unallowable urls. From the page contents of Relavant sub urls, the location of image file was identified. After this process the Document image categorization was performed

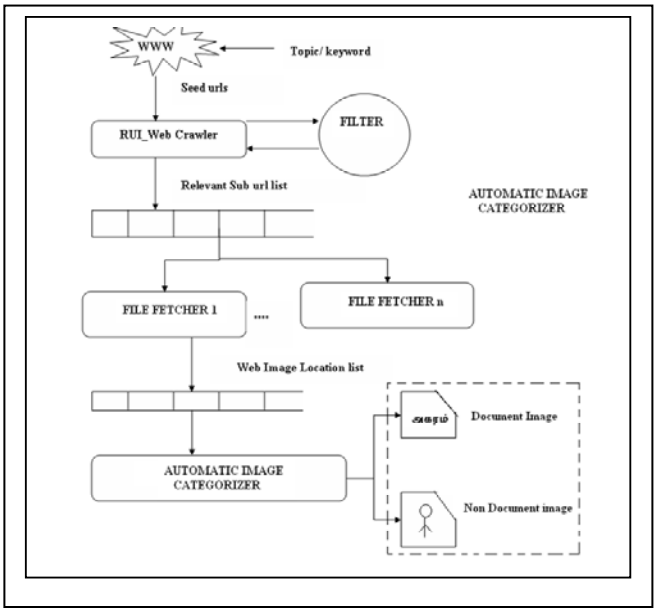

Figure 1. Automatic Web Image Categorization System

IV. AUTOMATIC WEB IMAGE CATEGORIZATION SYSTEN

The basic components of this Automatic Web image categorization systems' are,

1. Seed Extractor.
2. Web Image File Grabber.
3. Automatic Image Categorizer

A. Seed Extractor

Initilay Seed Finder is used which finds web sites that containing topic related documents or most likely lead to those web sites containing such document image files. It is critical to select the seed web page with knowledge acquirement, and this is done by Google API or AltaVista or suggestions from the experts. These seed urls remain to the starting point for the whole crawling process.

B. Web Image File Grabber

In order to grab the web images for the given seed urls we have to identify the relavant sub urls and image locations. This Web Image File Grabber consists two sub parts.

1. RUI Web Crawler.
2. File Fetcher.

1) RUI(Relavant URL Identfication)Web Crawler

Web Crawler identifies Relevant sub url list from the seed url set. Once the seed url is provided, a new URL with its contents has been crawled from the seed url.Once it is exhausted, next url would be obtained from the Queue.. This process will be repeated until no URL exists in Queue. This is summarized as follows.

Input : Seed url, keywords.

Output : Relevant Sub urls.

Process:

1.Begin RUI_WebCrawler

2. Get robots.txt for each seed url.

i. If robots.txt was found then add to unallowable url list.

ii. Else

Add to allowable_url_list.

3. For each url(allowable_url_list)

Enqueue(allowable_url_list,starting_url);

4. Next.

5. While(#url(allowable_url_list>0))

a. url:=dequeue url_with(allowable_url_list)

b. page:=crawled_pagecontent(url);

c. enqueue(crawled_url_list, (url, page));

d. sub_url_list = extract_urls(page);

e. For each page p in crawled_pagecontent

i. If [page p has similiarity with the keyword w in body or in title]

Enqueue(relavant_sub_url_list)

f. Next.

6. End While.

7. End RUI_Webcrawler.

In the above algorithm, operation Enqueue only adds a new URL to the queue if it is not already there. Operation Dequeue return the front element in queue and remove it out of queue. Operation extract_url(page), extracts all link context of url in the web page. Operation crawled_pagecontent(url), extracts the page content of the given url.

2) File Fetcher

File Fetcher uses sub url links discovered by RUI_Web Crawler to identify image files. The image files are identified by parsing through each sub url page content. The file name has been parsed using the "img=" and "src" keywords.

After identifying the image file name the path of the image location has been identified by merging the parent url and image file name. This process also eliminates duplicated files with undesired file names. Later locations of image are stored in Vector list for further image categorization process.

Input : Relavant sub urls.

Output : Image File Locations.

1.Begin Image_File_Fetch()

2. Dequeue(Relevant_sub_url_list, sub_url);

3. For each(#sub_url)

a. page:=crawled_pagecontent(sub_url);

b. Identify file name between the clue words in page p.

c. Merge filename with parent url to find web_image_file_location.

4. Next.

5. End Image_File_Fetch.

C. Automatict Image Categorizer

This module is intended to categorize into document image and non document image. Categorization takes place by calculating the Energy, and Entropy values of the web images. The Black –to-White (BWR) Transition rate also calculated from binarized web document image.

