International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 49 - 55, 2016, ISSN:- 2319–8656

An Effective Approach for Document Crawling With Usage Pattern and Image Based Crawling

Ankur Tailang
School Of Information Technology
Mats University
Raipur,India

Abstract: As the Web continues to grow day by day each and every second a new page gets uploaded into the web; it has become a difficult task for a user to search for the relevant and necessary information using traditional retrieval approaches. The amount of information has increased in World Wide Web, it has become difficult to get access to desired information on Web; therefore it has become a necessity to use Information retrieval tools like Search Engines to search for desired information on the Internet or Web. Already Existing and used Crawling, Indexing and Page Ranking techniques that are used by the underlying Search Engines before the result gets generated, the result sets that are returned by the engine lack in accuracy, efficiency and preciseness. The return set of result does not really satisfy the request of the user and results in frustration on the user's side. A Large number of irrelevant links/pages get fetched, unwanted information, topic drift, and load on servers are some of the other issues that need to be caught and rectified towards developing an efficient and a smart search engine. The main objective of this paper is to propose or present a solution for the improvement of the existing crawling methodology that makes an attempt to reduce the amount of load on server by taking advantage of computational software processes known as "Migrating Agents" for downloading the related pages that are relevant to a particular topic only. The downloaded Pages are then provided a unique positive number i.e. called the page has been ranked, taking into consideration the combinational words that are synonyms and other related words, user preferences using domain profiles and the interested field of a particular user and past knowledge of relevance of a web page that is average amount of time spent by users. A solution is also been given in context to Image based web Crawling associating the Digital Image Processing technique with Crawling.

Keywords: WebCrawler, Page Ranking, Indexer, Usage Pattern, Relevant Search, Domain Profile, Migrating Agent, Image Based Crawling.

1. INTRODUCTION:

World Wide Web Is the largest hub for getting data related to any field. From the past few years it has become the major and the biggest means of getting information. Each n every day millions of pages get uploaded in the web related to some field, adding to the humongous number of millions pages already on-line. [1]As the rapid growth of World Wide Web from past ten years, it becomes difficult to get the desired information which user wants. The relevancy guaranteed by search engine is lack in accuracy. Search engine have some issue that need to be addressed to make it more efficient for the user, so that they can have more relevant page according to their previous requests. "This issue of search engine is like big number of irrelevant or unwanted links, topic drift and l server load" [6] that causes server failure. As with the help of search engine user query for their desired information, they generally entered some query with specific keywords what they wishes to access and search engine returns the list of URL's that are related to user keyword. Page rank ranks all the web pages according to their ranking to present in straighten out manner. Search engine may suffer many difficulties like
sometimes crawler download the irrelevant links and due to this quality of search engine reduces. To overcome this multiple crawling instances is introduced but it may results in network congestion also put extra burden on server.
Many algorithm like Page ranking [2] and HITS [3] etc. are used for ranking. But there is no contingency given to rank pages on the basis or previous relevance or relation of the page with respect to the particular query and user feedback. This paper proposed the work through which crawler give only the relevant or appropriate links by using migrants. The document which gets download are being ranked according to user related field and past knowledge about user visit on a web page and how much time the user spent on it. Whenever we retrieve a webpage the page might be containing images as well some time the images that are fetched are not related to the text associated with it , Image Based Crawling concept is an attempt to get rid of this problem and can affect the Page Relevance score if the image is according to the text or not.

2. PROPOSED WORK:

This work has been done with the provision of satisfying the objective of downloading only the relevant or the matching pages that are according to the searched topic or collection of topics and these pages have the capacity of providing the information related to the user query. In contrast to already existing crawling and ranking techniques Irrelevant or non matching pages are going to be ignored and only the links that consist of large amount of data according to the user search are presented to the user.

Major components of the system are **User Interface, Crawling Manager, Page Relevance Score, Indexer, Migrating agent, Context Manager.**

2.1User Interface

There will exist an interface through which the user will write their queries and ask the web server to serve them with the best possible result. "The user interface is can be defined as the medium where communication between human beings and machines occurs. The goal of this interaction is impressive operation and control of the system i.e. machine on the user's end, and feedback from the machine that is retrieving the information regarding the query, which aids the operator in making operational decisions. It is the part of Web Search Engine establishing communication with the users and allowing them to request and view responses." [4]

2.2 Crawling Manager

It is responsible for providing relevant pages according to prerequisite topic or set of topic it supplied with the set of seed URL's. To earn seed URL the query is submitted to the search engine and from the first n pages the important term appears is stored in D-table. These first n pages are treated as seed URL.

