AN IMAGE CRAWLER FOR CONTENT BASED IMAGE RETRIEVAL SYSTEM

Purohit Shrinivasacharya[1], M V Sudhamani [2]

[1]*Siddaganga Institute of Technology, Tumkur, Karnataka, India,* ***purohitsn@gmail.com***
[2]*R.N.S Institute of Technology, Bangalore, Karnataka, India,* ***mvsudha_raja@hotmail.com***

Abstract

Images from the minute it was invented, has had an immense impact on the world we live in. The extracting the required images from the World Wide Web (WWW) is very difficult because web contains a huge number of images. To solve this problem we need a system that can retrieve the required images needed by the user. Image Crawler is a web based tool that collects and indexes group of web images available on the internet. This tool collects the keyword or phrase from the user to retrieve the images from the web. Then these collected keyword is applied to the different general search tools like Google, Yahoo, Bind etc,. The collected web page information is stored in the temporary file till 200KB file size from the server. Then this file content will be scanned and extract the image URL's and it is compared the URL which is present in the database to avoid the duplicate downloads. The extracted URL's images are downloaded and finally stores unique image and corresponding metadata like filename, url, size etc. in the database. In this paper we present the designing of an Image crawler tool. We build a search tool which is flexible, general-purpose image search framework and explore a directed result aggregating and removing of duplicates to achieve top results compared to other existing search tools. Finally this resulted images are used in the Content Based Image Retrieval (CBIR) system for extracting the relevant images need by the client using the content of the images rather than the text based information.

Keywords: *Image Crawler, Database, URL, Metadata, and Retrieval*

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

1. INTRODUCTION

Web crawlers are more or less as same as the web. The spring of 1993 Matthew Gray [6] writen a first web crawler World Wide Web named as "Wanderer" after a month the release of NCSA Mosaic, it was used since from 1993 to 1996 to accumulate statistics about the growth of the web. The David Eichmann [5] has written the first research paper the RBSE spider containing a squat explanation of a web crawler. The Burner has published a first paper that describes the web crawler architecture, it is the original Internet Archive crawler [7].The Google search engine architecture was presented in the Brin and Page's paper, this can be used as a distributed system of page-fetching method and a central database for coordinating the crawl. Brin and Page's paper becomes the blueprint for the other crawlers. A distributed and extensible web crawler designed by Heydon and Najork described Mercator [8,9], that has become the outline for a number of other crawlers. The literature includes the other distributed crawling systems PolyBot [10], UbiCrawler [8], C-proc [9] and Dominos [11]. The text retrieval systems use the ranking and reranking approach to extract the best result from the search copies[3,4].

Image retrieval is the process of searching and retrieving images from a huge dataset or WWW. As the image grows in the database or WWW, retrieval of the correct images becomes a difficult task and it is challenging. Most of the web based search engine uses the common methods of image retrieval exploit some method of accumulating the metadata such as file names, captioning, keywords or descriptions to the images constructed by human. Therefore, that retrieval can be performed over the annotation words rather than the content of the image. The method for finding the WWW images is nothing but browsing the several webpage and extracting the related text and file name extensions to identify the image. The well-known search engines [1] and directories are Google, Yahoo!, Alta Vista [2], Ask, Exalead and Bing etc. The text-based image retrieval systems only worry about the text described by humans,rather than the content of images. Our main aim is to implement the effective image search engine on WWW using the CBIR technique. To apply the CBIR method first we need the collection of images to construct the features database. In this paper we are presenting the technique to retrieve the images from the WWW using the text description, then these images are used for the CBIR system. The presently Google Image Search results are ranked on the bases of surrounding text of the image in a page.

1.1 Data Scope

The complexity decision of an designing the image search system is very difficult unless understanding the nature and scope of the image. The diversity of user-base and expected

user traffic on a search system is one of the influenced factors of designing the search system. Based on this dimension, search data can be classified as follows [12]:

- Archives: A collection of large numbers of semi-structured or structured homogeneous images relating to specific topics.
- Domain-Specific Collection: A collection of large homogeneous images allowing for access to restricted users with very specific objectives. Examples of such a collection are medical and geographic satellite images.
- Enterprise Collection: A collection of large heterogeneous of images that can be accessible to users within an intranet. Images are stored in different locations on the disk.
- Personal Collection: A large homogeneous collection of images and they are generally small in size, that can be accessible primarily to its holder or owner. These collections are stored on a local disk.
- Web- World Wide Web (WWW): A collection of large non-homogeneous of images, that can be easily accessible for everyone with an Internet connection. These image collections are semi-structured, and are usually stored in large disk arrays.

