An Improved Approach for Caption Based Image Web Crawler

Dhiraj Khurana[1], Satish Kumar[2]

[1]**Assistant Professor, CSE Department**
University Institute of Engineering & Technology
Maharshi Dayanand University, Rohtak(Haryana)
dhirajkhurana23@rediffmail.com

[2]**Assistant Professor, CSE Department**
Vaish College of Engineering, Rohtak (Haryana)
Krsk23@gmail.com

ABSTRACT

The World Wide Web [1] is a global, read-write information space. Text documents, images, multimedia and many other items of information, referred to as resources, are identified by short, unique, global identifiers called Uniform Resource Identifiers so that each can be found, accessed and cross referenced in the simplest possible way. It is a vast reservoir of information provides an unrestricted access to large inexhaustible pool of information, present in the form of hypertext documents formatted using Hyper Text Markup Language (HTML). These documents contain hyperlinks to other documents.

Keywords: Crawler, Optimization, Duplicate, Webpage, Prioritization

1. INTRODUCTION

1.1 Image Web Crawler

An I**mage Web Crawler** is a system for browsing; searching and retrieving images from a large database of web Images. 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 **[1]**.

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 (HREF, FRAME, AREA, and certain JavaScript constructs) and puts them in a queue; pages are subsequently considered in queue.

1.2 Types of Image Web Crawler:-

Based on method of crawling the images or matching the images on particular property, we defined the following types of image web crawler:-

1) Content Based web Image crawler
2) Keyword Based web Image crawler

1.2.1 Content Based web Image Crawler:-

Content-based web image Crawler is a type of crawler in which images described on the basis of their visual properties. In this procedure, images are analyzed by their low level features such as color, texture and light intensity **[10].**

1.2.2 Keyword Based web Image Crawler:-

Keyword- based web image crawler is a type of crawler in which image representation uses words to describe itself. Keyword- based retrieval is based on giving captions to images and retrieving images by querying with text and finding matches between query words and caption words **[10]**.

2. RELATED WORK

Here we are using the Distributed approach for A caption based Image Web Crawler. The focus on distribution occurred for crawling the images from web for three important reasons: scale, availability, and cost. Web-Crawler is a large system: it supports hundreds of queries per second against an index that takes days of processing to create. Once created, the index is read only, and can easily be distributed. Furthermore, the system must always be available downtime create will with searches and damages the web crawler business, finally because web crawler is at the core of business, minimizing cost is essential. The related work done by different authors on Web Crawlers are following:

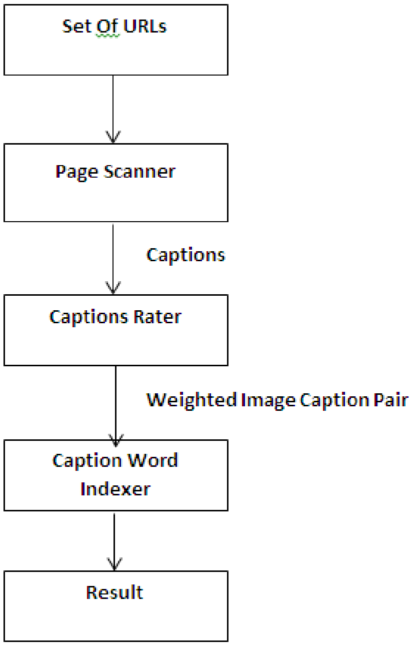
Neil C. Rowe (2002), Marie-4: A High-Recall, Self-Improving Web Crawler That Finds Images Using Captions, has as made important progress on general image indexing from the Web by intelligent information filtering of Web text. By looking for the right clues, large amounts of Web page text can be excluded 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.

Wei Ren, Maneesha Singh and Sameer Singh (2003), Image Retrieval using Spatial Context, has described as the retrieval of images is an important topic of research. With recent advances in multimedia technologies, a large amount of image and video data is becoming publicly available. However, without effective image retrieval it is not possible to make use of this information. A number of systems have been proposed commercially that allow for effective image retrieval on the basis of its texture, shape and color characteristics. These approaches are primarily based on low- level Characteristics of the image and their success depends on the quality of features used and the similarity criterion used. In this paper, we propose an image retrieval system based on modeling the spatial relationship between image contents.

Vadhri Suryanarayana1, Dr. M.V.L.N. Raja Rao, Dr. P. Bhaskara Reddy,Dr. G. Ravindra Babu (Feb 2012), 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.

