Web Mining using Page/Server Crash and Broken Links Ranking System (PCBRS) Algorithm

Pawan S. Hora[1], Nishant A. Jain[2], Aniket A. Gangurde[3], Sonal Balpande[4]
KC College of Engineering & Management Studies and Research,
Thane (E), India

ABSTRACT

In today's world not only big organizations but also a common man needs a lot of information or rather, big chunks of information for his/her work according to their area of interest from the web. So, in short web mining can be defined as a process in which the data is being extracted, filtered and given back to the end user, in order to satisfy his needs for the information that he need, in order to complete his work. In simple terms, it can be compared to the process of practical world mining, wherein, there is an extraction of a particular "RESOURCE" from a particular place. So hence, the web mining can be thought of as a simple DATA as a RESOURCE being extracted from the web, which has a huge reserve's of it. This paper also suggests some of the improvements in the world renowned search engine GOOGLE and also adds some of the features which can prove instrumental in the world of search engines and thus information search.

General Terms

Search Engines, Google, Algorithm, Page rank, Web Mining, Information Search.

Keywords

Search Engines, Page Rank, Page Crash, Broken Links, Dead Links, Ranking algorithm.

1. INTRODUCTION

Nowadays, the internet can be thought of to be as a complete "University" in its own, because a completely whole new package of information, relevant to a particular search term/topic is being found on the web nowadays, which is also being updated periodically which helps the users from different disciplines, in order to get an access to the updated information, which is mainly due to the DYNAMIC nature of the web .The rise of the search engines have brought about a totally new revolution in the field of Web Mining and also has opened the gates for a whole new world of information extraction, information exchange, updating of the information, semantic web mining etc.

2. LITERATURE SURVEY

2.1 Search Engine

Search engines can be thought of to be as a "KEY-TERM FINDER" or "QUERY RESOLVER" for a particular user, belonging to any field or referring to any context whatsoever. The first of all the search engines was "Archie", which was being created in the year 1990 by Alan Emtage [1]. In 1990, there was no World Wide Web. Nonetheless, there was still an Internet, and many files were being scattered all over the vast network. The primary method of sorting and retrieving the files was via the FTP. Archie had changed all this. It combined a script-based text data gatherer which fetched the site listings of anonymous FTP files, with a regular expression matcher for retrieving the file names matching a user query. As seen below, the Fig.1 shows the working of a typical search engine.

One of the world's first "full text" crawler based search engines was Web Crawler, which came out in the year 1994. Unlike its predecessors, it let the users search for any word in any web page, which has become a standard for all the major search engines since. Nowadays, the search engines not only provide the users with just a normal keyword search, but they also try to provide the users with a filtered search relevant to their search topic or their search term. They are one of the core parts of the Web Mining. The other special feature about the modern search engines is that they not only provides its

