ISSN:0975-9646

An Accurate Surfer Model to Enhance Web Search Criteria

Neerumalla Swapna^{1}, S.Bhavani^{2}
^{1,2}Associate Professor,
Department of Computer Science ,
VBIT,Gatkhesar

Abstract - **This paper enumerates about a Surfer Model which will improve a Search Engine's accuracy in fetching only the valid spam free pages. This will give a suitable solution for the present most unresolved problem which has been faced by the world net surfers for the last few decades i.e. 'spam pages'. Generally while we search most users tend to concentrate on the first few search results, so getting a place at the top of the search list is highly competitive. A Ranking System assigns a rank or score to every web page. The higher the page's score, the top the page will be placed in the search results list. This is simply based on the links going to and from a Web page. While there are a few tricks webmasters can use to improve their webpage ranks to get at top spot in the result list. With these tricks it not only wastes time of all web surfers, but even a spam page gets good rank. This is the most important concern of the present web surfers. This can be solved by improving the present ranking system of the major search engines which can be achieved by implementing this Accurate Surfer Model.**

Keywords: **Time factor, Page Ranking System.**

I. INTRODUCTION

This paper explains about a surfer model which helps to improve the present web page ranking system and how to stop spam and how to give valid ranks to web pages which have the necessary and useful content. Before going into the details let us outlook the present searching system.

How a search engine is assigning ranks to the web pages on the net?

- Based on the calculations of no. of in-bound links and out-bound links – Google's technique
- HITS Algorithm – Developed by Jon Kleinberg
- Trust Rank – A link Analysis Technique – Described in a paper by Stanford University and Yahoo Researchers
- No. of Hits per page

What are the major problems we are facing while searching for any content on the net?

- We may get spam pages with our desired search pages.
- And the content which we will actually need may be appeared in the fourth or fifth page to which we may not go.

What are the reasons of getting spam mails or spam pages while we searching for any content on the net?

- The Spam pages are appearing on the search engine top results because of their higher page rank values.
- We are getting spam emails because the sender of them either may want to increase the page rank for his website or he may want to advertise the items which he wants to sell.

What's Wrong with Spam?

Most spam messages on the Internet today are advertisements from individuals and the occasional small business looking for a way to make a fast buck. Spam messages are usually sent out using sophisticated techniques designed to mask the messages' true senders and points of origin. And as for your email address, spammers use a variety of techniques to find it, such as "harvesting" it from web pages and downloading it from directories of email addresses operated by Internet service providers (ISPs).

But spamming today could well be undergoing a revolution. Over the past year, AT&T, Amazon.com, and OnSale.com all have experimented with bulk email. Although the companies clearly identify themselves in the mail messages, these bulk mailings can cause many of the same problems as spam messages from less scrupulous individuals and companies.

Spammers often say that spam isn't a problem. "Just hit Delete if you don't want to see it." And many spam messages carry the tagline "If you don't want to receive further mailings, reply and we'll remove you." But spam is a huge problem. In fact, junk email and junk postings are one of the most serious threats facing the Internet today.

Spam messages waste the Internet's two most precious resources: the bandwidth of long-distance communications links and the time of network administrators who keep the Internet working from day to day. Spam also wastes the time of countless computer users around the planet. Furthermore, in order to deliver their messages, the people who send spam mail are increasingly resorting to fraud and computer abuse.

How Much Spam Is There?

Just how much spam is out there? Although it's hard to come up with exact numbers, the initial reports from the field show that there's a lot and that the problem is getting worse:

- According to America Online, which testified about spam in front of the Federal Trade Commission in 1997, roughly a third of the email messages AOL receives on any given day from the Internet are unsolicited spam.
- According to the first academic study of spam, by Lorrie Faith Cranor at AT&T Labs-Research and Brian A. LaMacchia at Microsoft, between 5% and 15% of the email received by AT&T Research

and Bell Labs Research between April 1997 and October 1997 was spam.

- According to Spam Hippo, an automated Usenet anti-spam system written by Kachun Lee for PathLink Technology Corporation, roughly 575,000 articles were posted to Usenet in June 1998, of which roughly 200,000, or 35%, were spam. (That's down from a high of 60%, or 300,000 spam messages out of 500,000 postings, before Spam Hippo began operation.)

