Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Blog Search Engine: Architecture, Types, Benefits and Limitation: A Brief Review

Romita
M.tech (CSE) Scholar
India

Vikas Chaubey
Asst. prof. in MVN University, Palwal
India

Abstract: ***This paper shows the basic conceptual working of the extraction process used by the blog search engine with its architecture and also put some highlights on the limitations or problems that blog search engines are faces. This paper also showssome of the benefitsof using blog search engine are also explained. Also some points are mentioned which shows that on which area of the blog search engine the more work will required to be done.***

Keyword: BSE, web log, blog search engine architecture, Blog, information retrieval.

I. INTRODUCTION

In today's scenario minute knowledge related to every topic or issue in this world we are getting from the internet. The Internet helps us to communicate globally. It is only the hardware and software infrastructure which provides connectivity between computers. Internet helps us to provide the solution or a hint to one's query. The new technologies are discovering day by day to provide with the fast and better results to user.

With the help of World Wide Web internet has developed a strong bond between the human being and technology. From the past decade the people are engaging more towards in publishing the journals [9]. Now the question arises that what is search engine and how does it works? This concept is cleared out in this paper.The bounteous content on the World-Wide Web is so much useful to millions of people. An information seeker utilizes the search engine like Google, Yahoo to commence with their Web activity for a fruitful result.

The blog is a combination of TWO words. "Web" a blog is formed on the World Wide Web. "Log" a blog is a record of someone's opinions, thoughts and activities [1]. One can also utilize the blog as a personal diary or discovering a hobby or an infatuation and sharing information by writers, reporters, political leaders, scientists and many more. The blog sites on a server are a website but not all the websites on server are blogs. It is only and only determined by the presentation of blog or the way the blogs are followed their own design and format but neither by the content of the blog and the creator of the blog.

A web log is also called as a blog or WeBlog [6, 2]. It is a website that people are allowed to create and write their observations, suggestions, feelings, experiences and ideas on the extensive range of subject matter. The very first blog was made in Spanish for the education purpose. The second one is made by the well know writer who invites the other writers to come on his blog and show the interest of their own writing skills and participate actively.

Some blogs do not make the full use of the special features like blog posts, archives, template etc. which is also followed by the bulletin boards [7].

People are engaging towards blogs because people want to discuss and learn something from the discussion on the current affairs and famous personalities and also providing their own thoughts related to the topic [8].

Good blogs also have a viewpoint section of outlook that enlightens the reader something about the blogger's qualities. One must engage its reader on a blog by sharing their own views personally, frankly, honestly, also with a fruitful result and host must be committed towards its blog permanently. The host must not sell up their blogs and also don't cover up their mistakes. Blog Search Engines are deliberate in order to take plus point of blog formations.

However, blog search engines are existing from 2000, the traditional blogs and modern blogs are different from each other and modern blogs are more information oriented than traditional ones [10,11,12]

Blog searching is the most remarkable task ever. A blog is a small web sites containing entries in reverse chronological order that is, latest post date wise appears on top. The entries in the blogs are regularly updated by daily or weekly and monthly [13] such as political remarks [14]. The majority of blogs carry little influence level and the content of their logs are mostly unimportant [15].Consequently, if a person is following a particular blog than he/she can reach to limitless friends on that blog which increases a social network to share information [16].

One of the best example is the Google Blog Search Engine. It is a multi lingual, multi-platform blog search engine which is currently maintaining the largest index size which makes available the results with best match methodology and provides the fastest updating speed.

Blogs are categorized on the basis of the topics like environment, education, business, political, etc.frequent updates are available on blogs to aware the people. On the blog, posting can be done daily or weekly, monthly.The order followed by the blogs is the reverse chronological, that is, the latest post appears on the top with the details of title, author name, day,

date and time.One can comment on the blog to stimulate interaction and conversation.Blog maintains a record of their post comment and conversation which is stored as an archive by the blogger

The people want to influence the other people in the way they are thinking.The people want to motivate other persons to join and entertain the people.The people want to establish new contacts and network. Some people can also blog so that they can earn money by advertising their product on the blog.

There are some other definitions which must be cleared out to know more about blog.

