A Framework for Focused Image Crawler and Face Harvester

[1]Shatashi Bansal, [2]Abhinav Goel, [3]Manav Bansal
[1,2,3]Sir Chhotu Ram Institute of Engg. and Tech., Ch. Charan Singh University, Meerut, India

Abstract
The World Wide Web is a global, read-write information space loaded with Text documents, images, multimedia and many other items of information. Search Engines are important tools of information gathering from World Wide Web, if information is in the form of picture than it plays a major role to take prompt action and easy to use. There is a human tendency to retain more images than text. This paper presents a development of a image crawler that will focus on gathering images from the World Wide Web and identifies and extract Human Faces and store them for social uses such as making Face databases .and interests specific human face collections. To achieve Image crawling, we designed two crawler modules: a Web Crawler that searches the relevant web URL and a Face Detector that identifies the faces, crop it and stores them in a particular directory.

Keywords
Web Crawling, Focused Crawler, Image Retrieval

I. Introduction

The WWW encompasses a large amount of visual information such as structured collections (e.g., web pages) or independent collections (e.g., Images, photographs, logos, and so on) that comprises of a large amount of visual information such as videos, movies, and comic strips. Tools that are used for effective retrieval of this information can be proved to be beneficial for many applications. Here we try to show why such tools are indispensable for users, what services users may ask them to accomplish and what applications people may need them for. In order to retrieve and process an image, image web crawler uses a web crawling [1] technique that simply contains many systematic and typical processes. The whole information is retrieved using an image crawler by the use of just a single picture. It is advantageous in the field of internet surfing.

A. Image Crawler

An Image Web Crawler is a Focused Web Crawler [2] for browsing and retrieving images from a large database of web images on websites. Most traditional and common methods of image retrieval utilize some method of adding metadata such as captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. Manual image annotation is time consuming, laborious and expensive; to address this, there has been a large amount of research done on automatic image crawling. Additionally, the increases in social web applications have inspired the development of several web-based image crawling tools. Generally a web image crawler fetches the HTML source code for a given page and scans it for image references. It also finds links to other pages and puts them in a queue. Then pages are subsequently considered in queue.

B. Types of Image Crawler

On the basis of crawling techniques the following types of Image Crawlers are defined:-

- Visual Content Based Image crawler
- Text Keywords Based Image crawler

1. Visual Content Based Image Crawler

Visual Content based Image Crawler [3] retrieves the images on the basis of their visual properties. In this procedure, images are analyzed by their features such as color, texture, shapes and light intensity.

2. Text Keyword Based Image Crawler

Text Keywords based Image Crawler retrieves the images on the basis of the captions and name given to the images. For example:

Above the image name is 'laptop1.jpg' and caption is 'Sony laptop'.

In this paper, the Visual Content based Image Crawling technique is used to crawl images and harvests the Human face out of the collected images database.

II. Related Work

A. Image Crawling

The related work done by different authors on Image Crawlers is discussed below:
Neil C. Rowe (2002), Marie [4]: A High- Recall, Self-Improving Web Crawler that Finds Images Using Captions, had made an important progress in general image indexing from the Web by intelligent information filtering of Web text. By looking for the right clues, large amounts of Web page text can be excluded as captions for any given image, and the captions in the remaining text can be inferred. Clues can include caption candidate wording, HTML constructs around the candidate, distance from the associated image, image-file name words, and associated image properties. These clues reduce the amount of text to examine to find captions, and the reduced text can be indexed and used for keyword-based retrieval. But so far, the selection of these clues has been intuitive, and there has been no careful study of the relative values of clues.
Vadhri Suryanarayana1, Dr. M.V.L.N. Raja Rao, Dr. P. Bhaskara Reddy,Dr. G. Ravindra Babu [5], Image Retrival System Using Hybrid Extraction Technique, has described, A content based image retrieval system allows the user to present a query image in order to retrieve images stored in the database according to their similarity to the query image. Content based image retrieval method is used on digital image data set. The author evaluates the retrieval system based on Hybrid features. The texture features are extracted by using pyramidal wavelet transform and the shape features are extracted by using Fourier descriptor. And the hybrid technique is the combination of both texture and shape. The major advantage of such an approach is that little human intervention is required. It is ascertained that the performance is superior when the image retrieval based on the Hybrid features, and better results than primitive set.