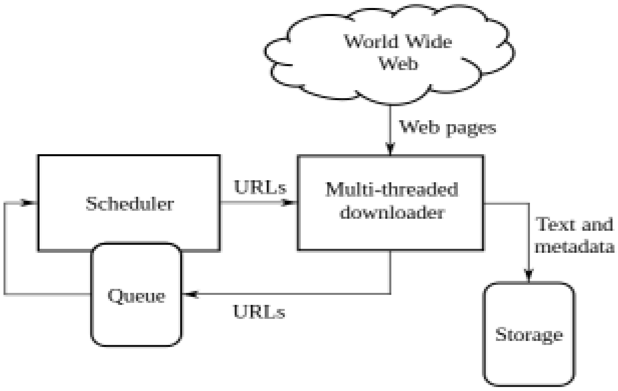


Figure1. **High-level architecture of a standard Web crawler [9]**

2.3 Working of Crawling Manager

It selects the URL from list of seed URL and calls the Migrating Agent [5], along with key table as it contains the keywords which are of high frequency in D Table and this is used to match with web page. The migrant extracts the web page from the web server and then relevancy score of each web page is calculated on the basis of how many terms from Key table appears in web page if value of relevant score of page is on the greater side against the decided threshold score page is considered to be a relevant page and if page is irrelevant migrant return to Crawling Manager. The Crawling manager will also search for the relevant images on the basis of image based crawling.

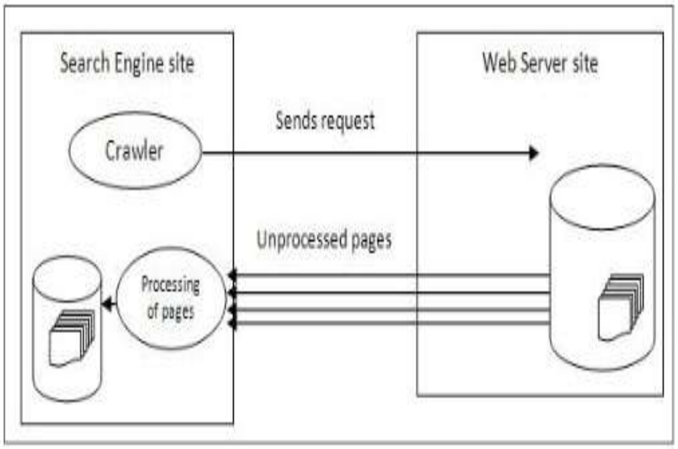


Figure 2.**Traditional Web Crawler [5]**

2.4 Migrating Agents

Migrating agents are computational software processes have the power of visiting large/wide area networks such as the internet, communicating with foreign organization, collecting information on behalf of its boss and reporting back after performing the duties assigned by its master. [5] Migrant returns the relevant web page to Crawler Manager, which further stores in local repository. Repository transferred the link to URL listed called REL URL's and indexed those pages. Using migrants (migrating agents), the procedure of gathering and filtration of web contents can be done at the server side rather than search engine side which can reduce the load on network caused by the traditional Web crawlers. Along with the Migrating Agents there can be another software process that keep on running this process is the process of Context manager that will keep an eye on the context in which the query has been asked, if the pages that are processed are related to context then its fine otherwise a message will be delivered to the Migrating Agent that a particular page is not related to the asked context.

Other components are:

