Review of Measures of Computing Page Importance in Web Crawling

Harjeetkaur and [1], Dr. Kanwal Garg[2]
[1]Scholar, M.Tech, Department of Computer Science and Applications, Kurukshetra University, India
[2] Assistant Professor, Department of Computer Science and Applications, Kurukshetra University, India

Abstract: Web crawling has a great literature of extensive research and optimization of various aspects of a crawler, but still there are different areas to explore to satisfy every user's need. Researches in the past emphasized on the relevancy and robustness of search results. In order to have in-depth knowledge of web crawling, various crawling strategies and algorithms along with their issues need to be studied. In this paper, the author has reviewed the different aspects of crawling millions of web pages on web. On the basis of this review, new strategies to improve the search results index of a search engine can be devised in future.
Keywords: Web Crawling, PageRank, Search Engine

1. INTRODUCTION

In today's highly competitive and connected world, internet has become an important source of information retrieval. Most of the users rely heavily on search engines for a useful web content search. The responsibility of search engines therefore has increased to maintain this trust of users by an efficient web crawling. This in turn demands the use of a crawler serving various optimization needs like relevancy of content, freshness of web page, vast coverage etc.

In its simpler form, a web crawler is a program to crawl a number of web pages through a graph of links on the web and downloads them to a local repository of search engine. In other words, a crawler start from a random page, also known as seed page, download it, parse it and check for external link and follow another external link. The process repeats until a sufficient amount of web pages supported by local repository are crawled.The different search engines employ different crawling strategies and page importance measures along with certain other attributes like anchor text, URL etc to rank the search results. Due to competitive business and to avoid spamming, page ranking algorithms are kept as secrets. This paper tried to explore the literature of web crawling sciences to aid academic research in subsequent sections.

2. LITERATURE REVIEW

To carry out this review work, papers from year 1998-2010 have been studied and it has been concluded that every crawler is designed to achieve some objectives according to some constraints. However the ultimate goal is to satisfy the user's need but researches, as described in the next paragraph, had been done either in the light of relevance or freshness of search results.

Every crawler is designed to cover the most part of web in order to identify the all possible requested pages. In [3] J. Cho and U Schonfeld deal with the two important goals of crawler namely coverage and efficiency. This work was important as it is costly to house a large corpus of web pages. As the web is very wide the efficiency of search result measured by search engine indexing is sometime misleading. So an algorithm was devised to provide guarantee that a crawler has downloaded the most important part of web before it stop crawling. The rankmass metric, a variant of personalized PageRank was proposed for comparing the quality of search engine indexes. Their crawler was focused on the user relevant pages and can prioritize the crawl downloading high personalized PageRank first and ultimately high rankmass is achieved when the crawl is over. The result of their experiments showed that their rankmass metric improves the PageRank of every page and allows search engine to specify the end of the crawl based on specific conditions. It ultimately reveals the fact that search engines can have smaller sized indices and still can be very effective in search results. Out of the number of web pages, only the most relevant and important pages are to be fetched by crawlers and ordered accordingly by search engines. There are number of criteria to estimate the importance of a page as discussed later. In [10], S. Pandey and C. Olston suggested query dependent approach, to compute the page importance of pages with high impact, to improve the quality of search results. In [11], Wenpu Xing and Ghorbani Ali also suggested a variant of PageRank algorithm to improve the relevancy of search results. In [7], S. Abiteboulet. al, Proposed an Online Page Importance Computation(OPIC) algorithm that aims to save the extra usage of available resources, by computing the importance of a page online during the crawling process. It can be focused towards fetching important pages and adapts dynamically with changing web through its adapted OPIC version

The technology is evolving day by day. It is therefore necessary to maintain the latest copies of web pages in a crawler repository to avoid producing stale results. In [12] W. Liu et al presented an approach to face this challenge through query related graph model that can fetch new web pages without crawling the entire web. In query related graph model, web database is represented as undirected graph where every record is represented by vertex in graph. An edge exists between two records if both the records have at least one common query in their query interface. The deviation between history version of web database and randomly generated samples of web database is analyzed through this model by generating their graph and measuring the vertex selection criteria 'sel(v)'.The larger value of this criterion denotes connectivity of vertex v with new vertexes. The appropriate query is thus generated to explore new records by estimating the effectiveness of a query. The experimental results proved significant reduction in crawling cost preserving coverage rate at the same time by this approach.