The Automatic Image Categorization (AIC) Algorithm was summarized as follows.

Input : Image File Location.

Output : Categorized Images.

1. Process AIC

2. Dequeue (sub_url_list)

3. For each(web_image_file)

a. Calculate Energy, Entropy for each web_images.

b. If energy> threshold then and entropy < threshold then

Set as document image.

c. Else

Set as Non Document image.

d. Convert Web image into binarized image.

e. Scan the binarized image in horizontal direction.

f. Calculate Average Black-to-White transition (BWR) count for each row of binarized image.

h. If count> threshold then

Set as Document image.

i. Else

Set as Non Document Image.

4. Next.

5. End AIC process.

Whereas entropy and energy has been calculated in Equation (1) and (2).

$$\text{Entropy} := \Sigma i\, h_i \log_2 h_i \quad (1)$$

$$\text{Energy} := \Sigma_i\, hi^2 \quad (2)$$

hi - Value of the ith histogram bin(i Є [0-255] for all histograms)

Entropy value measures the average bits per pixel. Small entropy indicates the presence of homogeneous regions in the image. Therefore, Document images tend to have smaller entropy than other type images.

Energy value is higher for more homogeneous images. Therefore, Document images are generally characterized by large energy values.

V. EXPERIMENTS AND RESULTS

In the Experimental Phase, The seed urls and keyword were given as input for RUI_Web crawling process. Figure 2 shows the result of RUI_Web crawling process, which is the list of relevant sub urls. This process filters out visited suburls and irrelevant sub urls. The relevant sub urls are then stored in a Vector list. The list is saved in a temporary file named crawler.txt for further processing. Figure 2 the list of relevant sub url for the seed url http://www.palani.org.

Figure 2. RUI Web Crawler process

After Image File fetching process the source file (.jpg) has extracted and the full image path has been identified. Figure 3 shows Web Image File Grabbing process. It shows the list of source file names and the absolute path of image locations list for the seed url http://www.palani.org .

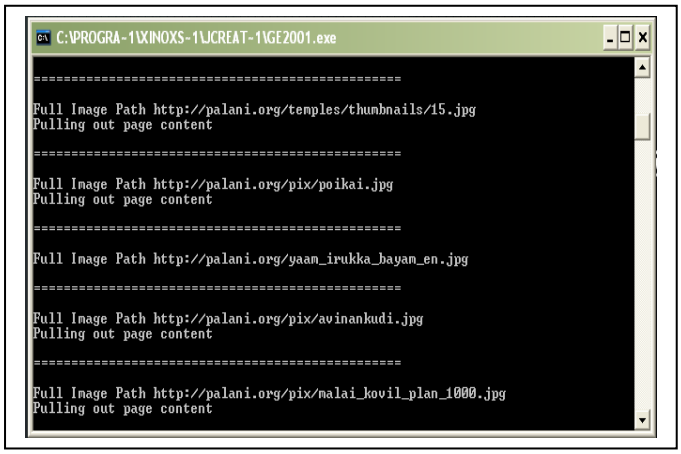

Figure 3. Web Image File Fetching process

After identifying the web image locations the document image categorization process was performed. Figure 4 shows the output of Automatic Image categorization algorithm. Based on the Energy, Entropy, BWR values, the Web images were Categorized.

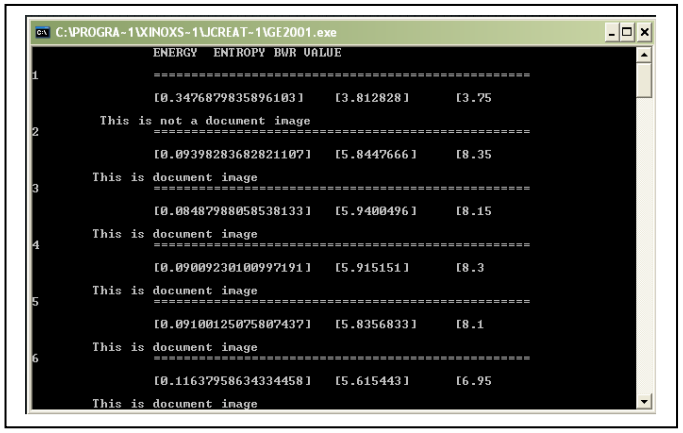

Figure 4. Automatic Image Categorization process

Figure 5 shows the Categorized Document image which has been identified properly. As the entropy value is less than the threshold and energy is greater than threshold and Black to white Transition rate is greater than threshold this image was categorized into Document image.