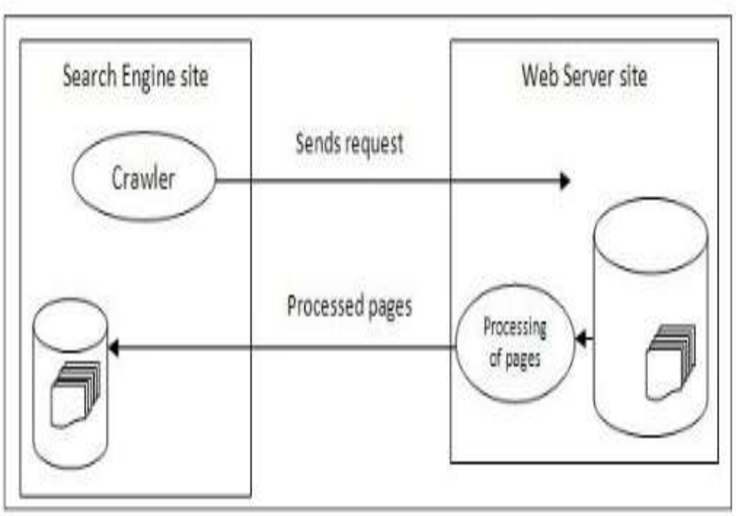
2.5 D–Table

There will be existing a structure that will be able to store all the terms that are present in the submitted query, D-Table (short for Data table) contains all the terms that appear in first n-pages for the submitted query. This table will contain all the words that are extracted from the user query after the stop words are removed.
The frequency for each term is also calculated using formula (1)
Fte = total number of presence of te in D-Table / Total number of words present in D-Table …. (1)
Where Fte = frequency of term te ,te=Term

2.6 Combination table

A Combination Table is a data structure that stores synonyms or related words (directly or indirectly) with respect to keywords. Sometimes Combination table can also be called as the Prospective table because it contains all the prospects on which a keyword can be identified.

Figure 3. **Web Crawler using Migrating Agents [5]**



2.7 Keyword table

It contains the keywords to be verified. This table is sent with the Migrant Agent. Topmost n-terms with
highest number of occurrence in D-Table are selected as keywords. Terms from the Combination table that are conventionally or unconventionally associated with each of the keywords selected in D-Table are included in the key-Table.

2.8 Page Indexer

The Page Indexer indexes the web pages in a five table column containing information about the caption, prospective captions for a given caption, URL's where the caption and combination captions appear,
The work of Indexer is to collect the data, analyze the collected data and store the data into the proper format that enables a fast and efficient retrieval of data whenever required."**Stores and indexes information on the retrieved pages**."[1]

2.9 Page Relevance score

Sometimes some WebPages contains all the needed information in the form of pdf, doc files or sometimes in image files as well, so when calculating the page relevance score there might occur a page that might be containing a pdf file or an image file that contains the data related to the user's query. So we will also derive a formula that will match the terms with the keyword-Table by extracting the terms from an Image file, any Pdf file, any Ppt or Word file. The numbers 0.2,0.4,0.1 and 0.3 are taken on the basis of the content importance and also remembering the fact that their summation should be equal to 1.0 , so whatever the number it gets multiplied to should be equal to that number itself.

a)PageRelScore=0.2*TextOfURL+0.4*TextOfMeta +0.1*TextOfHead+0.3*TextOfBody

Another approach that can be applied to find out the usage pattern of the user is to allow the search engine to have a look at the user's search history or the links that have been bookmarked my the user, or the most recent pages that have been visited by the user, or the pages that have been made as the welcome page by the user. The Query that has been entered by the user is taken and the words or the terms are extracted from the query and then the words are matched again the terms that are present there in the Bookmarked or the links present in the history tab and the pages that are most frequently visited by the user gets retrieved or the pages that are related to those pages gets retrieved. Adding User's Personalization and the structure of the page residing on the web can help in better and more prominent Information Retrieval as it is the sole motto of the Crawler and the Search Engine.

b)PageRelScore=0.5*BookmarkLinktext+
0.3*historylinktext+0.2*newtablinktext

Text of URL: It contains the text of the outgoing links that are associated to a particular page. The text it is containing should be clear and should be relative to the Link it is pointing to.

Text of Meta: Meta Text generally contains all the keywords that are present in the document and these keywords play a very prominent role when the user searches for a topic on the web. It also contains description regarding the document.

Text of Head: The title of the page is placed inside the Head, and this is the area where one of the important keyword related to our document should be presented, it helps the search engine to search the page easily if the title is according to the corresponding document.

Text of Body: This is the area where all the information regarding the document is present or we can say that it is the place where the actual content resides.

Book Mark Link Text: Text associated with the bookmarked links.

History Link Text: Text that is associated with the links present in the History tab.

New Tab Text: The users tend to personalize their search engine home window so the pages that are present there can be taken and from their links the text can be extracted.