1.2 Input Query

The basic problem is the communication between an information or image hunter or user and the image retrieval system. Therefore, an image retrieval system must support different types of query formulation, because different needs of the user and knowledge about the images. The general image search retrieval must must provide the following types of queries to retrieve the images from the web.

1. Attribute-based : It uses context and or structural metadata values. Example:
    - Find an image file name '123' or
    - Find images from the 17th of June 2012
2. Textual: It uses textual information or descriptors of the image to retrieve. Example:
    - Find images of sunsets or
    - Find images of President of India
3. Visual: It uses visual characteristics (color, texture, shapes) of an image. Examples:
    - Find images whose dominant color is orange and blue
    - Find images by taking the example image.

As we mentioned the above query types uses the different image descriptor and requires a different processing method for searching the images. The image descriptor can be classified into the following types:

- Metadata descriptors: It depicts the image, as recommended in various metadata standards, like MPEG, CIDOC/CRM and Dublin Core, respectively. The metadata descriptors are again classified as:
    1. Attribute-based: context and structural metadata, such as dates, genre, (source) image type, creator, size, file name etc.,
    2. Text-based: semantic metadata, like title/caption, subject/keyword lists, free-text descriptions and/or the text surrounding images. The example html document contains the images and its related information.
- *Visual descriptors:* These descriptors are extracted from the image while storing and retrieving with related images.

2. IMAGE CRAWLER SYSTEM

A general image crawler system consists of the user interface model to accept the user query and web interface model to connect the WWW to collect the web pages that contain the images. From the collected web pages it extracts the text and metadata and stores in the database for further uses. The Fig -1 shows the general image crawler system.

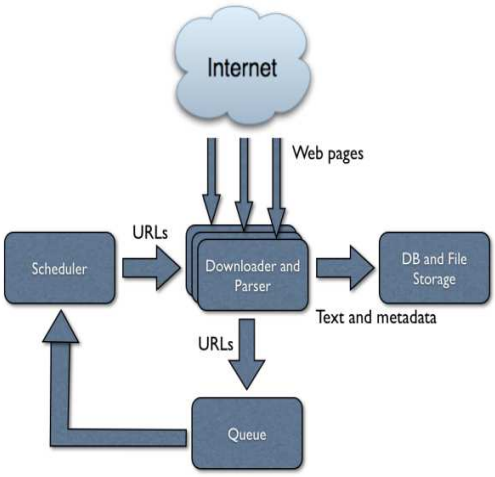

Fig -1. General Image Crawler Architecture.

3. PROPOSED IMAGE CRAWLER ARCHITECTURE

The proposed image crawler architecture consists the user interface to collect the query in the form of text or images itself. Once the keyword or image is taken from the user is fed into the web as a URL to Yahoo Image Search and Google Image search to collect the images from the WWW. The Fig -2 shows the proposed architecture.

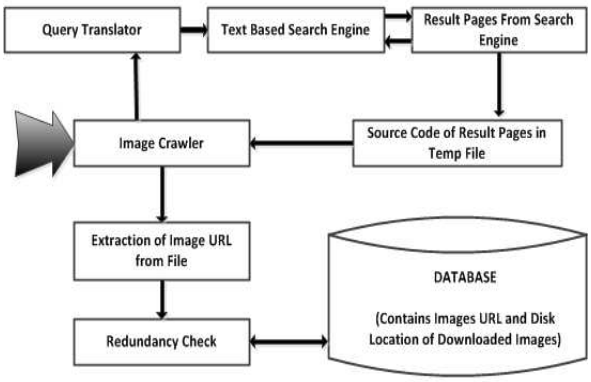

Fig -2. Proposed Image Crawler Architecture.

The description of the each module of the Fig -2 as follows:

- IMAGE CRAWLER: it is a search based tool where it requires only a keyword or phrase from the user to present the relevant images according to the user requirements.
- The tool "crawls" or "spiders" the web and then the user can browse through the search results.
- QUERY TRANSLATOR: The query is converted into the format specific to the search engine it is dealing with the object and the results are obtained in the form of an HTML page.
- TEXT BASED SEARCH ENGINE: The tool requires only a keyword or phrase from the user to present the relevant images according to the user requirements.
- REDUNDANCY CHECKER: Extraction of different urls leads us to the same content. As the check needs to be fast, all URLs are kept in memory, and are comparing character by character quickly
- DATABASE: These results are entered into a database sheet with the key as the url and the corresponding disk path.