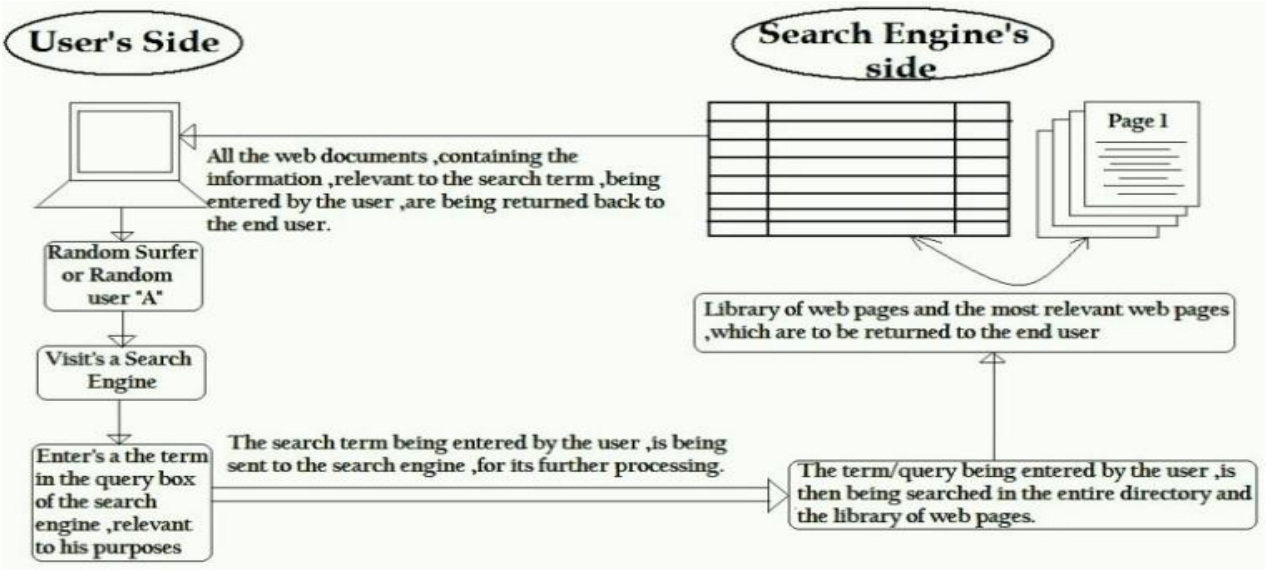

Fig 1: Typical working of a Search Engine

users in order to find their bit of information in a simple format but they also provide the users with the choice to find the information related to their sections of interest.

For Example:-GOOGLE has options for the user to find the information, relevant to their search term, in their sections such as Images, Books, Web, Latest News etc which are present in the search engine itself. And not only this, Google also allows the users other services, such as you tube, Gmail, Orkut social networking site, Language Translators etc.

2.2 Working of Google

The Google search engine, is today one of the world's greatest search engine's which faces about billions of searches, in a single day. Now basically what Google does, or how it works, is that it first takes a search query or the search term (which is to be entered by the user who wants to find the information relevant to that term, in the search engines query box), then it takes that term as its "Index" and then it searches its entire web directory and all the web documents, in order to find the information, relevant to that search term and then, it returns the links and web documents, which contain that search term and also the other pages, which are being linked to those pages also being known as the "Suggestions" to the end user. Thus, in this manner a random surfer can visit any webpage and search any information which he/she wants to extract from the web. This is how the Google search engine works. Another very interesting concept, about the search engines is the concept of Page Rank of a webpage.

2.3 Page Rank Algorithm

Page Rank of a particular webpage mainly refers to its position or standing in the list of suggestions being produced by a web search engine relevant to a particular topic or a search term [2].

If a back link comes from an "important" page, then that back link is given a higher weighting than those back links comes from non-important pages. It a simple way, link from one page to another page may be considered as a vote [3][4][5]. However, not only the number of web pages receives is considered as important, but the "importance" or the "relevance" of the ones that cast these votes as well.

Now the main use of the page rank for the search engines is that it tries to provide the "Best possible bits" of information which is most closely related to their search term. Nowadays, most of the search engines typically use a particular algorithm to calculate the page rank of a particular web page. There are many algorithms which calculate the rank of a web page based on the various factors, such as the Time Rank which means that ranking the web page based on the amount of time being spent by a particular user on that web page, means that more the time spent by the user, the greater is the rank of that web page. Another factor is the URL Tag Rank which suggests that if the length of the URL of a particular web page is bigger the rank is bigger of that web page and also on many other factors such as the length of the web document the Euclidian's mathematical formula for calculating the page rank etc.

2.4 Google's Page Rank Algorithm

The page rank algorithm being used by the Google search engine was basically being created by Larry Page and Sergey Brin in the year 1996. This algorithm was based on a simple mathematical web graph which suggests that if there are a larger number of links pointed to a particular webpage, then its page rank is high. The mathematical web graph can be represented by the following Fig.2 as shown below:

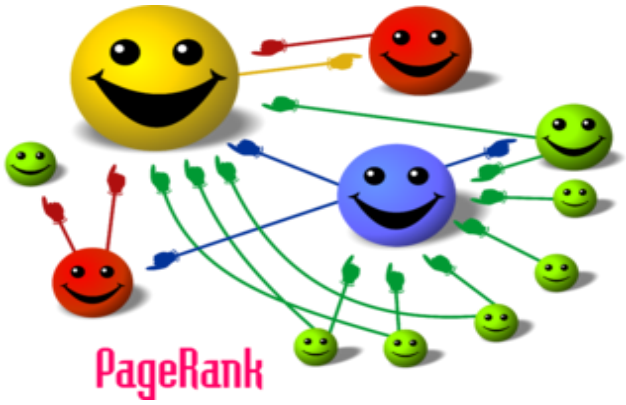

Fig 2:- Web graph indicating the working of Google's page rank algorithm

The Page rank algorithm being used by Google, can be given in the following manner:

$$PR(A) = (1-d) + d(\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)})$$

Where,

- PR(A) is the Page Rank of page A,
- PR(Ti) is the Page Rank of pages Ti which link to page A,
- C(Ti) is the number of outbound links on page Ti and
- d is the damping factor which can be set between 0 and 1.

The basic concepts about this algorithm are as follows:

2.4.1 In-Links:

The links which are pointing inwards, that is towards a particular web page from an external place, is being called as the in links of that web page. Or we can also say that an external web page is pointing towards a particular webpage.

2.4.2 Out-Links:

The links to which your web page is pointing towards are being known as the out - links of that particular webpage; they are also being known as the external links. In this case, it can also be said that a particular web page is pointing to an external web page.

2.4.3 Damping factor:

The Damping factor is any random value ranging between 0 and 1. The Damping factor is being mainly chosen because of the dead link issues of a web page.

The working of the Google's page rank algorithm can be shown by the following example:

PR(A) = 0.5 + 0.5 PR(C)
PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

These equations can easily be solved. We get the following Page Rank values for the single pages:

PR(A) = 14/13 = 1.07692308
PR(B) = 10/13 = 0.76923077
PR(C) = 15/13 = 1.15384615

In the Google's page rank algorithm there is a phenomena, that is if there is web page, with a high page rank, which is pointing towards a web page which has a lower page rank, then that page automatically gets a high page rank value.

3. PROPOSED SYSTEM

Improvements of the proposed search engine in comparison to the Google search engine:-

a) Displays a screenshot of the links suggestion along with the suggestions page after user makes a search on the search engine. This saves the time of the users to visit that link as he can already view the contents of that web page on the list of the suggestions, relevant to his search query.
b) Takes up and displays the backup files, for the data which is being updated for a particular web site, to the users of the search engine. This means that if a particular web site restructures its links or the updating of some content on the website then this search engine would also be providing an access to the files and the data of the website, which was present on that website, before there had been any data updating on that website.
c) The search engine provides its user with the facility of the file based or paragraph based search engine, wherein the users can enter a big query, store it in a file (for the purpose of future reuse of the file, if any), and search for this query on the web, which avoids the restriction of only a specific keyword search, thus allowing to search for a larger and a combined query, made up of 3-4 smaller queries.
d) The search engine would also provide with the context based searches, that is if a user searches for a search term such as "HEART" then this term would have different context's for different users according to their needs such as some user would mean "HEART" as a romantic term or "HEART" as a body part etc. So this graph based search would provide a more filtered search.
e) "Popularity meter" and the concept of "Branded Websites" is another step forward in order to provide additional help to the end users with a more accurate information from the billions of web pages in which he is searching his/her relevant information. Basically this means that the users can rank the web pages in a custom manner, that can be handy during crucial times, in a manner that he can view all his top suggestions related to a search term, and irrelevant of the search engines ranking system, he can get his favourite websites above the other websites, by ranking them in a custom manner.
f) The search engine would also provide an additional feature of Image Based Searches that is, input as an image and its output suggestions as a text based suggestions (including snippets). This is an image based search engine.

As shown in the Fig.3, these are the major factors, which must be considered during the ranking of a particular website or a web page. This is because these are the factors, which can prove to be a game changer, and can give an edge above, to those websites, which are well designed, which do not face problems like broken links, and server crashes, but still their page rank is low. These factors decide that how closely, the user of a search engine can reach his/her information, and in a minimal amount of time.

