ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 4, April 2016**

# Review of Ranking Algorithms

Rinki Tyagi  Deepali Dev

M. Tech Student, Dept. of C.S.E, A.B.E.S Engineering College, U.P.T.U, Ghaziabad, India

Asst. Professor, Dept. of I.T, A.B.E.S Engineering College, U.P.T.U, Ghaziabad, India

**ABSTRACT:** Focused crawler is basically a type of crawler which navigates the web and collects information from the web according to the requirement entered by the user. This paper is just a review paper in which we will focus only the types of Ranking algorithms used by focused web crawler, their advantages and disadvantages and how relevancy is calculated by using various Ranking algorithms.

**KEYWORDS***:* Ranking Algorithms.

## I. INTRODUCTION

As we all know that we are living in 21st century and whenever we need any information, our first choice is to use web. In traditional time when web crawler was not there then it was very difficult to collect relevant information but when web crawler was developed then it becomes easy to collect relevant information but there were many problems faced by web crawler but our first question is what is meant by web crawler??? Web crawler is a set of instructions that collects only relevant information. Now next question is what is focused web crawler??? Focused web crawler is a type of web crawler which only searches the information according to the topic or keywords given by user. So another name of focused web crawler is topic specific crawler [9]. Some of the features of focused web crawlers are it should be robust, scalable, performance, efficiency, quality. Today's youth spend most of the time on internet and let us understand it by seeing a graph that is given in this section:
Here graph shows the time spent by users on the internet in different countries. Philippines is the country in which mostly people spend time on internet by using desktops and mobile phones in a month and Japan is the country where less number of people spend time on internet in a month.

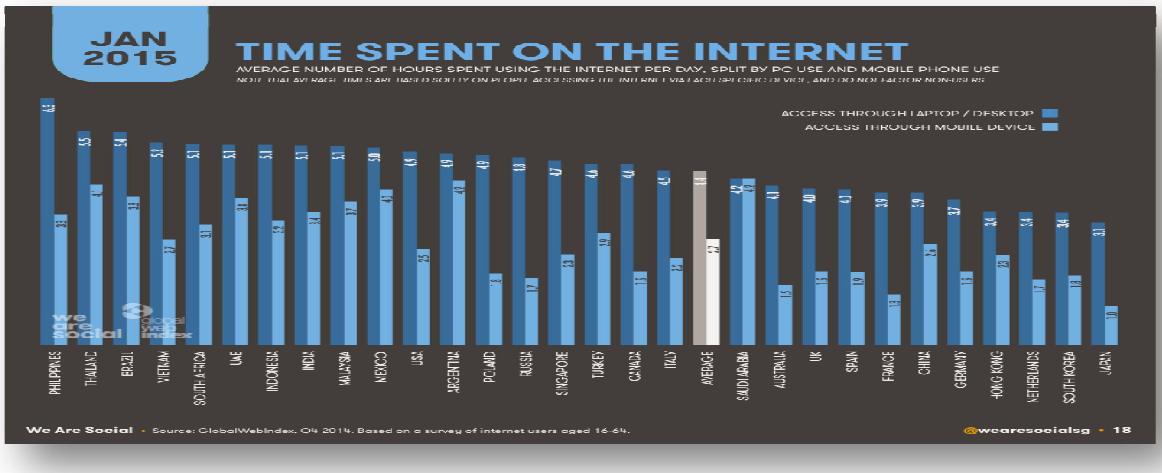


Figure (1): Graph showing the time spent on an internet.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 4, April 2016**

Web crawler uses the concept of graphs [8] i.e. vertices of a graph are taken as the web pages of a crawler and edges between the nodes of a graph are taken as hyperlink between the graphs. There are many algorithms used by a web crawler which we will define in next section of this paper.

## II. TERMS USED IN RANKING ALGORITHMS

Some of the terminologies used by ranking algorithms based on links are defined below in this section:

**INLINKS:**
They are also known as inbound links. These links are from outside the website to inside the website [5]. These links are used to increase the value of a page.

**OUTLINKS:**
These are also known as outbound links [5]. These links are present from the page to that page i.e. is present outside the website.

**DANGLING LINKS:**
This is the link that gives the link to the page that has no outgoing link [5].

**DEAD ENDS:**
These are the pages without any outgoing links[3].

**SPIDER TRAPS:**
When there is no link found from inside the group to outside the group [3] then it is known as spider trap.