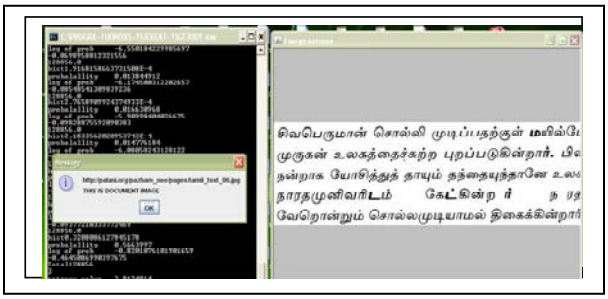

Figure 5. Automatic Image Categorization- Document Image

Figure6 gives the Output of AIC when the image is an non Document image. As the entropy value is greater than the threshold and energy is less than threshold and Black to white Transition rate is less than threshold this image was categorized into Non Document type image.

Also, when the performance of the system has been analysed over a set of seed urls, results are more promising.

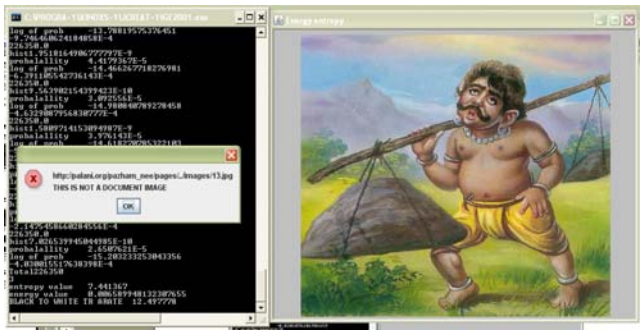

Figure 6. Automatic Image Categorization – Non Document Image

VI. CONCLUSION AND FUTURE WORK

This system provides an architectural framework for the development of Automatic Image Categorization System for Document Images which are available in the Internet. This system mainly performs Automatic image categorization process in web image files. The major benefit of the system is it filters out the Document images from Non Document images from the Internet. As a new attempt this system tries to process within the meta data and also inside the image content for filtering the images.

This system could be extended for Information Retrieval from Web document images.

REFERENCES

[1] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan."Searching the Web",ACM Transactions on Internet Technology, Vol. 1, No. I,August 2002, pp 2-43

[2] M. Najork and I. N. Wiener. "Breadth-first search crawling yields high-quality pages." In Proceedings of the 10th International World Wide Web Conference, 2001.

[3] S. Chakrabarti, M. van den Berg, and B. Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." In Proceedings of the 8th International World Wide Web Conference, 1999

[4] Neil C. Rowe, "Marie-4: A High-Recall, Self-Improving Web Crawler That Finds Images Using Captions ",IEEE Intelligent Systems archive, Volume 17 , Issue 4 (July 2002), pp: 8 – 14.

[5] Kompatsiaris, E. Triantafyllou, M.G. Strintzis, "A World Wide Web Region-Based Image Search Engine," iciap, pp.0392, 11th International Conference on Image Analysis and Processing (ICIAP'01), 2001.

[6] J. R. Smith, S. F. Chang "An Image and Video SearchEngine for the World-Wide Web", IS&T/SPIE Proceedings, Storage & Retrieval for Image and Video Databases V, February 1997.

[7] M. J. Swain, Charles Frankel, Vassilis Athitsos, "WebSeer - An Image Search Engine for the World Wide Web, IEEE Conference on Computer Vision and Pattern Recognition, 1997.

[8] M. Diligenti, F.M. Coetzee, S. Lawrence, C.L. Giles, and M. Gori. "Focused crawling using context graphs." In Proceedings of the 26th International Conference on Very Large Data Bases, 2000.

[9] S. Sclaro, "World wide web image search engines," in NSF Workshop on Visual Information Management, Cambridge, MA, June 1995.

[10] J. R. Smith and S.-E Chang, "Visually searching the Web for Content," IEEE Multimedia Magazine, vol. 4, no. 3, pp.12-20, Summer 1997.

[11] G. S. Jung and V. N. Gudivada, "Autonomous tools for information discovery in the world-wide web," School of Electrical Engineering and Computer Science, 1995.

[12] Shen Jin-Xing, An ontology-based adaptive topical crawling algorithm, IEEE, 2008.

[13] Baratis.E., "Automaticwebsite Summarization By Image Content : A case study with logo and trademark images", IEEE Transactions on knowledge and Data engineering, Vol 20. N0.9. Sep 2008.

[14] C. Jorgensen, "Attributes of Images in Describing Tasks," Information Processing and Management, vol. 34, nos. 2–3, 1998, pp. 161–174.

[15] 3. R.K. Srihari, "Use of Captions and Other Collateral Text in Understanding Photographs," Artificial Intelligence Rev., vol. 8, nos. 5–6, 1995, pp. 409–430.