Each search brings about 40 search results i.e. first page for each search engine. It can be updated based on option to query the next time. The search engine has its own error messages for when no results are found.

The tool accepts the user query and fed into the text based search engine to build the web page from the Yahoo and Google and this source code is extracted and finds the images URL's present this source code collected from the web. Then this URL is sent into the redundancy check tool to check whether this URL is already present in the database. If database consists the URL that URL will be rejected and finds other URLs. If the URL is not found in the database then that image is downloaded and stores that URL and image in the specified folder for the use of CBIR system. These processes will be repeated until some number of pages from the web. The process of finding images will not be always if the query is given, because first it checks is there any related images information is present in the database. If related information is not found in the local database then it will search from the web using the above mentioned method. The experiment was conducted in 7 to 8 pages to download the images.

3.1 Flow Chart and Pseudocode

The Fig -3 shows the flow chart of the proposed work.

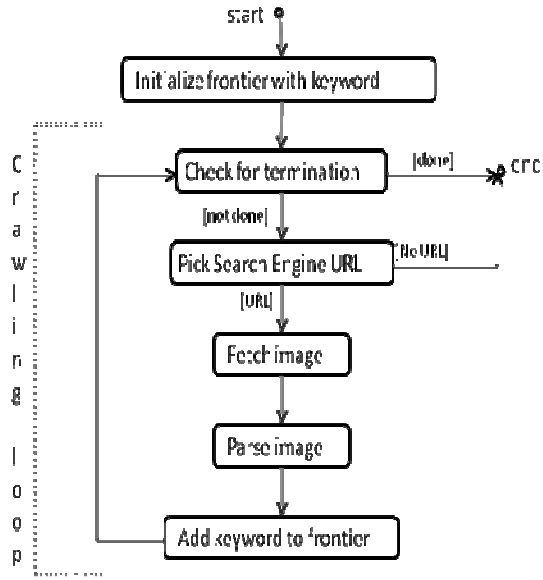

Fig -3. Flow diagram of Proposed work.

```
START
Enter the search query
Check for connection errors
Prepare the query string
Create the new image search
Start request to search engines
If Search is found
{
    For each source page
    {
        Extract the Source code and parse the result
        For each Source code
        {
          Extract image URL for corresponding page
        }
        End For
        Connect to database
        If no response
          Set connection error
```

```
        Else
        {
                Prepare database
                Check for updates if required
                Check for redundancy
                Insert URL and disk address
                Check for redundancy
                Download corresponding images
        }
        End If
        Check for termination
        If the termination condition reached
                Exit
        Else
                Increment the next source page
        End if
}
End for
}
Else
Display an error message
End If
END
```

4. EXPERIMENTAL RESULTS

This work is enforced mistreatment JAVA and Oracle SQL in Windows XP software. The analysis of the Image Crawler system is completed by submitting a text question to retrieve pictures from numerous classes of web pictures. We tend to conducted experiments on giving totally different keywords to extract the photographs from web. Once the keyword is submitted, it'll check its connected pictures area unit gift within the information or not. If information consists the connected pictures then it'll raise to update the information or terminate. If the choice is change the information then it'll search {the pictures|the pictures|the photographs} within the web to gather the new images and stores its data within the information. The table one shows the knowledge gift within the information/database once downloading the new image.

Table 1. Information present in the database

Image URL	Keyword	Image Loaction in Disk	Features for CBIR

The experimental results for different query text and corresponding resultant images are showed in Fig. 7, Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 12 and Fig. 13.

