Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# An Approach for Fetching User Relevant Pages Using Backlinks: A Review

**Sanjeev Dhawan[1], Vinod[2]**
[1]Assistant Professor, U.I.E.T. KurukshetraUniversity, Kurukshetra, Haryana, INDIA
[2]Research Scholar, U.I,E.T, Kurukshetra University Kuruksheta, Haryana, INDIA

***Abstract: A search engine to determine how many backlinks a web page has though is not as accurate as building in a modification that can assist in counting the number of backlinks to the particular webpage or website. A web page that has more backlinks than another with similar content will rank higher than the other page, simply because it seems to be more popular with visitors and other websites. Unfortunately, this means that many websites have engaged in paid linking, which boosts their backlink numbers. This has caused search engines to add in specifications to the algorithms used to determine backlinks that now research whether the backlinks have been paid for, or are genuine. Only genuine backlinks help web pages rank well on most search engines.***

***Keywords: Internet, Backlinks, Web page, Rank, Search engine***

## I. Introduction

Internet is "network of networks" that transmits data by packet switching using the standard TCP/IP protocol suite. It consists of millions of smaller domestic, academic, business, and government network, which together carry information and services, such as electronic mail, file transfer, news, and the interlinks web pages and other resources of the World Wide Web(WWW). The World Wide Web is a huge situate of interlinked images, documents and other resources, linked by hyperlinks and URLs. The documents are formatted in a markup language called HTML (Hyper Text Markup Language). [1]. According to using HTTP (Hypertext Transfer Protocol) store originals, and cached copies of, these resources to distribute, allow the web servers and other machines. The communication protocols used on the Internet such that HTTP is Purely. In normal use, web browsers, such as Internet Explorer, Fire fox, Opera, and Google Chrome, access web pages and allow users to navigate from one to another via hyperlinks. Web documents may contain almost any combination of computer data including graphics, sounds, text, video, multimedia and interactive content including games, office applications and scientific demonstrations. World Wide Web has information and data Compared to encyclopedias and traditional libraries. Search engines are useful for finding information on the World Wide Web, such as Google, Yahoo! and AltaVista. However, these general-purpose search engines are subject to low accuracy and/or low reporting. Manually-generated directories such as Yahoo! provide high-quality references, but cannot keep up with the Web's explosive growth. Although crawler-based search engines, like AltaVista, cover a larger fraction of the Web, their automatic indexing mechanisms often cause search results to be imprecise[2][3]. This paper presents world wide web, search engine and its architecture. The rest of the paper is organized as follows. Section 2 discusses the Web crawler and its classification & various modes. Section 3 presents the backlinks and its different types & working and finally Section 4 presents our conclusions.

**WORLD WIDE WEB (WWW)**

Most people use *Internet* and *World Wide Web* interchangeably, but fact in the two terms are not synonymous. The World Wide Web is a huge situate of interlinked images, documents and other resources, linked by hyperlinks and URLs. The documents are formatted in a markup language called HTML (Hyper Text Markup Language). According to using HTTP (Hypertext Transfer Protocol) store originals, and cached copies of, these resources to distribute, allow the web servers and other machines. Web documents may contain almost any combination of computer data including graphics, sounds, text, video, multimedia and interactive content including games, office applications and scientific demonstrations[1]. World Wide Web has information and data Compared to encyclopedias and traditional libraries. The world wide web was created in 1989 by Sir Tim Berners Lee, Working at the European Organization for Nuclear Research (CERN) in Geneva, Switerland and released in 1992. Since then, Berners-Lee has played an active role in guiding the development of Web standards (such as the markup languages in which Web pages are composed) [2].