To explore the tradeoff between crawl size and effectiveness, [1] provided an evaluation framework for comparing different crawl policies. The stability of crawling policies over multiple iterations was also investigated through this framework. The maxNDCG metric was used to represent the effectiveness of different policies over a baseline breadth first crawl. Also evaluation was based on real human based relevance judgment along with click based relevance from Microsoft live search engine. The experimental results favor the best performance by PageRank strategy both in individual and iterative crawl selection. Also combination of trans-domain inlink and PageRank policy outperforms the PageRank alone.

3. WEB CRAWLING ISSUES

Web crawling can be seen as a three step process of a) searching, b) retrieving and c) maintenance of web pages. At each step a crawler has to consider certain issues like being polite to server, parallelization of crawling process, network and local hardware issues etc. An important issue is to divide the crawling resources among these tasks[10] The following two important aspects of crawling present the various issues involved

3.1 RECRAWLING

Since information on web changes continuously, recrawling the web is a necessity for all search engines. As certain links gets removed, added or updated with time, their preference in search process also gets changed. So search engines need to recrawl the web according to predefined recrawling schedule. The immense size of web does not allow exhaustive crawling approach to find the updates. The incremental crawling approach presented by [12] however did not presented the scheduling problem. The primary issue in scheduling recrawl is to determine the upper bound on the crawl for a given page as crawlers also cause performance problems [4]. The update patterns of different web sites need to be analyzed more efficiently in future to determine the recrawling schedule effectively.

3.2 EFFICIENCY

The success of a search engine depends on the efficiency of crawling policy employed by it. However evaluating various crawling strategies is not a simple task because what makes a crawler 'good' is not universally defined. Also user's perception of effective search result varies from user to user. The very large scale of web is the main issue to compute the page importance efficiently. OPIC algorithm computed page importance that depends on the entire web by looking at one page at a time but the algorithm needs to be generalized for link matrices other than google's link matrix[7]. Inspite of the coverage guarantee provided by[3], upper bounds on rankmass was not discussed that can ensure higher quality documents earlier in the crawl. The recrawling schedule of the proposed algorithms was also not considered. The crawling strategies can be evaluated only on a subset of web due to infinite web and finite resources. The continuously changing web pages and time constraints of crawling real web also makes comparisons of different crawling strategies difficult.

The other experimental barriers for evaluation are large communication and computational cost for conducting multiple crawls and network related issues.

The evaluation framework of [1] provided avenue for comparing different other crawling strategies in future. Discovering the suitable domain limiting criteria to balance between shallow and deep crawling is a significant area for future work.

4. CRAWLING STRATEGIES

This functioning of crawler can be achieved through various strategies. Some of the worth mentioning strategies are:

4.1 BREADTH FIRST SEARCH [6]

According to this strategy, search is started from seed page and continues to all the immediate neighboring pages. Only after crawling all the first level web pages, it moves to the next level web pages. It is a kind of natural crawling but is biased towards retrieving important pages earlier as reported by [6].

4.2 DEPTH FIRST SEARCH [4]

In this strategy, search is done across the depth of web graph. Only after reaching the deepest link after which no link is present, the neighboring pages can be crawled.

4.3 BEST FIRST SEARCH

This strategy aims to retrieve the most relevant result to the query. Such type of strategy is used in focused crawling. The goal of a focused crawler is to selectively seek out pages that are relevant to a pre-defined set of topics .