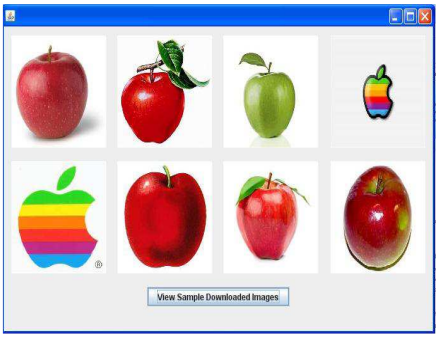

Fig- 4. Apple as a text query and its results on page 1

Fig- 5. Apple as a text query and its results on page 5

Fig- 6. Rajkumar as a text query and its results on page 1

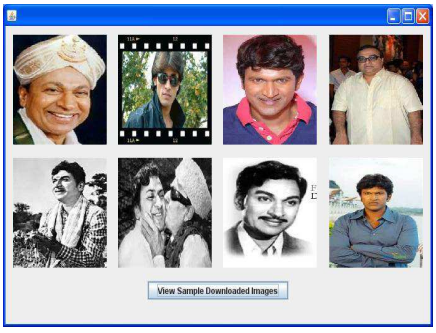

Fig- 7. Rajkumar as a text query and its results on page 3

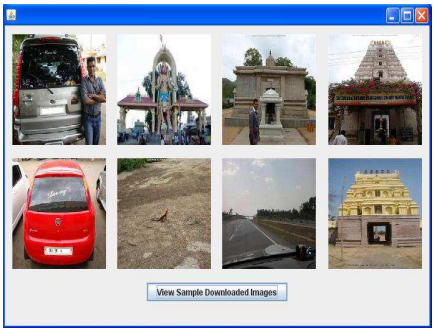

Fig- 8. Visiting Places in Tumkur as a text query and its results on page 1

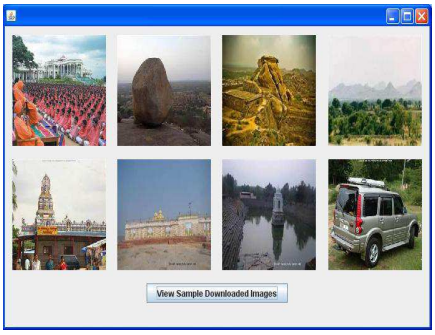

Fig- 9. Visiting Places in Tumkur as a text query and its results on page 4

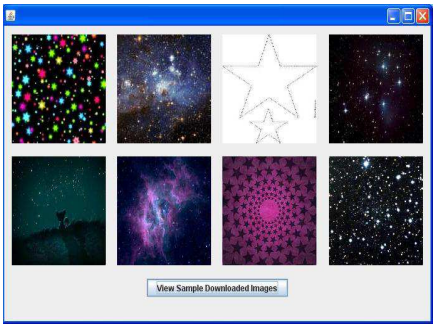

Fig- 10. Stars as a text query and its results on page 1

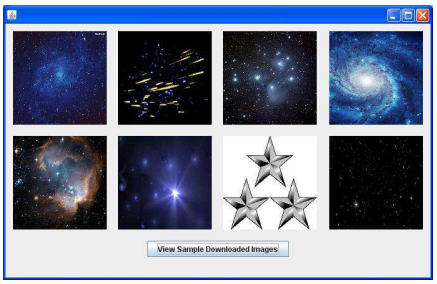

Fig- 11. Stars as a text query and its results on page 1

CONCLUSIONS

This paper presented an effective image crawler to crawl the images from the WWW by using different search engines. This tool collected the images and its corresponding metadata for later uses. The crawled images were best input for the content based image retrieval systems. It was observed that the performance this crawler was best for the CBIR system. The experiment was conducted with 1000 different text query for downloading the images from the different web sites. The enhanced reranking technique and giving the image itself as a query to extract the images from the Google and Yahoo needs to be adapted to get the attractive performances for feature work.

REFERENCES

[1] http://www.20search.com/image.php 20 SEARCH The Web's Best Image Search Engine List!

[2] http://www.altavista.com/ Altavista

[3] G. Salton and C. Buckley, Term weighting approaches in automatic text retrieval, Information Processing and Management, 24 (5): 513--523, 1988.

[4] Gerard Salton and Christopher Buckley Term Weighting Approaches in Automatic Text Retrieval, 323-327,1988

[5] Eichmann D. The RBSE Spider – Balancing effective search against web load. In Proc. 3rd Int. World Wide Web Conference, 1994.

[6] Gray M. Internet Growth and Statistics: Credits and background. http://www.mit.edu/people/mkgray/net/background.html

[7] Burner M. Crawling towards eternity: building an archive of the World Wide Web. Web Tech. Mag., 2(5):37–40, 1997.

[8] Boldi P., Codenotti B., Santini M., and Vigna S. UbiCrawler: a scalable fully distributed web crawler. Software Pract. Exper., 34(8):711–726, 2004.

[9] Cho J. and Garcia-Molina H. Parallel crawlers. In Proc. 11th Int. World Wide Web Conference, 2002, pp. 124–135.

[10] Shkapenyuk V. and Suel T. Design and Implementation of a high-performance distributed web crawler. In Proc. 18th Int. Conf. on Data Engineering, 2002, pp. 357–368

[11] Hafri Y. and Djeraba C. High performance crawling system. In Proc. 6th ACM SIGMM Int. Workshop on Multimedia Information Retrieval, 2004, pp. 299–306.

[12] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, Vol. 40, No. 2, April 2008, pp. 5:1 – 5:60.