***SEARCH ENGINE:*** Search engines are useful for finding information on the World Wide Web, such as Google, Yahoo! and AltaVista. However, these general-purpose search engines are subject to low accuracy and/or low reporting. It is thus difficult for a single search engine to offer both high reporting and high precision. This problem is exacerbated by the growth in Web size and by the increasing number of naive users of the Web who typically issues short (often, single word) queries to search

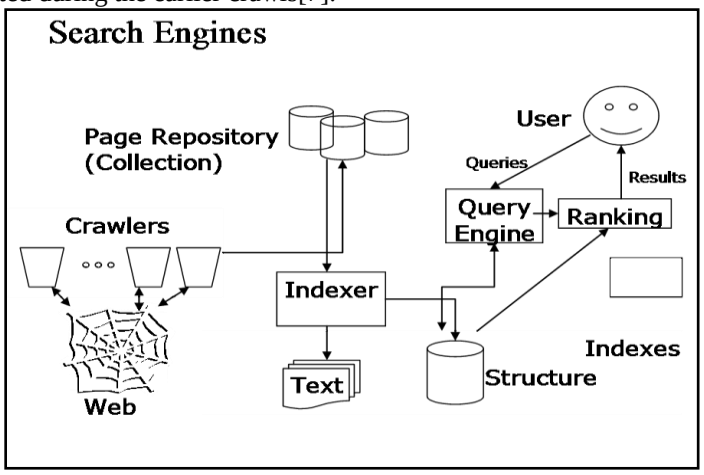
engines. Topic-specific search engines often return higher-quality references than broad, general-purpose search engines for several reasons[5]. First, specialized engines are often a front-end to a database of authoritative information that search engine spiders, which index the Web's HTML pages, cannot access. Second, specialized search engines often reflect the efforts of organizations, communities, or individual fanatics that are committed to providing and updating high quality information. Third, because of their narrow focus and smaller size, word-sense ambiguities and other linguistic barrier to high-precision search are ameliorated[6].

Search engine presents a novel query routing method, called *topic-centric* query routing, which compensates for lack of unfettered access to search engine databases by using two key techniques:

***Neighborhood-based topic identification***: a technique for collecting the abstract topic terms relevant to a search engine from existing Web documents.

***Query expansion***: a technique for obtaining the terms relevant to a query. For the purpose of topic-centric query routing, it is used mainly for evaluating the relevance of a query to the identified topic terms of search engines.

***General Architecture of a Search Engine:*** Search engines have typically used exhaustive crawler to build and update large collections of documents. The cost of crawling and indexing the collections is amortized over millions of queries received by the search engines. The General Architecture of a search engine is shown in Figure. This includes the HTML interface where the users submit their queries and the mechanism for serving these queries. The database stores all crawled data from the web crawlers. The crawling system is the subsystem responsible for maintaining the search engine database and incorporating the changes from the Web. Once the search engine, has been through at least one complete crawling cycle, it may be formed by several indexes that were created during the earlier crawls[7].




**Search Engine General Architecture [7]**

## II. Web Crwler

A crawler is a program that downloads and stores Web pages, often for a Web search engine. Roughly, a crawler starts off by placing an initial set of URLs, *S*0, in a queue, where all URLs to be retrieved are kept and prioritized. From this queue, the crawler gets a URL, downloads the page, extracts any URLs in the downloaded page, and puts the new URLs in the queue. This process is repeated until the crawler decides to stop until repeated by that process. For other applications used to collected pages are later, such as a Web search engine. Crawlers are typically programmed to visit sites that have been submitted by their owners as new or updated[9].

***CRAWLER TYPES:*** There are main three type of crawler. Each is targeted at different types of crawling tasks[10]:

***Standard crawler:*** Crawler with one thread that uses 'depth first' searching when searching a repository. This crawler type is aimed at searching repositories with moderate CPU usage.

***Web crawler:*** Crawler with one thread that uses 'breadth first' searching when searching a repository. This crawler type is aimed at searching Web repositories with moderate CPU usage.

***Crawler with several threads:*** Crawler with more than one thread that uses 'breadth first' searching when searching a repository. This crawler type is aimed at searching repositories with high CPU usage.

***GENERAL ARCHITECTURE OF CRAWLER:***

There are two configurations of crawling architectures with dynamic assignments.

A small crawler configuration, in which there is a central DNS resolver and central queues per Web site, and distributed downloader's[11]:

- A large crawler configuration, in which the DNS resolver and the queues are also distributed. With this type of policy, there is a fixed rule stated from the beginning of the crawl that defines how to assign new URLs to the crawlers.