Where, TextOfURL= No.of keywords in Keyword-Table that occur in Web Page URL / Total number of terms in TextOfURL

Text Of Meta= No. of keywords in Keyword-Table that occur in Meta tag of web page / Total number of terms in Meta Text

Text Of Head= No. of keywords in Keyword-Table that occur in Head tag of web page / Total number of terms in Text Of Head

Text Of Body= No. of keywords in Keyword-Table that occur in Body tag of web page / Total number of terms in Text Of Body.

Book Mark Link Text: No. of keywords in Keyword-Table that occur in Body tag of web page / Total number of terms in Text Of BookMark Link.

History Link Text: No. of keywords in Keyword-Table that occur in Body tag of web page / Total number of terms in Text Of History tab links.

New Tab Text: No. of keywords in Keyword-Table that occur in Body tag of web page / Total number of terms in Text Of New Tab Links.

In order to define an initial value for the page relevance, let us assume that at least 50% of the contents will get matched to the contents of Keyword Table.

Then by formula:
c)PageRelScore=0.2*TextofURL+
0.4*TextOfMeta+0.1*TextOfHead + 0.3*TextOfBody[7]
=0.2*(0.5)+0.4*(0.5)+0.1*(0.5)+0.3*(0.5)
=0.1+0.2+0.05+0.15
=0.5
And
d)PageRelScore=0.4*BookMarkLinktext
+0.3*HistoryLinktText+0.3*NewTabText
=0.4*0.5+0.3*0.5+0.3*0.5
=0.2+0.15+0.15
=0.5

This value of 0.5+0.5/2= 0.5 mean, of PageRelscore is being used as the initial value and will act as a threshold for all the pages for their relevancy.
If the fetched web page contains any pdf file or ppt file or any document file (word):

e)PageRelScore=0.1*TextofURL + 0.2*TextOfMeta+0.1* TextOfHead +0.3*TextOfBody+0.3(textinpdf+textinppt+textinword)

3. IMAGE BASED WEB CRAWLING:

Search engines are some of the most popular sites on the World-Wide Web. However, most of the search engines today are textual; given one or more key-words they can retrieve Web documents that have those keywords. Since many Web pages have images, selective image search engines for the Web are required. There are two major ways to search for an image. The user can specify an image and the search engine can retrieve images similar to it. The user can also specify keywords and all images relevant to the user specified keywords can be Retrieved [8].Here in this paper I am proposing a new concept for the purpose of Image Based Crawling over the web. In this concept first of all the traditional web crawler will find out the major source of Image from where the images can be taken according to the need and requirement of the user.
1) The rich source of image will be found out by the crawler according to the content that has been specified in the page regarding the resultant images.

2) The second task will take place by performing the Image segmentation using the Image processing techniques. The image that has been retrieved the large number of times and is very much popular among the use in a particular category will be taken and gets segmented.

3) In order to segment an image mostly the Morphological Image Processing will be practiced.
Morphology in image processing is a tool for extracting image components that are useful in the representation and description of region shape, such as boundaries and skeletons. This is middle level of image processing technique in which the input is image but the output is attributes extracted meaning from an image.

4) Once the Boundary of The Image gets extracted the next task is to be performed is to represent the detected shape using the chain codes. The chain codes will create a representation of an image and they help a system (in our case the crawler) that the resulting image is of a particular object and will return only those images that satisfies the shape that is stored in the disk. This will help in retrieving only the relevant images to the query made by the user and won't allow anyone to make a fake label to some other image. For example, the image in the page is of DOG and in the "alt" "title" or "name" attribute of image tag "Tiger" is written.

Above proposed work can be added into the existing crawler or can be used for inventing a new crawler for crawling images only.

4. PROPOSED ALGORITHM:

A step-by step working of the proposed system is given below:

1: First of all a query will be fired by user to any of the popular search engine to retrieve first n pages. The URL's of these pages will serve as Seed URL's.

2: The next step is to remove the Stop words from each page and each term appearing on each page is extracted and stored in D-Table.

Note: The Stop Words are very large in number and affect the page rank of a particular webpage, thus it is the duty of crawler to leave out the stop words while crawling. Examples are of, and, the, etc., that provide no useful information about the documents topic. The process of removing these words is called Stop word removal. Stop-words account for about 15-20% of all words in a typical document. These techniques immensely reduce the size of the search engines index. [4]

3: Then, the Frequency of each term will be calculated using formula (1).