Blogging- the act of reading and contributing towards a blog or the process of placing entries on the blog site [5].It is interactive in nature.It is just like a social networking site for other peoples where the people interact with others and explore themselves.Blogger- a person who reads and contributes to, perhaps maintains a blog (designs, formats, technical programming) [6]. The person has the authenticity to write on the blog and utilize the existing content management system such as WordPress, Google Blogger, TriPad, Drupal, Squarespace or uses the coding of web to write postings in a chronological order about a particular subject.

Blogosphere- the World Wide Web environment in which bloggers communicate with each other. It is a community of blogs and bloggers the individuals in community who blog

Post-it is also known as "Tweet" or "Postings" or a "Comment" or a "Feed". An individual entry of content on a blog

II. BASIC CONCEPT OF BSE

A search engine is a computer program that takes user query as an input. Or we can say that search engine is developed to find the information on www. It searches from database and returns a set of results, where database is created by the crawler. The result contains a list which is known as the "HITS" in the Information Retrieval term. These HITS contains the address link of the user query's solution. The searched information may be a collection of web pages, images, audio clips, video clips and may consist of some other types of files. Many search engines also searches data from the available database or also from the open source of directories. Search engines works fully on algorithmatically step by step. The result obtained with the help of crawler. The crawler is an employee just or an agent of the search engine which retrieves the pages. Indexing is applied on these retrieved pages which are done with the help of ranking algorithm. Ranking system ranks the each web page on the server and then fetches the best one for the user. The problem related with ranking is the maintenance, as there is tremendous variety of web pages on web server related to single topic and thousands of web pages are adding on the server day by day, so it's a crucial task for the ranker system to rank each pages and update itself. This whole searching process is done simultaneously and is also a recursive process, which repeats again and again when searching is done by the user.

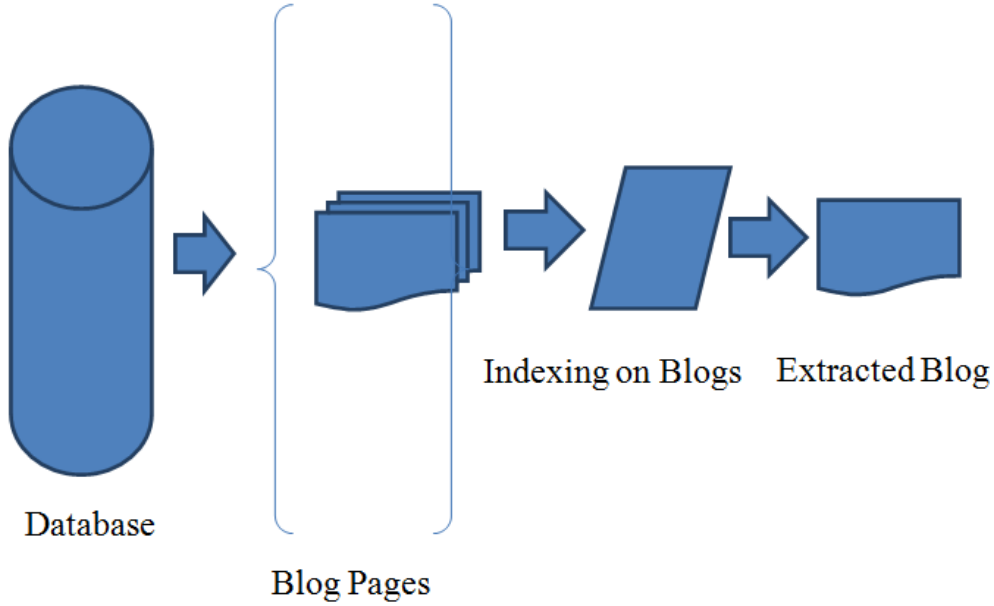

Figure1. Extraction Scenario

Search Engine is resided on the user side computer system which is further connected to Internet. When a user inputs a query into a search engine (typically by using key words), the engine check up its database which is created by web crawler and indexes each page and provides a listing ofbest-matching web pages according to its criteria.