Crawlers are software programs that automatically pass through the web, retrieving pages to build a searchable index of their content. Conventional crawlers receive as input a set of "start" pages and recursively obtain new ones by locating and cross their outbound links.

***Parallel Crawler:*** A parallel crawler is a crawler which is used when the size of the Web grows. It becomes more difficult to retrieve the whole or a significant portion of the web using single process. Therefore, many search engines often run multiple processes in parallel to perform the above task, so the download rate is maximised this type of crawler is referred as a parallel crawler, which works parallel. When multiple processes run in parallel to download pages, it is possible that different processes download the same page multiple times .One process may not be aware that another process has already downloaded the page. Clearly, such multiple downloads should be minimized to save network bandwidth and increase the crawler's effectiveness.

***Distributed crawler:*** Distributed web crawling is a distributed computing technique whereby Internet search engines employ many computers to index the Internet via web crawling. The idea is to spread out the required resources of computation and bandwidth to many computers and networks.

## III. User Relevant Pages Usin Backlinks

**Backlinks**, which are sometimes called inbound links, are incoming links to a web page or the entire website. Search engines have measure the number of backlinks a website or web page has, and ranks those web pages with more backlinks in a higher position as inbound links are important to search engine rankings[12].

A web page that has more backlinks than another with similar content will rank higher than the other page, simply because it seems to be more popular with visitors and other websites. Many search engine use backlinks to determine page rank too. This means that many websites have engaged in paid linking, which boosts their backlink numbers. This has caused search engines to add in specifications to use to determine backlinks that now research whether the backlinks have been paid for, or are real. Only genuine backlinks help web pages rank well on most search engines. This action has caused many websites to lose their page rankings, but has also allowed other pages with genuine backlinks to rise in the search engine results with major search engines such as Google. In search engine terminology a backlink is a hyperlink that links from a Web page, back to your own Web page or Web site. Also called an ***Inbound Link*** (IBL) these links are important in determining the popularity (or importance) of your Web site. Some search engines, including Google will consider Web sites with more backlinks more relevant in search results pages[12][13].

***TYPES OF BACKLINKS***

There are many types of backlinks which is given below[14]:

***Links from scraper sites:*** Before submitting your website to a directory listing, be sure the site is a legitimate directory. Scraper_directories_steal traffic rather than driving more and typically use frames, no-follow tags, or simply omit the backlinks intended for your site. Gaining links from these websites will benefit their rankings rather than your own.

***Backlinks from link farms, link exchanges, and similar groups:*** In Google's eyes, any types of link schemes are not approved and can seriously decrease your rankings. This includes buying, swapping, and giving away links simply to generate links. There are some reputable programs that have developed acceptable methods but traditional link farms and backlink exchanges are noticed by search spiders.

***Links generated by scripts and software:*** There are a variety of programs available online that promise to increase backlinks to your website. Typically this is done with a software script that can be set to run automatically, leaving comments on forums and blogs that include your links. These scripts are annoying to other webmasters, are often ineffective because the comments are blocked by spam catchers, and can get your website blacklisted by search engines.

***Backlinks from "bad" sites:*** Just as backlinks from respected, high ranking websites can provide great benefits to your own, backlinks from blacklisted websites or those deemed unsuitable by search engines can damage your standings. Avoid incoming links from gambling sites, adult entertainment, and other questionable niche communities.

***Website directories designed to increase rankings:*** Directories are a great way to drive traffic to your website, but not all are effective link building tools. If you're hoping to gain a solid backlink from a directory listing, look for the following characteristics: The directory page is included in the website's navigational links and/or sitemap.

- The page is not using the no follow or no index attribute.
- The directory should be a static database rather than dynamic so it is easily read by search robots.
- Organized links arranged in categories or grouped by similar topics. A mish-mash of URLs thrown on a page generally signifies spam to the search engines.

Remember that website directories are not bad and can increase traffic to your site - but most will not generate quality backlinks.

***Reciprocal links:*** Once upon a time in internet land, everyone traded links. Blogrolls and link lists are still quite common in the blogosphere but "trading links" is not an acceptable way to build links. Google actually states that "Excessive reciprocal links or excessive link exchanging ("Link to me and I'll link to you.")" is in violation of their webmaster agreement.