4: The Top n terms having maximum term frequency are selected as keywords. The keywords and their prospective/related terms are stored in key table.

5: Now, The crawler starts with seed URL's. The crawler manager calls the Migrating agents and transfers them to the respective web sites of these
Seed URL's. The migrant also takes the Key table with it for grouping purpose.

6: At the Web-site the Migrant agent retrieves the web page and with the help of HTML Parser parses the document. Then it calculates Page score to determine whether the page is relevant or not. If the retrieved pages contain Images as well then the Images will be checked morphologically and the contents will be matched according to the found images.

7: If the Mean is greater or equal to the initial score that we have already been calculated, the page qualifies to be a relevant or an appropriate page otherwise page is irrelevant page and Migrant Agent ignores the page. The Images that have been checked above if found relevant then they will add positives to the page relevance score otherwise if they are unrelated then the page relevance score will go down hypothetically.

8: The Migrant Agent then transfers the pages that satisfy the user objective and extracted links to a local repository i.e. the database at server side.

9: The Indexer then indexes the pages in the local repository.

10: At the last step the page rank according to the relevancy will be assigned.

5. FUTURE PROSPECTS :

1. The Segmentation of the image can be done on the basis of different-different sizes of the images and from different angles.
2. A Crawler that will only crawl the images over the internet can be developed.
3. Now a day's number of websites are existing over the web, that are having the domain names in Hindi, and when a person types in the name in Hindi the existing crawling technique doesn't really been able to fetch the page and in the URL bar the URL gets written in the form of Unicode's, so this problem can be addressed in near future.[10]

6. ADVANTAGES

1. This proposed technique of using Migrant technology for downloading Relevant document help to reduce load on server.
2. Improve quality of result.
3. Only the relevant pages that are containing the information related to the query are fetched and improve the response time.
4. Image based crawling gets improved as the shape gets checked and reduces the no. of unwanted images gets retrieved , if the name in img tag is mentioned wrong.

7. RESULT

Provide better quality result based on user's preference and provided requirements.

The process of Crawling is implemented In such a way that the pages that are only Relevant will be retrieved.

The Images will be retrieved are more related To the contents that they have been before.

8. CONCLUSION

Web Crawler act as the most important technique of the existing applications that helps in the process of Information Retrieval over the web, thus providing a methodology to improve the quality and working of the Crawler, it can provide much better Results as it are based on user's preference.A technique to improve the retrieval of images is also proposed that alongside with Image processing technique can really prove helpful in the future.

9. REFERENCES

[1] Hema Dubey, Prof. B. N. Roy "An Improved Page Rank Algorithm based on Optimized Normalization Technique" of Department of Computer Science and Engineering Maulana Azad National Institute of Technology Bhopal, India.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. Technical report, Stanford University, 1998.

[3] Kleinberg, Jon (December 1999). "Hubs, Authorities, and Communities". Cornell University. Retrieved 2008-11-09.

[4] Ms. Nilima V. Pardakhe[1], Prof. R. R. Keole[2] "An Efficient Approach for Indexing Web PagesUsing Page Ranking Algorithm For The Enhancement Of Web Search Engine Results" of

1) S.G.B.A.U., Amravati, H.V.P.M. College of Engg., Amravati, McMahons Road, Frankston 3199, Australia and

2) Department of Computer Science and Engineering, S.G.B.A.U.,Amravati, H.V.P.M. College of Engg. Amravati,

[5] Managing Volatile Web Contents Using Migrating Agents.

[6] A. Gupta, A. Dixit, A. K. Sharma, "Relevant Document Crawling with Usage Pattern and Domain Profile Based Page Ranking", IEEE, 2013.

[7] ReviewonDocument Crawling With Usage Pattern and Page Ranking Akanksha Upate1 and Surabhi Rathi2

1) Final Year, Computer Science Department, J.D.I.E.T, Yavatmal, India,
2) Final Year, Computer Science Department, J.D.I.E.T, Yavatmal, India,

[8] Crawling for Images on the WWW
Junghoo Cho1 and Sougata Mukherjea
Department of Computer Science, Stanford University.

[9] https://en.wikipedia.org/wiki/Web_crawler#/media/File:WebCrawlerArchitecture.svg

[10] http://hindi.thevoiceofnation.com/technology/hindi-domains-could-soon-become-a-widespread-reality-if-digital-india-club/