2.1 **Proposed Architecture of Image web Crawler**

Each Crawl System contains the following parts:



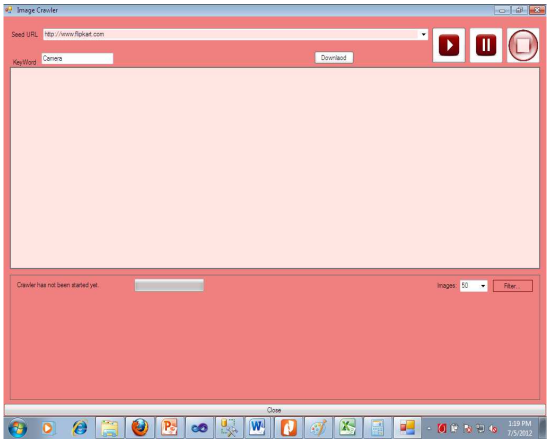
- **Set of URLs: -** Set of URLs contains all the urls those are given by the crawl manager to this particular crawl system.
- **Page Scanner:-** Page Scanner scans all the page locates all images on each page and good candidate captions for them. The page scanner searches for captions near each image reference on the HTML Page.
- **Caption Rater: -** A caption rater that assigns a likelihood to each image–caption pair on the basis of a weighted sum of factors [9].
- **Caption Word Indexer: -** The indexer indexes the inferred caption words in the form of keywords by providing images that match those keywords, sorted in order of decreasing likelihood of match.
- **Result:-** Finally the resultant images send back to the user via crawl manager.

The result of the proposed research is a caption based image web crawler. The analysis is being done on the basis of the total no of Images that the system found, Total no of images that the system retrieved and Total No of images that are relevant to the given keyword.

In the proposed system we have to give a seed URL and a Keyword to the system from the user interface as shown in various snap shot below.

3. RESULTS

Snap-Shot of Seed URL and Keyword Selection- In the following snap shot a user is giving the Seed URL and a keyword through the user interface-



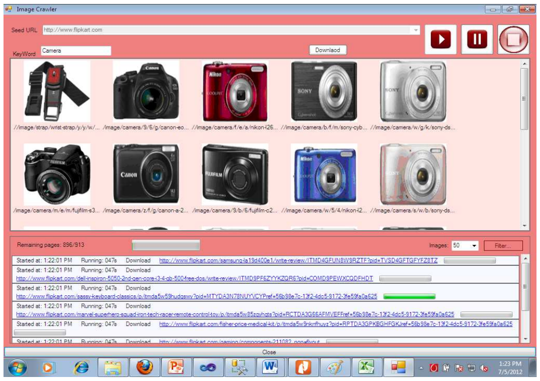
Seed URL http://www.flipkart.com
KeyWord Camera
Crawler has not been started yet.
Images: 50
Filter...
1:19 PM 7/5/2012

Fig. 1 Seed URL and Keyword Selection

Snap-Shot of Downloading Web Pages and Retrieving Images:- In the following snap- shot a Proposed Image Web Crawler started to Crawl the web and started to retrieve images relevant to the given keyword.

Snap-Shot of Detail Image View- On clicking on a particular image from among the retrieved images we get the a complete view of that particular image and also find the URL of the website from which that particular image is retrieved.

.



Seed URL http://www.flipkart.com
KeyWord Camera

//image/strap/wrist-strap/y/y/w/... /image/camera/9/6/g/canon-eo... //image/camera/f/e/a/nikon-l26... //image/camera/b/f/m/sony-cyb... //image/camera/w/g/k/sony-ds...

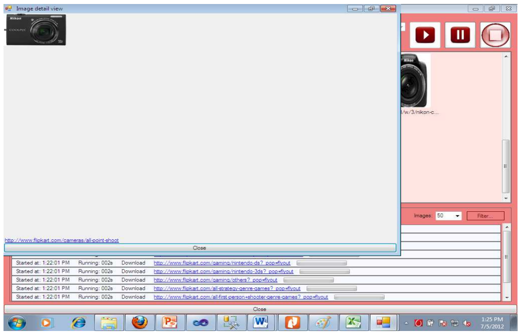
/image/camera/m/e/m/fujifilm-s3... /image/camera/z/f/g/canon-a-2... /image/camera/9/b/6/fujifilm-c2... //image/camera/w/5/4/nikon-l2... //image/camera/s/w/b/sony-ds...

