DOI: 10.17148/IJARCCE.2023.12814

Page Relevance Computation Techniques

Dr. Kompal

Govt. College, Panchkula

Abstract: In today's world, the rising popularity of the web has drastically expanded the probability of sharing significant information and knowledge on a large scale certainly not seen before. Owing to the availability of the bulk amount of data, the search services on the World Wide Web (WWW) are fetching more demand among users. Regardless of its beneficial part by conventional term-based search engines, precise filtering and retrieving relevant data from the web is considered as a challenging task. In fact, page relevance is the fundamental aspect for the web search, as it supports the current and novel search engines, indexing, crawling, and ranking. Relevancy and popularity of a website are two different things in the world of search engines.

I. INTRODUCTION

Search engines generally involve bulky web indexes and they are the most commonly and significantly utilized tools [1]. The necessity for the allocation of resources for downloading, indexing, harvesting and storing the web has turned out to be the opaque requirement. For the purpose of indexing the significant pages into dispersed repositories and varied topics of human knowledge, focused crawlers serve as inevitable tools that are aimed for detecting pages, thereby making the fast retrieval process via the merged index. Since maintaining an older version of the web is old-fashioned, the practice of indexing helps to survive with the alterations done in the active contents on the web. This is the main motive for the service of reinforcement learning strategies or machine learning within a few search engines to conscientiously observe the web. Generally, multiple powerful spiders find a way of collecting web information space using web search engines. For searching data pertain to a concerned topic, a web crawler searches through the entire web servers. Owing to continuous progression and refreshment of the web, it is not realistic to identify the entire web pages and servers. Dealing with such a huge amount of data is still a tough task by the engines. Because of the network resources and the required hardware, navigating the web completely in quick is an unrealistic, expensive goal [2] [3]. For maintaining the copies of data afresh, all search engines utilize web crawlers internally. Since the Google crawlers perform speedy parallel processing, the incorporation of numerous low-priced computers makes it possible to execute on a distributed network, and provides the output in a short duration.

II. IMPORTANCE OF PAGE RELEVANCE

The first thing a user must perform when a query is given to a search engine is to decide which pages in the index are associated with the query and which are not.

2.1 Computation of Page Relevance

The computation of page relevancy in an enormous dynamic graph has just concerned high attention. The fixpoint of a matrix equation is defined as the page rank or page importance [4]. Previous algorithms utilize additional disk and CPU resources and calculate it off-line. A new technique called OPIC utilizes much fewer resources and works on-line. Storing the link matrix is not needed in common. While the web/graph is visited, it constantly refines its estimate of page importance in on-line. Most interesting pages can thus be utilized to focus on crawling. In the IR model, the most significant point is the requirement of numerous phases for performing the pagerank computation. Initially, it is required to match the query with the diverse pages on the web, and extract the relevant data. Further, the values attained for the page rank arranges the relevancy set in order. In almost all cases, the pagerank provides first-class approximation for query relevance. Moreover, the quality of searching in IR system is enhanced by more new techniques pertain to text searching. For instance, a weight is assigned to terms in pages of web by the Google. The bold font usually takes higher weight, which is mostly for the titles of the article. The external data termed as meta-data is involved with data regarding the status of the "source, frequency of updates, and usage statistics". As the inlink text usually provide precise data related to the page content, the text corresponding to the outgoing link is connected with both linkage and linked pages [5]. Even though pagerank method is extremely hard to spam, it is not impermeable. In the link farms, the entire pages in the network connect to each other, which then connects to another or separate page. Further, the addition of the pages with high page rank is done with the corresponding inlink. A web crawl should determine the transition probability matrix before computing the pagerank for the known web.