The Price Users Pay

It may take a spammer just five or ten minutes to program his computer to send a million messages over the course of a weekend. Now it's true that each of these messages can be deleted with just a click of the mouse, which takes only three or four seconds: a few seconds to determine that the message is in fact spam plus a second to click Delete. But those seconds add up quickly: one million people clicking Delete corresponds to roughly a month of wasted human activity. Or put another way, if you get six spam messages a day, you're wasting two hours each year deleting spam.

The price users pay for spam increases if you include the cost to the business or organization that operates the computer that holds your mailbox. These computers, called *mail servers*, require full-time connections to the Internet that can cost anywhere from $250 to $2,000 per month or more. The cost of the connection is determined, in part, by the amount of data it can carry. If a company's Internet connection is filled with spam, that company will be forced to spend more money on a faster Internet connection in order to handle the rest of its email traffic. Likewise, the company will be forced to buy faster computers and more disk drives. These costs must eventually be passed on to end users.

This scenario is not theoretical. In July 1997, spam mail overwhelmed AT&T WorldNet's outgoing mail system, delaying legitimate email by many hours.

Now before going to the actual solution for the above all problems let us observe the present Ranking system.

II. THE PRESENT PAGE RANK SYSTEM

Google describes Page Rank as:

Page Rank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives. It also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important".

In other words, a Page Rank results from a "ballot" among all the other pages on the World Wide Web about how important a page is. A hyperlink to a page counts as a vote of support. The Page Rank of a page is defined recursively and depends on the number and Page Rank metric of all pages that link to it ("incoming links"). A page that is linked to by many pages with high Page Rank receives a high rank itself. If there are no links to a web page there is no support for that page.

Before moving into the description we come across some frequent questions while talking about the search engine providers like GOOGLE, YAHOO and MSN etc. They are,

- How the page results are displayed in the search engine?
- What are the factors that affect the display of results?
- Does my results provided are actually based on its Rank?
- What are the different algorithms I can use for implementing the page rank for the web pages?
- Does my page rank technique handle my current Internet Traffic?

The answers for all the above queries can be found by the end of the below description on Page Ranking System.

Google assigns a numeric weighting from 0-10 for each webpage on the Internet. This Page Rank denotes a site’s importance in the eyes of Google. The Page Rank is derived from a theoretical probability value on a logarithmic scale like the Richter scale. The Page Rank of a particular page is roughly based upon the quantity of inbound links as well as the Page Rank of the pages providing the links. Let us see them in detail.

Simplified Page Ranking Algorithm

Page Rank is a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. Page Rank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided between all documents in the collection at the beginning of the computational process. The Page Rank computations require several passes, called "iterations", through the collection to adjust approximate Page Rank values to more closely reflect the theoretical true value.

A probability is expressed as a numeric value between 0 and 1. A 0.5 probability is commonly expressed as a "50% chance" of something happening. Hence, a Page Rank of 0.5 means there is a 50% chance that a person clicking on a random link will be directed to the document with the 0.5 Page Rank.

How Page Rank Works

Assume a small universe of four web pages: A, B, C and D. The initial approximation of Page Rank would be evenly divided between these four documents. Hence, each document would begin with an estimated Page Rank of 0.25.

In the original form of Page Rank initial values were simply 1. This meant that the sum of all pages was the total number of pages on the web. Later versions of Page Rank (see the below formulas) would assume a probability distribution between 0 and 1. Here a simple probability distribution will be used- hence the initial value of 0.25.

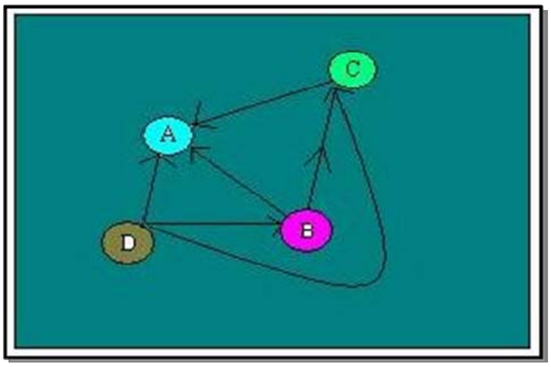

Fig. 1 Figure corresponds to the links in the WEB

If pages B, C, and D each only link to A, they would each confer 0.25 Page Rank to A. All Page Rank PR( ) in this simplistic system would thus gather to A because all links would be pointing to A.
This is 0.75.That is

PR (A) = PR(B)+PR(C)+PR(D)

Again, suppose page B also has a link to page C, and page D has links to all three pages. See the Figure 1 for correspondence. The value of the link-votes is divided among all the outbound links on a page. Thus, page B gives a vote worth 0.125 to page A and a vote worth 0.125 to page C. Only one third of D's Page Rank is counted for A's Page Rank (approximately 0.083).