***USER RELEVANCE PAGES***

All search engines determine what will be returned in the SERPs (Search Engine Results Pages) by using two measures authority and relevance. Relevance means the page contains the keywords and authority means the page has back links to it from other web pages.[13]. The search engines determine the order in which web pages are indexed on the results pages by the number of back links to the page and their respective authority. Back links play two important roles they direct traffic to your web site and factor in the search engines deciding the position of your web pages in the index of results.[14]. Browsers who discover and click on back links containing keywords associated with their interests will be directed to your web pages. This text in the back link is known as 'anchor text' and this too is taken into consideration by the search engines. Whilst all back links are of value some are more valuable than others. The search engines give authority to web pages which can be passed on to your web pages through the back links. The higher the authority of the web page 'sending' the back link the more authority is likely to received by your web pages[15].

## IV.Conclusion

The backlink search enginge results are the web pages that point to the compact URL. The search results will represents the firm's Web communities. The backlink search results will be fetched in real-time to the local computer will examine the fetched web pages and perform text analysis to extract the important phrases from the stored Web pages**.** This paper describe the search engine architecture and web crwalers. Baclinks and its various types are also discussed.

**References**


[1]. Tim Berners Lee, *"The World Wide Web: Past, Presnet and Future"* Draft response to invitation to publish in IEEE Computer special issue of October: August 1996 Cambridge MA 02139 U.S.A.

[2]. Sergey Brin, *"Extracting Patterns and Relation From The World Wide Web"* Springer-Verlag London, UK 1999 ISBN: 3-540-65890-4.

[3] C. Aggarwal, F. Al-Garawi, and P. S. Yu, *" Intelligent Crawling on the World Wide Web with arbitrary predicates"* In *WWW10*, Hong Kong, May 2001.

[4]. B.D. Davison, "*Topical Locality In The Web"* In Proc. 23rd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2000.

[5]. Sergey Brin, Lewrence Page "*The Anatomy of a Large-Scale Hypertextual Web Search Engine"* Volume:30, Issue:17, Publisher: Elsevier, Pages:107-117: Computer Network and ISDN System .ISBN:9781424453658.

[6]. G. Pant, S. Bradshaw, and F. Menczer, *"Search Engine-Crawler Symbiosis: Adapt-ing to Community Interests"* In Proc. 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2003), Trondheim, Norway, 2003.

[7] Sergey Brin and Lawrence Page, "*The Anatomy Of a Large-Scale Hyper Textual Web Search Engine"*. Proceedings of the Seventh International World Wide Web Conference, pages 107—117, April 1998.

[8] L. Page and S. Brin, "*The Anatomy of a Search Engine*", Proc. of the 7th International WWW Conference (WWW 98), Brisbane, Australia, April 14-18, 1998.

[9]. J. Johnson, T. Tsioutsiouliklis, and C.L. Giles, "*Evolving Strategies For Focused Web Crawling"* In Proc. 12th Intl. Conf. on Machine Learning (ICML-2003), Washington DC, 2003.

[10]. F. Menczer, G. Pant, M. Ruiz, and P. Srinivasan, *"Evaluating Topic-Driven Web Crawlers"* In Proc. 24th Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2001.

[11]. F. Menczer, G. Pant, and P. Srinivasan, *"Topical Web Crawlers: Evaluating Adaptive Algorithms" To appear in ACM Trans. on Internet Technologies*, 2003. http://dollar.biz.uiowa.edu/~¯l/Papers/TOIT.pdf.

[12] J. Cho, H. Garcia-Molina, and L. Page, *"Efficient Crawling Through URL Ordering"* in Proceedings of the Seventh World-Wide Web Conference, 1998.

[13] Qu Cheng, Wang Beizhan, Wei Pianpian, "*Efficient Focused Crawling Strategy Using Combination of Link Structure and Content Similarity"* IEEE 2008.

[14] F. Yuan, C. Yin and Y. Zhang, *"An Application of Improved PageRank in focused Crawler"* IEEE 2007.

[15] J. Rennie and A. McCallum, *"Using Reinforcement Learning To Spider The Web Efficiently"* in Proc. International Conference on Machine Learning (lCML),1999.