DOI: 10.17148/IJARCCE.2023.12814

The page relevance is considered as the main feature to be considered for web search, due to the availability of modern search engines, in which indexing, ranking and crawling are the general steps to concentrate. Instead of downloading the entire web pages like existing search engines, the web crawler generally downloads only the most relevant or significant web pages mostly connected with the requirement of the user. Therefore, the preliminary aim of focused crawler is to search for the most informative web pages that satisfy the user's requirements. Depending upon the selection and ranking strategies used for downloading the relevant web pages, the URLs are utilized to prioritize the page ranking method, which is one of the link analysis algorithms. The technique utilized by search engines for gathering from the net is referred to as the web crawler. A major challenge that happens within the field of IR Systems is the need for an online crawler that downloads most relevant web content from an oversized internet. For retrieving the knowledge from web, most internet crawlers utilize keyword base techniques. The major constituent of the search engine is the software program termed as the web crawler. Crawler is also termed as a computer code agent or spider. Generally, a web crawler initiates its operating victimization seed address by acting as an associate initial address for the creep method. Once going to the net page of seed address it converts that online page and accumulates the entire links to the queue and thus collects every hyperlinks available there in the downloaded online page. This method is also called a frontier that recursively continue the process till it attains the optimal outcomes [6]. Regarding the priorities that are arranged in the queue, the major idea of a web crawler is to visit the significant pages and to download only important pages.

The overall significance of the pages relative to the location and occurrence of the searched term or word estimates the search engine employing its algorithm and the algorithm used by individual search engines is unknown. For example, the search algorithm utilized by Google, varies frequently as Google tries to increase its search engine's relevance and results. In the earlier period, for indicating a page's relevance for a particular query, search engines observe keywords in assured locations of the HTML code. Nowadays, keyword-based algorithms and relevancy utilized by the Bing and Google ranked and evaluated pages that are extremely more difficult. The user's experience with the site relies the overall rankings by attaining a trivial assistance in a keyword [7] placement-based algorithmic element. Generally, an optimized page is the one that concerns distinctive value and content. Search engines search this distinctive value and they are more intelligent in which links, positive associations, social shares, and branding come together for producing all the right signals to impel the website to the top.

III. CONCLUSION

Search engines are more concerned with an optimized page rather than a page completed with keywords. The off-topic areas are not navigated by the crawlers. With the charge of navigating the fewer links, it is possible to search for more relevant pages with a suitable navigational order using topic specific crawlers. Moreover, the exact ordering of the unvisited pages that could be visited by the crawler afterward is the main challenge faced by the topic specific crawlers.

REFERENCES

[1]. H. Ishii and R. Tempo, "Distributed Randomized Algorithms for the PageRank Computation," IEEE Transactions on Automatic Control, vol. 55, no. 9, pp. 1987-2002, Sept. 2010.

[2]. Takayuki Yumoto, Ryohei Tada, Manabu Nii and Kunihiro Sato, "Finding Rare Web Pages by Relevancy and Atypicality in a Category," 2013 Second IIAI International Conference on Advanced Applied Informatics, Los Alamitos, CA, pp. 284-288, 2013.

[3]. Jayendra Singh Chouhan and Anand Gadwal, "Improving web search user query relevance using content based page rank," 2015 International Conference on Computer, Communication and Control (IC4), Indore, pp. 1-5, January 2015.

[4]. Hassan Artail and Kassem Fawaz, "A fast HTML web page change detection approach based on hashing and reducing the number of similarity computations", Data & Knowledge Engineering, vol. 66, no. 2, pp. 326-337, August 2008.

[5]. M. M. El-Gayar, N. E. Mekky, A. Atwan and H. Soliman, "Enhanced Search Engine Using Proposed Framework and Ranking Algorithm Based on Semantic Relations," in IEEE Access, vol. 7, pp. 139337-139349, 2019.

[6]. Chih-Ming Chen, Hahn-Ming Lee, Yu-Jung Chang, "Two novel feature selection approaches for web page classification", Expert Systems with Applications, vol. 36, no. 1, pp. 260-272, January 2009.

[7]. Ying Zhang, Weinan Zhang, Bin Gao, Xiaojie Yuan, Tie-Yan Liu, "Bid keyword suggestion in sponsored search based on competitiveness and relevance", Information Processing & Management, vol. 50, no. 4, pp. 508-523, July 2014.