## III. FOCUSED CRAWLER ALGORITHMS

In this section we will discuss about various algorithms used by focused web crawler.

**BLIND SEARCH ALGORITHM:**
Second name of this algorithm is BREADTH FIRST SEARCH ALGORITHM [6]. It is known as blind search because here FIFO queue is used which will navigate some useless pages also because there is no priority queue.

**NAÏVE BEST FIRST SEARCH:**
This algorithm is an advance version of Breadth First search. Only difference between Breadth First Search and Naïve Best First Search is that it uses the concept of priority queue. So, priority is assigned to the URL's in crawler frontier.

**FISH SEARCH ALGORITHM:**
It is a dynamic search algorithm which uses the only 0 and 1 to check relevancy [7]. When algorithm gives 1 as output then crawler predicts that the data is relevant and 0 represent that the value is relevant but it gives only 0 and 1 as output and cannot differentiate about most and less relevant pages.

**SHARK SEARCH ALGORITHM:**
It is a second dynamic search algorithm and it is used to check the relevancy and uses the values between 0 and 1 [6]. So it overcomes the disadvantage of fish search algorithm.

**PAGE RANK ALGORITHM:**
This algorithms works on the concept of predicting the relevancy by using links [6]. It does not depend on any query input by user. It calculate score by using rank but it's disadvantage is that it calculates the rank by using the popularity of page. We will define various types of ranking algorithms later in this paper from which page rank algorithm is the one so we will study it deeply in next section of this paper.

## IV. TYPES OF RANKING ALGORITHMS

Various types of rank algorithms are as follow:

1. HITS (Hyper Text Induced Topic Search) Algorithm.
2. Page Rank Algorithm.
3. Weighted Page Rank Algorithm.
4. Topical Page Rank Algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

**HITS (HYPER TEXT INDUCED TOPIC SELECTION) ALGORITHM:**

It is an algorithm which is completely dependent on links not on content. It is used to calculate the ranks for the pages that are extracted from WWW (World Wide Web). Some of the difficulties faced by this algorithm is that overhead cost increases because of gain in weight by those web sites that are not highly relevant [2]. Major disadvantage of this algorithm is that it assumes that all the links have equal weight and because of this assumption it fails. So to overcome the problem of HITS algorithm CLEVER and PHITS comes into existence. CLEVER is an advance version of HITS algorithm. Gives more efficient solution as compared to HITS and assigns weights to each link and PHITS solves the problem of HITS that links consist of equal weights. Some of the advantages of HITS algorithms are as follows:

1. It is a sensitive algorithm.
2. Pages are extracted using hubs and authorities values.
3. It gives vertices of hubs and authorities in a web graph more precisely.

Some of the disadvantages of HITS algorithms are as follows:

1. Query time is more because of which this algorithm become expensive.
2. Sometimes, it gives irrelevant authorities and hubs.
3. It is less feasible.
4. Topic drift takes place.

**PAGE RANK ALGORITHM:**

It is the most important algorithm and that is used to calculate the rank of page. Surgery Brin and Larry Page were the first to discover this algorithm [2]. Damping factor is generally set to 0.85.

$$PRank(p) = (1-d) + d\left(\frac{PRank(T1)}{c(T1)} + - - - - + \frac{PRank(Tn)}{c(Tn)}\right)$$

Where:

PRank(p)= Page rank of p.

c(T1)= Total number of outgoing links from page T1.

d= Damping factor.

Some of the advantages of page rank algorithm are as follows [3]:

1. Query time is less, gives response to the query of user very quickly.
2. When talking about localized links, it is less susceptible.
3. More efficient.
4. More feasible than HITS algorithm.

Some of the disadvantages of Page Rank algorithms are as follows[3]:

1. Less relevant to the query of user as it ignores that the page is relevant or not.
2. Here Rank Sink problem occurs because of infinite network.
3. When compared to the huge internet this algorithm is not too fast.
4. It cannot handle dead ends.
5. Spider trap basically occurs in Page Rank algorithm. When there is no link found from inside the group to outside the group then it is known as spider trap.

Implementation of Page Rank Algorithm is defined as follows:

1. Set the rank of page 1/m.
   Where:
   m= total no. of pages.
   With the help of an array we can represent these pages:
   A[j]=1/m where $0 \leq j < m$.
2. Set the value of damping factor $0<d<1$.
3. Repeat for each vertex j such that $0 \leq j < m$, Prank[j] ← 1-d.