The Page Crash Ranking System Algorithm mainly takes into account 4 factors in order to calculate the page rank of a particular web page and those are:

- Number of Page Crashes or Server Crashes of a particular web page.
- Number of Broken links or Dead links present in the web page.
- The Total number of links which are present in a particular web page that is both the in links and the out bound links.
- The Balancing factor which is a constant value which is 0.1.

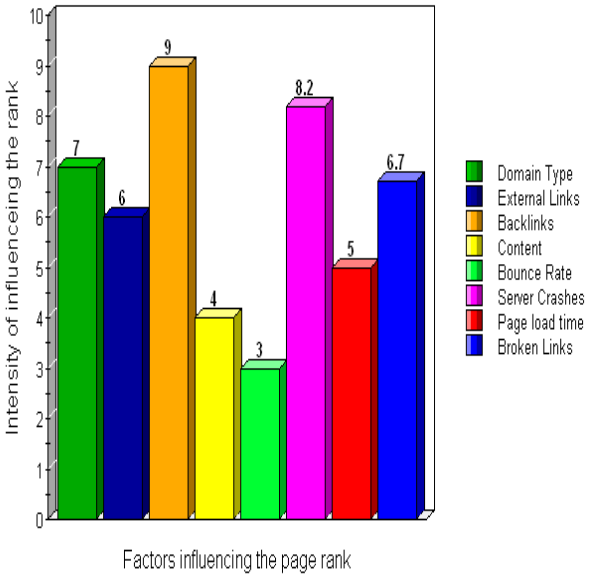

Fig 3:-Factors which can affect the Page rank of a particular web page.

The Page Crash and Broken Links (PCBRS) Ranking System Algorithm can be given in the following manner:

$$PNR(N1) = \left[\frac{p(N1)}{n(N1) + r(N1)\text{^}2}\right] * B$$

Where,

- PNR (N1) = Overall page rank calculated of the web page.
- (N1) = Page number 1.

- P (N1) = Total number of links of a web page.
- N (N1) = Total number of the web page crashes.
- R (N1) = Total number of Broken Links and dead links of a particular web page.
- B = The Balancing Factor(usually the value is taken to be as 0.1)

Some of the concepts which are present in this algorithm can be explained in the following manner:

1) Broken Links: The broken links can be mainly referred to as the links, which are not directly associated or linked to the other web page. Let us take an example of 3 web pages A, B, C; so now among these 3 web pages, there are bidirectional links between Page A and Page b, and also a bidirectional links exists between the Page A and Page C, but there is no links, which are present in between the Page B and Page C; so hence, this can be said to be as a broken link between Page B and Page C. There are no specific reasons, for why the broken links occur, but mostly they are an outcome of the changing page rank values. The Fig.4 depicts a visual example and a sample, in order to better explain the concept of the broken links, how it takes place, and also the problems which the users encounter when they face a broken link in a website or two. The broken links are a major problem, being observed in many of today's websites and thus must be taken into account for, during the ranking of the web pages. The given below example, indicates the existence of a broken link between 3 web pages, Page A, Page B, and Page C. These all web pages, can b a part of a single website, or of 2 or 3 different websites.

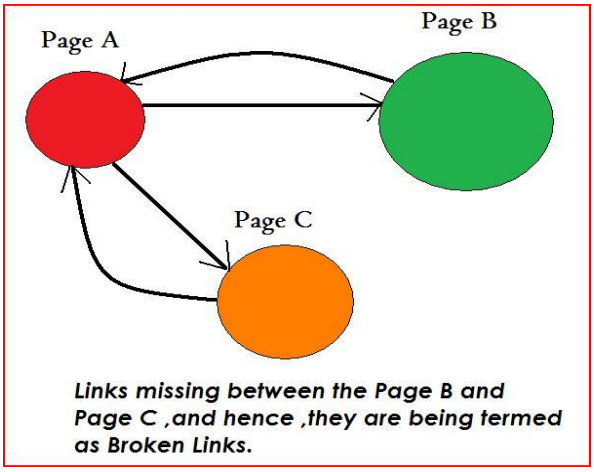

Fig 4:- A typical example of a Broken Link

2) Dead Links: The Dead Links are basically those web pages which do not reach or which do not link to any other web pages. Dead links can also be referred to as the last links.