The expediency of a search engine is totally depends upon the relevancy, that is, the result which a search engine provides is how much useful or helpful or relevant for the user. While there may be millions of WebPages that contains a particular word or phrase, so to some extend those pages may be supplementary appropriate, well-liked, or convincing than other result pages. Most of the search engines utilize the "best mach" results method to rank the pages

III. ARCHITECTURE OF BLOG SEARCH ENGINE

The working of the architecture is divided into following modules and is described as follows:

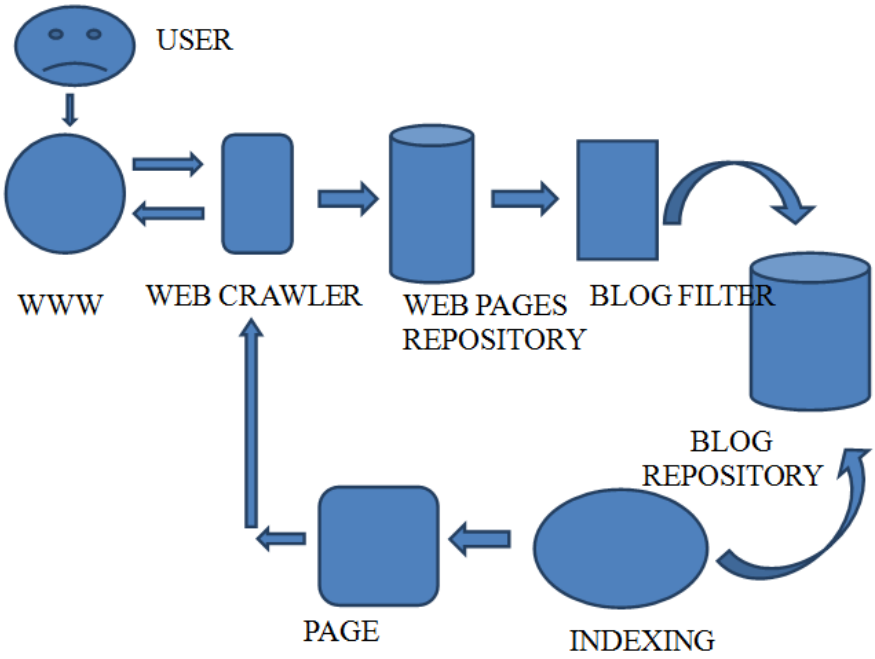

Figure 2: internal working of blog search engine

Blog Filter:- The blog filter is actually associated with the task to filter out or to remove the unwanted, superfluous, redundant and surplus and duplicated links of the blogs on server that are related to the specific topic or subject matter.
Blog Repository: The blog repository contains a queue having only blogs list in it which is related to the user defined query.Only theupdated url's of the blogs are stored in this repository.
Indexing: Search engine indexing is responsible to collect, parses, and stores the data to smooth the progress of fast and accurate information retrieval. The main aim of storing an index is to optimize the speed and performance in discoveryof the relevant documents for each search query. Without an index, the search engine would scan every document in the corpus, which would require a lot of time consumption and computing power which results in failure [3].
Page Rank: The page rank is basically related to the rank the pages on server which are up-to-date. They work with the help of algorithms. These algorithms are not publically available. They do not allow access to the text, but allow only to the only indices. Sometimes this concept brings too many relevant pages for a simple query. It is very hard to compare the quality of ranking for the two search engines together. For better results high ranking algorithms are preferred.
Web Crawler:A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are ants, automatic indexers, bots, Web spiders, Web robots [4]. Web worms, wanders, walkers. The web crawler runs on local machine and sends the request to remote server. The Web crawlers are used in a purpose to create a copy of all thevisited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. They obey in the breath-first search and depth-first search manner.
Web Page Repository:the web page repository working is same as that of blog repository instead of that it works on web pages not on blogs which are available on server. It available with the result of images, audio files, text files, video files and other files.
WWW: The www stands for the "World Wide Web". One can connect through the internet from anywhere and anytime. It provides the user friendly environment, so that user doesn't feel sophisticated while knowing the answer of their query.