For all pages p so that R links to Prank[j], do,

Prank[j] ←Prank[j]+d*A[R]/$R_m$.

$R_m$= total number of outgoing links from page R.

4. Perform modifications in the value of A.
   A[j] =Prank[j] for $0 \leq j < m$.
   Repeat step 3 unless you will get the same values of two consecutive iterations in terms of page rank.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 4, April 2016**

**WEIGHTED PAGE RANK ALGORITHM:**

This algorithm was proposed by Wenpu Xing and Ali Ghorbani. This algorithm is an improved version of Page Rank algorithm. Here rank value is not divided among outgoinglinks, here greater rank value is given to large important pages. Here two types of weights are used outgoing weights and incoming weights [1].

$W^{in}_{(p,a1)}=I_{a1}/(I_{a1}+I_{a2})$

$W^{out}_{(p,a1)} = O_{a1}/(O_{a1}+O_{a2})$

Where:

Page p has two references a1 and a2.

$I_{a1}$ and $I_{a2}$ are incoming links from references a1 and a2.

$O_{a1}$ and $O_{a2}$ are outgoing links from references a1 and a2.

So modified formula for weighted page rank is as follows:

$PRank(p) =(1-d) +d \sum_{v\in B(p)} PR(v) W^{in}_{(v,p)} \quad W^{out}_{(v,p)}$

**TOPICAL-PAGE RANK ALGORITHM:**

It is an extended version or we can say that improved version of Page Rank Algorithm. It is related to the topic so it uses the concept of topical random surfer model. Concept of content vector is used in it, content vector is represented by $C_i$ [4], best first concept is used in this crawler, pseudo code 0f this algorithms is as follows:

```
T-Prank (top, start_url's, freq) {
For each link (start_url's) {
  Enqueue (Front_link)}
While (visited< Max_Pages){
 If (multipiles(visited,freq)){
  Recomputed-score-T-Prank}
Link=dequeue-top-link (frontier)
Doc=fetch (link)
Score-similarity=similarity (topic.doc)
Enqueue(buffered-page,doc,score-similarity)
If(buffered-pages= Max_buffer){
Dequeue-bottom-links (buffered-pages)}
Merge(frontier,extracted-links(doc),score-T-prank)
If(frontier>Max_buffer){
 Dequeue-bottom-links (frontier)}
   }
}
```

## V. CONCLUSION & FUTURE WORK

This paper contains deep study of various algorithms used by focused web crawler. This paper contains mostly all the linking based algorithms. Various advantages and disadvantages of linking based algorithms. So this paper can be used by researcher for literature review. So, in future we can further improve performance of weighted page rank algorithm or can make it content as well as link based algorithm by implementing it with content based algorithm.

## REFERENCES

1. Weighted Page Rank Algorithm, Wenpu Xing and Ali Ghorbani, 0-7695-2096-0/04©2004IEEE.
2. Comparative Study of Page Rank Algorithm With Different Ranking Algorithms Adopted By Search Engine For Website Ranking, Mridula Batra and Sachin Sharma, IJCTA| Jan-Feb 2013.
3. Comparative Study of Page Rank and Weighted Page Rank Algorithm, Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, IJIRCCE Vol. 2, Issue 2, February 2014.
4. Improvement of Page Rank for Focused Crawler, Fuyong Yuan, Chuniax Yin, Jian Liu,0-7695-2909-7/07©2007IEEE.
5. http://en.wikipedia.org/wiki/Page Rank.
6. Shark Search Algorithm Using Page Rank Algorithm, Jay Prakash and Rakesh Kumar/ Procedia Computer Science 48(2015)210-216.
7. Focused Web Crawling Algorithms, Andas Amrin, Chunlei Xia, Shuguang Dai, Journal of Computers, Volume 10, Number 4, July 2015.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 4, April 2016**

8. An Improved Topic Relevance Algorithm for Focused Crawling, Hong-Wei Hao, Cui-Xia Mu, Xu-Cheng Yin, Shen Li, Zhi-Bin Wang,978-1-4577-0653-0/11©2011 IEEE.
9. Web Crawler: Extracting the Web Data, Mini Singh Ahuja, Dr. Jatinder Singh Bal, Varniea, ISSN:2231-2803(IJCTT).