4.4 WEB FORUM CRAWLING STRATEGY

To crawl a web forum is difficult from crawling general websites in two ways. Firstly a single post in web forum, presented through multiple pages, creates the problem of page-flipping. Secondly, the access control of a web forum leads to crawling of invalid pages like login portal. [13] Proposed a different traversal strategy for crawling web forum. They recognized that a generic crawler only consider out-links based information but a web forum has a complex in-site link structure. So while crawling a web forum through generic crawler, many duplicate results are shown. To overcome this problem, the web forum crawling strategy focused on which links to follow and how to follow these links. Prior to this, a sitemap was required by the strategy. The traversal strategy works in following two steps:

i) Skeleton Link Identification: Out of number of links present in web forum, only the most relevant links present in web forum, only the most valuable and valid links are identified based on the criteria of coverage and informativeness.
ii) Page-flipping link detection: The selected skeleton links are further observed to detect page-flipping links. These links helps a crawler to recognize different threads. A measurement called connectivity was defined which score page flipping links higher than other loop back links.
This strategy was evaluated successfully in terms of crawling quality and crawling effectiveness and efficiency. However the recrawling schedule of a highly dynamic web forum has been left as future work. Also the coverage criteria set for identifying skeleton links is too general and need to be refined against useless pages.

4.5 IMPACT-DRIVEN CRAWLING

The impact of a page depends on the query for which page is relevant, the rank achieved by page and user’s interest in that page. The overall goal is to fetch the pages with highest impact by estimating the neediness and relevance of queries to a page.As some of the web pages might not have matching query content but are more impactful due to high PageRank, so this approach alone is not as fruitful and needs to be supplemented with query-independent information.

5. PAGE IMPORTANCE METRICS

The present state of web crawling considers various matrices of computing page importance as studied by [2]. The literature is more biased towards link based importance as it is simpler to compute than other measures and scale well with the growth of web. Some of the algorithms using link based metric are as follows:

5.1 PAGE RANK

In [8] the importance of web pages was based on a PageRank metric. It state that if a page has important links to it , its link to other pages also contribute to their importance and a page with high PageRank is most relevant page to be downloaded. The PageRank of a page A is given by:
$PR(A)= (1-d)/|D| + d(PR(T1)/C(T1)+….+PR(Tn)/C(Tn))$
Where PR (A) = PageRank of page A
T1....Tn= inlinks to page A
C(A) = no. of links going out of page A
D = set of all web pages
d= damping factor which is often assumed to 0.85.

This metric measures the importance of a page very effectively as shown by [2] [6] but require multiple calculations over a large web graph and can be easily spammed. The two important variations of PageRank proposed in literature are

5.1.1 WEIGHTED PAGERANK

In [11] an improved version of PageRank which assigns more value to more important pages instead of dividing the rank value of a page evenly among its entire outgoing links. The weighted PageRank is thus given by:
PR(u)= (1-d)/|D| + d(PR(V1)Win(V1,u)Wout(V1,u)+….+PR(Vn)Win(Vn,u) Wout(Vn,u)
Where Win(v,u) = weight of link(v,u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v
Wout(v,u) = weight of link(v,u) calculated based on the number of outlinks of page u and the number of inlinks of all reference pages of page v
PR(u)=PageRank of page u

5.1.2 PERSONALISED PAGERANK [3]

To determine the importance of pages, their PageRank was computed using personalized PageRank that assumed that a user goes to a trusted site rather than to a page of equal probability. So the PageRank of page A is then defined as
PR(Ai)= (1-d)(Ti) + d(PR(R1)/C(R1)+….+PR(Rn)/C(Tn))
Where PR (A) = PageRank of page A
R1....Rn= inlinks to page A
Ti = trust score of page i
C(A) = no. of links going out of page A
d= damping factor which is often assumed to 0.85
As users are unlikely to go a single trusted page a new form named windowed rankmass algorithm was adapted in [3]. This new form batches together sets of probability calculation and downloading sets of pages at a time thereby reducing the computational overhead.