B. Focused Crawling:

The main objective of focused crawling is to only crawl on a small set of the Web to discover the set of pages covering a certain topic. Because of the finite crawling resources such as time, network bandwidth and storage, focused crawling is essential. Focused crawler is composed of the two hypertext mining programs which are based on the keyword relevancy evaluation, they are:

- The classifier component evaluates the relevance of the page.
- The distiller identifies the hypertext links that points to many relevant pages.

III. Proposed Work

The architecture of our Focused image crawler and Face harvester is proposed in the following figure,

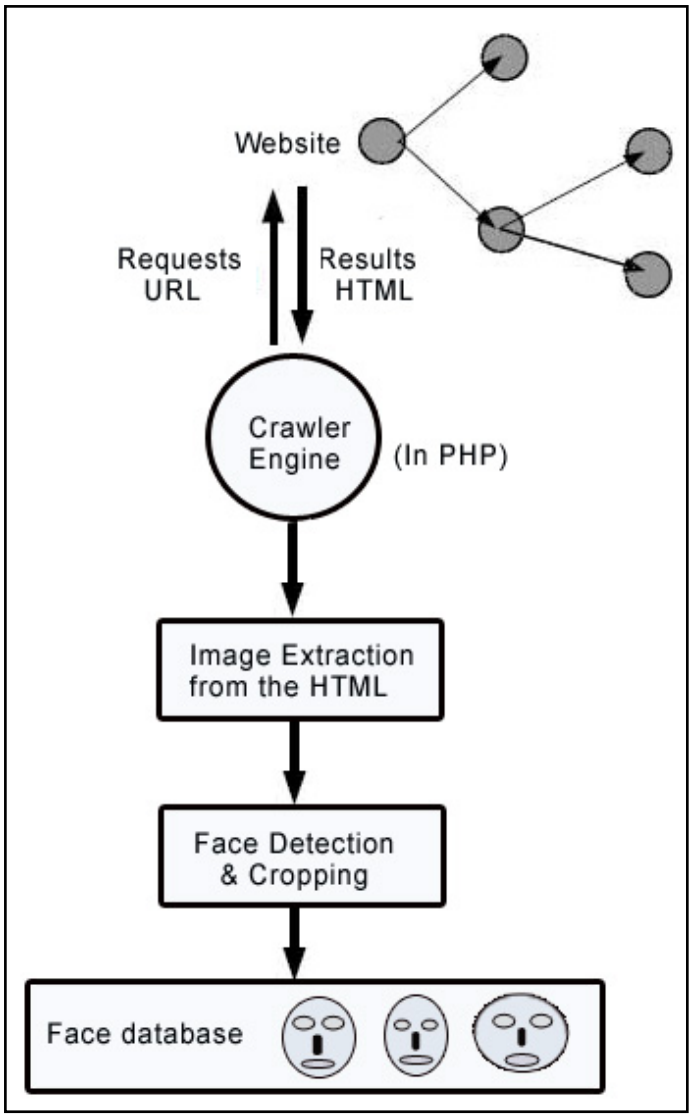

Fig. 1:

The mechanism of Image Crawler is technically same as that of a Web Crawler.

Web crawling involves interacting with thousands of Websites and Web Servers. The process starts from requesting the website URL from the web server by the crawler program and then extracting the image from the resulting HTML.

The Crawler program follows all links found in that HTML page. This leads to more links, which will be followed again, and again. A site can be seen as a tree-structure, the root is the start-URL, All links in that root- HTML-page are direct sons of the root. Subsequent links are then sons of the previous sons. Noting any hypertext links on that page that point to other Web pages, web crawler starts by parsing a specified Web page. They then recursively parse those pages for new links, and so on. Each resulting HTML from the pages is then passed to a Image Extraction Module which extracts images from the HTML.

Now the retrieved images are passed through the Face Detection routine and picture of Human face is extracted. The extracted Images are then stored in the directory.

IV. Working Methodology

The Web crawler engine and the image extraction and processing routine are designed as two independent modules with a minimal interface. The crawler application which handles the interface with the user and the Web is implemented using the PHP programming language.

The crawler engine is installed as Server Side Script which is accessible through a Web Interface.