IV. BENEFITS OF BLOG SEARCH ENGINE

There are many fundamental benefits of using blog search engine, some of them are described here:-
Blog allows almost all people to utilize the Web Site and writes their own observation, suggestions, thinking, ideas, experiences, and feelings on the ample choice of subject matter.It acts like one's own press which is absolutely free of cost. The posting is known as the "posts" or "comment" or" tweet"
One can get to know an idea of what is appreciating by the other peoples in your writing, one can comment on the same point or issue or topic and also takes interest in spreading their views by writing freely.
There are variety of blogs that are offering a debate section where one can get to know about the opinion and suggestions and idea of different types of topics.
The blogs are sorted by the date wise.The searching on blogs is done on in blog title, in title, in url and in authors name also. That is, the searching scans the query of a user from the title, abstract, content of the blog and within tweet.The recent tweet is visible on the top level of the blog posting column.It provides visibility and hence become a good book for others to learn the things.Overall result displayed by the blog search engine is very much relevant to the user.A blog is one of the best marketing tools for the web.There is no limit where a blog is taking you. One can reach to limitless possibility to meet people on a blog.

V. LIMITATIONS

This search engine has very low precision rate. This means the information needed by the user. It also related to the exactness or quality extent related to query.The Search engines provide efficient results if they come to know which indexing or crawling technique should be suitable for searching. A very less work has been done on page rank algorithm. Need to be furnishing a lot for better and fast results.Also, not a much more work is done on blog filters which actually provide the updated and most recent knowledge to the person.As compared to web search engine, blog search engine is not fully developed, the much more work is required to be done on BSE.

VI. CONCLUSION AND FUTURE WORK

The Limited work has been done on page rank and new and efficient algorithms can be constructed. Or better ranking algorithms have to be made.The work can also be done to provide better information filtering facts.Efficient blog filters algorithms are required more attention.Better Algorithms are required to construct that choose which pages to be indexed.New techniques are to invented to increase efficiency of the search engine.

REFERENCES

[1] BLOGRANGER: a multifaced search engine , IJERST by Ko Fujimura, Hiroyuki Toda, Takafumi Inoue, Nobuaki Hiroshima, RyojiKataoka, Masayuki Sugizaki

[2] Glance, N. S., Hurst, M., &Tomokiyo, T. (2004). *BlogPulse: Automated trend discovery for weblogs*, from http://www.blogpulse.com/papers/www2004glance.pdf

[3] "Information Retrieval in 21st Century" Text Book by PrabhkarRaghwan.

[4] A.K Sharma, J.P Gupta, D.P Aggarwal, "PARCAHYD: A Parallel Crawler based on Augmented Hyper text Documents", communicated to IASTED International Journal of computer applications, May.2005.

[5] http://www.merchantcircle.com/corporate/press/2010-11-04-MerchantCircle-Revitalizes-Bloglines.html

[6] Mohammad J. Kargar, Abd. R. Ramli, Humaidh, FatemehAzimzadeh "Formulating Priority Of Information Quality Creteria on the Blog", World Appled Sciences Journal4(4). PP- 586-593, 2008, ISSN 1818-4952.

[7] R. Ramakrishnan and A. Tomkins. Toward a People Web.COMPUTER, 2007.

[8] G. Mishne and M. de Rijke. A study of blog search.Proceedings of ECIR, 2006.

[9] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In WWW '03: Proceedings of the 12th international conference on World Wide Web, pages 568–576, New York, NY, USA, 2003. ACM Press

[10] Notess, G. R. (2002). The blog realm: News sources, searching daypop, and content management. *Online, 26*(5), 70-72.

[11] Bradley, P. (2003). Search engines: Weblog search engines. *Ariadne, 36*, available: http://www.ariadne.ac.uk/issue36/search-engines/.

[12] Curling, C. (2001). A closer look at weblogs.*LLRX.com*, available: http://www.llrx.com/columns/notes46.htm

[13] Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. In *Proceedings of the Thirty-seventh Hawaii International Conference on System Sciences (HICSS-37)*.Los Alamitos: IEEE Press.

[14] Trammell, K. D., &Keshelashvili, A. (2005). Examining new influencers: A self-presentation study of A-list blogs.*Journalism & Mass Communication Quarterly, 82*(4), 968-982.

[15] Weiss, A. (2004). Your blog?who gives a @*#%! *netWorker, 8*(1), 38,40.

[16] Bar-Ilan, J. (2005). Information hub blogs.*Journal of Information Science, 31*(4), 297-307.

BIOGRAPHIES

He received M. Tech and B. Tech. Currently, he is working as an Assistant Professor inComputer Science Department at MVN University, Palwal. His areas of interest include Operating System, Computer Network and Object Oriented Programming, Digital Image Processing.