5.2 OPIC

The change in importance of a page as the web changes is considered as important aspect of page importance in [7]. According to this algorithm, two values are computed for each page. First is ‘cash’ which is a value distributed equally to all pages. Second is credit history of the page which is cash accumulated until last crawl of that page. The cash value is stored in main memory to avoid disk access. For each retrieved page ‘i’, its cash value is added to ‘history’ and distributed equally among all its outlinks. The cash is then reset to zero. The whole process is repeated with continuous crawl. The page with higher amount of cash is considered important for crawling. It is faster than PageRank and is simpler to converge. However this algorithm is more expensive than other off-line algorithms. Other variant like adaptive OPIC are in consideration to tackle its disadvantages. Better importance estimates are in research like tuning of algorithm and choice of time windows.

5.3 HITS

At the time PageRank was developed, [5] also proposed Hypertext Induced Topic Selection (HITS) algorithm. It was a precursor to PageRank so can be considered as an alternative importance metric. According to this algorithm certain web pages that point to many hyperlinks are known as hubs other web pages that are pointed by many hubs are known as authorities. An authority pointed by highly scored hubs get high score and at the same time a hub pointing to number of authorities should have high score. This paper was also based on the link structure of web but fail during more focused queries as it assigns equal weights to all out links.

6. Conclusion

The web crawling emerged as a subject of research due to the wide growth of web and the increasing number of queries to be handled by search engines .The researchers have contributed towards the improvement of search engine technology by proposing various crawling approaches. The experimentation barrier however hinders the high quality research work. The goal to improve efficiency is therefore still in process. The various algorithms studied in this paper are thus needed to be refined. However there are different issues of importance metrices, dynamicity, scheduling that need to be considered somewhere in future research to aid efficient web crawling.

References

[1] Dennis Fetterly, Nick Craswell, VishwaVinay, The impact of Crawl Policy On web Search Effectiveness, in Proceeding of SIGIR, July 2009.

[2] J. Cho, H. Garcia-Molina, and L. Page. Efficient crawling through URL ordering. Computer Networks and ISDN Systems, 30(1-7):161–172, 1998.

[3] J. Cho and U. Schonfeld.Rankmass crawler: a crawler with high personalized PageRank coverage guarantee. In Proceedings of VLDB, pages 375–386, 2007.

[4] J.L. Wolf, M. S. Squillante, P. S. Yu, J. Sethuraman and L. Ozsen. Optimal Crawling strategies for web search engines. In proceedings of the 11thintenationalworld wide web conference, 2002.

[5] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, September 1999.

[6] M. Najork and J. L. Wiener, Breadth first crawling yields high-quality pages, In Proceedings of the TenthConference on World Wide Web, pages 114, May 2001.

[7] S. Abiteboul, M. Preda, and G. Cobena.Adaptive on-line page importance computation. In WWW '03: Proceedings of the 12th international conference on World Wide Web, New York USA, pages 280–290, 2003

[8] S. Brin and L. Page.The anatomy of a large-scale hypertextual Web search engine.ComputerNetworksand ISDN Systems, 30(1{7):107{117, April 1998.

[9] S. Chakarbarti, M. van den Berg, and B. Dom. Focused Crawling: A new approach for Topic-specific Resource Discovery. In Proc. 8th WWW,1999

[10] S. Pandey and C. Olston.Crawl ordering by search impact. In Proceedings of WSDM, pages 3–14, 2008

[11] Wenpu Xing and Ghorbani Ali, Weighted PageRank algorithm, In Proceeding of the second annual conference on Communication Networks and Services Research(CSNR' 04),IEEE,2004

[12] W. Liu, J. Xiao and Jianwu Yang. A Sample-guided Approach to IncreamentalStrustured Web Database Crawling.Proceedings of the IEEE international conference on Information and Automation, june 20-23,2010.

[13] Y. Wang, J. yang, W. lai, R. Cui, L. Zang and Wei-ying Ma. Exploring Traversal strategy for web forum crawling. SIGIR, JULY 2008.