3) Server Crashes: The Server Crash can be defined as the phenomena in which there is a failure of server of a particular web site or a web page from where it has been supplied so in such case the users cannot get an access to that website temporarily, permanently or for some period of time, depending on the reasons for the server crash, which can be various such as if the web page has been shifted to a new server, or if there are some technical problems at the server side, or if there are a lot of users who are trying to access the same website at the same instant of time, thus Jamming that particular web site, resulting in the crashing of that web page.

4) Balancing Factor: The balancing factor as the name suggests mainly does the work of balancing the very high page rank values and thus it decreases the value of the page rank and also ensures that the value being calculated by the algorithm, of a particular web page is feasible. The value of the Balancing factor is 0.1.

The Page Crash and Broken Links Ranking System Algorithm has a very low time complexity that is O $(1/n^2)$. Thus it is a very efficient algorithm. This algorithm is being created such that it is being mainly used in order to reduce the page rank of a web page, if it has a larger number of the negative factors such as the page crashes and also the number of broken links and dead links, so in our algorithm, these are the factors which affect the page rank of a particular web page. The PCRS algorithm can be demonstrated with the help of the following example: Suppose if there are 30 links in total, and if the number of times, the page has crashed is 1, and the total number of broken links in the webpage are 4 then the page rank of this web page by the Page Crash and Broken Links (PCBRS) Ranking System algorithm can be given in the following manner:

PNR (1) = $30/1 + (4^2)* 0.1$

$=30/1+16 * 0.1$

$=30/17 *0.1 = 0.176$

As we can see that the number of links are, 4 in total so they gave a very minor rank of 0.176 to our webpage. Similarly we can try other examples also for the demonstration of this algorithm.

There is a very simple methodology about this algorithm is that this algorithm basically works in 2 passes that is, in the first pass it would find out the total number of links which are present in a web page and in the next pass, it would find the total number of broken links and dead links and also if (there are any of the server crashes have occurred on that page) and then calculate the final page rank according to the parameters as mentioned above for that particular web page.

4. CONCLUSION

With the increasing number of links, and also with the increasing number of web pages in the World Wide Web there are also a number of users increasing day by day, needing more amount of updated information on a regular basis, so there is a need to provide a "Quality Information" to the end users and also the information which is very much closely relevant to a particular topic. The Google's page rank algorithm is not enough to detect and remove, or guide the users for the purpose that if there are various broken links in a particular webpage, then the page rank is less of that web page and also that they would not find the information of their interest in such websites. The Spam Links and the Broken Links of the web page are thus the negative factors of a web page which reduces the page rank of a web page, and hence the same phenomena can be depicted by the Page Crash Ranking and Broken Links System Algorithm (PCBRS). This can prove to be as a very instrumental algorithm for the

filtering of the web pages, which are to be sent to the end user, which contains the information related to their search term or query.

5. REFERENCES

[1] P.T Joseph and S.J, E-Commerce-An Indian perspective, New Delhi, India :PHI Publications, 2012

[2] L. Page, S. Brin, R. Motwani, and T.Winograd, "The Page rank Citation Ranking: Bringing order to the Web". Technical report, Stanford Digital Libraries SIDL-WP-1990-0120, 1999.

[3] L.Page, S.Brin, "The Anatomy of a large scale Hypertextual Web search Engine", Computer Networks and ISDN Systems, Vol 30 ,issue 1-7, pp. 107-117 ,1998.

[4] R. Kumar, P. Raghavan, S. Rajagopalan,D.Sivakumar, A. Tompkins and E. Upfal, "Web as a Graph", Proceedings of the Nineteenth ACM SIGMODSIGACT-SIGART symposium on Database Systems, 2000.

[5]R. Cooley B. Mobasher and J. Srivastava, "Web Mining :Information and Pattern Discovery on the World Wide Web" . "Proceedings of the 9th IEEE International Conference on tools with artificial intelligence, pp. (ICTA197), 1997.

[6] http://computer.howstuffworks.com/internet/basics/search-engine.htm

[7] http://www.youtube.com/watch?v=BNHR6IQJGZs