Crawler fetches URL from a queue filled with websites URL's. Crawler process every URL contained in the queue and download the pages from URL. If the web page is downloaded successfully then the resulted HTML content is parsed for image tags. The images are then stored temporarily into a directory. These images are then passed to the Face detector routine.

Face detector routine uses a Face detection algorithm [6] [7] which detects whether or not an image contain frontal faces, even if the background is complex and cluttered. The faces detected in the images are then cropped and stored independently into the Faces directory.

A. Crawler Module

Steps:

1. Fetch website URL from a text file
2. Enter URL in the Crawler Routine
3. Requests the URL from the web server

The resulted HTML is temporary saved and passed to the Image Extractor routine

B. Image Extractor

The Image Extractor routine extracts the images from resulted HTML from the Crawler Module by detecting image tag(<img />).

These images are temporarily saved and passed to Face detector routine

C. Face Detector and Harvester/Cropper

The images saved from the Image Extractor module is passed to the Face Detector program.

This face detector program detects faces in the picture and crops the face part of the image.

D. Demo Screen Shots

Fig. 1: Shows the Screen of a Website

Fig. 2: Images are Extracted from the Web Page. Relevant Images Containing the Faces are Then Passed to the Cropping Routine

Fig. 3: Faces are Extracted and Stored in the Directory

V. Conclusion

Mostly Crawlers and Search engines are generic tools and try to satisfy everyone's search needs. They can't be used for specific work, especially for searching the image of a person. Internet users are feeling a need for highly specialized and filtered image based search where they can explore their curiosity for a specific image or person's image like famous people or celebrities.

This paper presents a prototype for focused image crawler and face detector which help in creation of the face database.

VI. Future Work

Future work will include extending the crawler efficiency by adding new objects (such as pedestrians, pets and cars) detection algorithms in the crawler.

References

[1] Web Crawler Introduction: [Online] Available: http://www.en.wikipedia.org/wiki/Web_crawler

[2] Chakrabarti, Dom B., Van den berg M., "Focused crawling: a new approach to topic-specific Web resource discovery", Journal of Computer Networks: The International Journal of Computer and Telecommunications Networking Computer Networks, Vol. 31, Issue 11-16, pp. 1624-1640, May 17, 1999.

[3] Boyd D, Heer J.,"Proceeding of the 2005 IEEE Symposium on Information Visualization", pp. 5, 2005.

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

[5] Vadhri Suryanarayana1, Dr. M.V.L.N. Raja Rao, Dr. P. Bhaskara Reddy, Dr. G. Ravindra Babu,"Image Retrival System Using Hybrid Extraction Technique".

[6] K.K. Sung, T. Poggio.,"Example-based learning for view-based human face detection", MIT AI Lab-Memo, No. 1521, 1994.

[7] Edgar Osuna, Robert Freund, Federico Girosi,"Support vector machines: Training Support Vector Machines: An Application to Face Detection", Proceedings of CVPR-97, June 1997.

Shatashi Bansal received his MCA from Punjab Technical University. She had also done 'O' Level and 'A' Level from DOEACC society. Currently she is pursuing M.Tech. in Computer Engineering from Shobhit University, Meerut. Her research interests include Web development and Programming.

Abhinav Goel received his B.Tech. degree in Computer Science from Uttar Pradesh Technical University, M.Tech. degree in Information Technology from Karnataka State Open University, and M.B.A. in Information Technology from Punjab Technical University, M.Phil from Manav Bharti University and pursuing Ph.d. in Computer Science from Monad University. He is working as Assistant professor in Sir Chhotu Ram Institute of Engineering and Technology, CCS University Meerut. His research interests include Web development, Object Oriented Programming, Robotics, Open Source Systems.

Manav Bansal received his B.Tech. degree in Computer Science from Uttar Pradesh Technical University, M.Tech. degree in Information Technology from Karnataka State Open University, and M.B.A. in Information Technology from Punjab Technical University, M.Phil from Manav Bharti University and pursuing Ph.d. in Computer Science from Monad University. He is working as Assistant professor in Sir Chhotu Ram Institute of Engineering and Technology, CCS University Meerut. His research interests include Design and Analysis of Algorithms, Theory of Formal Languages.