PR(A) =PR(B)/2+PR(C)/1+PR(C)/3

In other words, the Page Rank conferred by an outbound link is equal to the document's own Page Rank score divided by the normalized number of outbound links L( ) (it is assumed that links to specific URLs only count once per document).

In the general case, the Page Rank value for any page u can be expressed as:

PR(u) = ∑PR(u)/ L(u)

Now we will see the new addition to the above said ranking system to make it more accurate and reliable. I name it as “An Accurate Surfer Model”.

III. An Accurate Surfer Model

Now we will see how this model increases the reliability of any ranking system.

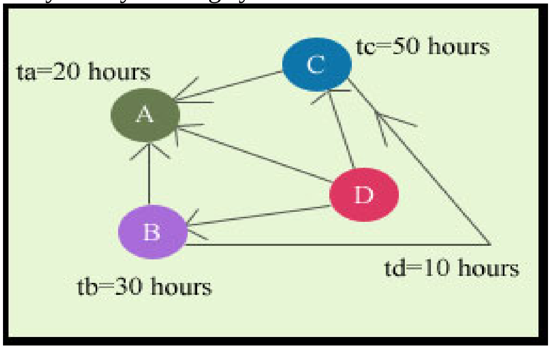

Fig. 2 Corresponds to the links in the WEB with average spent time

If you observe the above figure it is same figure which we earlier used to calculate the page ranks for the pages but here the new addition to this figure is the average time spent by all users on each page. This is the extra parameter which will increase the reliability on assigning the correct ranks for the deserved pages.

What is the extra Parameter?

Ans. Time factor

In what way it is useful?

Case i: Let us take an imaginary surfer surfing the net for some content and if he saw a page with some valid data then he waits and spend some amount of time, the amount of time spent by him will be dependent on the importance of the content. If the content is more useful then the surfer will spend more time on it.

Case ii: If he sees any spam page then he immediately closes the page.

But in the above two cases the pages visited by the surfer are getting the same ranks because the ranks are given on the basis of number of hits. The surfer has hit the two pages one time. So despite of having unnecessary information the spam page is also getting a rank same as that of the page which has useful content.

This has to be avoided which can be done if we consider the time spent by a surfer on a web page as one parameter to page rank and adding that to the rank evaluation system. So the pages which have the valid data will get the appropriate rank.

And another most important advantage is that the spam pages are getting less importance and less rank compared to the original pages. Then the amount of generation of spam pages in the search results will drastically decrease .

Introducing this term in the present ranking system then the new formula to calculate Page ‘A’ rank will be

PR(A) =PR(B)/2*tb+PR(C)/1*tc+PR(D)/3*td

tb - Average Time spent on Page ‘B’ by all surfers
tc - Average Time spent on Page ‘C’ by all surfers
td – Average Time spent on Page ‘D’ by all surfers

The new proposed formula to calculate page rank for any web page can be specified as,

PR(u)= ∑PR(u)*t / L(u)

Here‘t’ is the average time spent by all surfers on that particular page.

IV. Advantages

- The pages which have the original and useful content for the net surfers will get good ranks.
- By default Spam pages will get less rank when compared to the original and useful pages.
- Generation of Spam pages will be drastically decreased in the search results
- The amount of time will be decreased in searching for useful content by the net surfers because of lack of spam pages.
- The Efficiency and Reliability of the Search Engines will be increased

V. CONCLUSION

By implementing the above said "Accurate Surfer Model", most of the problems of the search engines will be reduced and the amount of time spent by the surfers on the internet for information will be reduced to a large extent, the pages which have the original and useful content will only get good rank which will increase the efficiency and reliability to the present day search engines because they don't display the false or spam pages in its search results. By using this we will not have any 1000 pages search result but a unique 4-8 pages only.

REFERENCES

[1]. http://en.wikipedia.org/wiki/Page_ranking
[2]. http://oreilly.com/catalog/spam/chapter/ch1.html