Remaining pages: 896/913
Images: 50
Filter...

```
Started at: 1:22:01 PM   Running: 047s   Download   http://www.flipkart.com/samsung-la19d400e1/write-review/ITMD4GFUN8W9RZTF?pid=TVSD4GFTGFYFZ8TZ
Started at: 1:22:01 PM   Running: 047s   Download
http://www.flipkart.com/dell-inspiron-5050-2nd-gen-core-i3-4-gb-500-free-dos/write-review/ITMD9PF6ZYYKZQR6?pid=COMD9PEWXCQDFHDT
Started at: 1:22:01 PM   Running: 047s   Download
http://www.flipkart.com/sassy-keyboard-classics/p/itmda5w59hudgswy?pid=MTYDA3N78NUYVCYPref=56b98e7c-13f2-4dc5-9172-3fe59fa0a625
Started at: 1:22:01 PM   Running: 047s   Download
http://www.flipkart.com/marvel-superhero-squad-iron-tech-racer-remote-control-toy/p/itmda5w85zgyhqts?pid=RCTDA3G66AFMVEFFref=56b98e7c-13f2-4dc5-9172-3fe59fa0a625
Started at: 1:22:01 PM   Running: 047s   Download   http://www.flipkart.com/fisher-price-medical-kit/p/itmda5w9nknfhuyz?pid=RPTDA3GPKBGHFGKJref=56b98e7c-13f2-4dc5-9172-3fe59fa0a625
Started at: 1:22:01 PM   Running: 047s   Download   http://www.flipkart.com/gaming/components-21108?_pop=flyout
```

1:23 PM 7/5/2012

Fig. 2 Downloading Web Pages and Retrieving Images



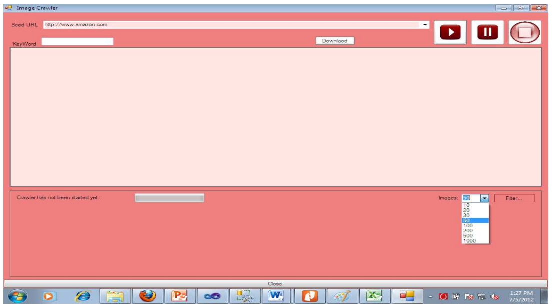
/w/3/nikon-c...
Images: 50
Filter...
http://www.flipkart.com/cameras/all-point-shoot

```
Started at: 1:22:01 PM   Running: 002s   Download   http://www.flipkart.com/gaming/nintendo-ds?_pop=flyout
Started at: 1:22:01 PM   Running: 002s   Download   http://www.flipkart.com/gaming/nintendo-3ds?_pop=flyout
Started at: 1:22:01 PM   Running: 002s   Download   http://www.flipkart.com/gaming/others?_pop=flyout
Started at: 1:22:01 PM   Running: 002s   Download   http://www.flipkart.com/all-strategy-genre-games?_pop=flyout
Started at: 1:22:01 PM   Running: 002s   Download   http://www.flipkart.com/all-first-person+shooter-genre-games?_pop=flyout
```

1:25 PM 7/5/2012

Fig. 3 Detail Image View

Snap-Shot of Setting No of Images to be Downloaded:- In the following snap-shot a user can set the No of images that is to be downloaded for a particular keyword.



Seed URL http://www.amazon.com
KeyWord
Crawler has not been started yet.
Images: 50
10
20
30
50
100
200
500
1000
Filter...
1:27 PM 7/5/2012

Fig. 4 Setting No of Images to be downloaded

3.1 PERFORMANCE ANALYSIS OF THE PROPOSED WORK:

1) Seed

http://www.flipkart.com

Keyword: - Mobile

Explanation:- Table 1 shows the performance analysis of the system for a keyword "Mobile" on the seed URL-http://www.flipkart.com, all the entries in the table corresponding to the first No of Images to retrieved, such as after retrieving the First 10 Images , The total no of images in the database are 90, and out of these 10 images 9 images are relevant to the given keyword.

Table 1. Performance Table Of keyword "Mobile

Total no of Images(a)	No of First retrived Images(b)	Relevent Image Retrived(c)	Precision P=(c/b)	Recall R=(c/a)	F-Measure F=2PR/(P+R)
100	10	9	0.90	0.10	0.18
150	20	18	0.90	0.15	0.26
180	30	28	0.93	0.20	0.33
200	50	43	0.86	0.22	0.42

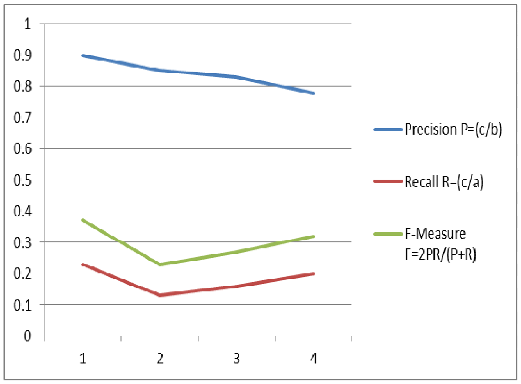


Fig. 5. Performance Graph Of keyword "Mobile"

3) Seed URL:

http://www.flipkart.com

Keyword: - Laptop

Explanation: - Table 2 shows the performance analysis of the system for a keyword "**Laptop**" on the seed URL-**http://www.flipkart.com**, all the entries in the table corresponding to the first No of Images to retrieved, such as after retrieving the First 10 Images, The total no of images in the database are 90, and out of these 10 images 9 images are relevant to the given keyword.

Table 2. Performance Table Of keyword "Laptop"

Total no of Images(a)	No of First retrived Images(b)	Relevent Image Retrived(c)	Precision P=(c/b)	Recall R=(c/a)	F-Measure F=2PR/(P+R)
90	10	9	0.90	0.10	0.18
119	20	18	0.90	0.15	0.26
137	30	28	0.93	0.20	0.33
163	50	46	0.92	0.28	0.42

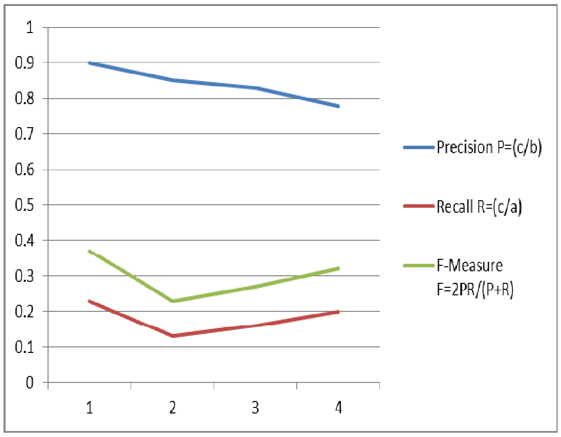


Fig. 6. Performance Graph Of keyword "Laptop"

4. CONCLUSIONS

We see that image retrieval by caption description achieves high precision and fairly high recall. The size of or database and the plethora of images in it guarantee that the results are very good. So the combination of a Distributed Approach of A caption based image web crawler results a high precision and high recall crawler with a good speed.

REFERENCES

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

[2] Gautam Pant, Padmini Srinivasan, and Filippo Menczer3, "Crawling the Web", 2004

[3] Leigh Dodds, "Slug: A Semantic Web Crawler," February 2006.

[4] Junghoo Cho, "Crawling The Web: Discovery and Maintenance of Large Scale Web Data", Ph.D. thesis submitted in November 2001 at Stanford university, USA.

[5] S.S. Dhenakaran1 and K. Thirugnana Sambanthan2, "Web Crawler – An Overview" International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267

[6] Sandhya, M. Q. Rafiq ," Performance Evaluation of Web Crawler" , IJCA Journal, 2011.

[7] Sunil M Kumar and P.Neelima. "Design and Implementation of Scalable, Fully Distributed Web Crawler for a Web Search Engine". International Journal of Computer Applications 15(7):8–13, February 2011.

[8] Zhixing GAO, Kunhui LIN, " Design and Implementation of an Efficient Distributed Web Crawler with Scalable Architecture", Journal of Computational Information Systems5:6(2009) 1817-1823,2009

[9] B. Dinakaran1, J. Annapurna2, Ch. Aswani Kumar1 "Interactive Image Retrieval Using Text and Image Content", Cybernetics And Information Technologies,Volume 10, No 3 Sofia , 2010.

10] Praveen Bandikolla ,Keshi Reddy, Vishwanath Reddy, "Image Retrieval Using a Combination of Keywords and Image Features", Master Of Science, thesis submitted in 2008, at Blekinge Institute of Technology, Swedan.