October 2013

WEB GRAPH BASED SEARCH BY USING DENSITY OF KEYWORD AND AGE FACTOR

GAURAV AGARWAL

Computer Science & Information Tech. Dept., Krishna Institute of Management & Technology, Moradabad, India, meetgaurav1@rediffmail.com

SACHI GUPTA

Computer Science & Information Tech. Dept., Krishna Institute of Management & Technology, Moradabad, India, sachiagarwal@rediffmail.com

SAURABH MUKHERJEE

Computer Science Department, Banasthali Vidyapith, Niwai, India, mukherjee.saurabh@rediffmail.com

Follow this and additional works at: <https://www.interscience.in/ijcsi>

Part of the [Computer Engineering Commons](#), [Information Security Commons](#), and the [Systems and Communications Commons](#)

Recommended Citation

AGARWAL, GAURAV; GUPTA, SACHI; and MUKHERJEE, SAURABH (2013) "WEB GRAPH BASED SEARCH BY USING DENSITY OF KEYWORD AND AGE FACTOR," *International Journal of Computer Science and Informatics*: Vol. 3 : Iss. 2 , Article 4.

DOI: 10.47893/IJCSI.2013.1124

Available at: <https://www.interscience.in/ijcsi/vol3/iss2/4>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer Science and Informatics by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

WEB GRAPH BASED SEARCH BY USING DENSITY OF KEYWORD AND AGE FACTOR

GAURAV AGARWAL¹, SACHI GUPTA² & SAURABH MUKHERJEE³

^{1&2} Computer Science & Information Tech. Dept., Krishna Institute of Management & Technology, Moradabad, India

³Computer Science Department, Banasthali Vidyapith, Niwai, India

E-mail : meetgaurav1@rediffmail.com, sachiagarwal@rediffmail.com & mukherjee.saurabh@rediffmail.com

Abstract - Today, web servers, are the key repositories of the information & internet is the source of getting this information. There is a mammoth data on the Internet. It becomes a difficult job to search out the accordant data. Search Engine plays a vital role in searching the accordant data. A search engine follows these steps: Web crawling by crawler, Indexing by Indexer and Searching by Searcher. Web crawler retrieves information of the web pages by following every link on the site. Which is stored by web search engine then the content of the web page is indexed by the indexer. The main role of indexer is how data can be catch soon as per user requirements. As the client gives a query, Search Engine searches the results corresponding to this query to provide excellent output. Here ambition is to enroot an algorithm for search engine which may response most desirable result as per user requirement. In this a ranking method is used by the search engine to rank the web pages. Various ranking approaches are discussed in literature but in this paper, ranking algorithm is proposed which is based on parent-child relationship. Proposed ranking algorithm is based on priority assignment phase of Heterogeneous Earliest Finish Time (HEFT) Algorithm which is designed for multiprocessor task scheduling. Proposed algorithm works on three on range variable its means the density of keywords, number of successors to the nodes and the age of the web page. Density shows the occurrence of the keyword on the particular web page. Numbers of successors represent the outgoing link to a single web page. Age is the freshness value of the web page. The page which is modified recently is the freshest page and having the smallest age or largest freshness value. Proposed Technique requires that the priorities of each page to be set with the downward rank values & pages are arranged in ascending/ Descending order of their rank values. Experiments show that our algorithm is valuable. After the comparison with Google we find that our Algorithm is performing better. For 70% problems our algorithm is working better than Google.

Keywords- Search Engine, Density of keywords, Number of successors, Age of web page (Freshness value), Task Scheduling.

I. INTRODUCTION

The World Wide Web has become a new communication medium with informational, cultural, social and evidential values after a few decades since its inception. Search engines are widely used for Web information access and they are making more information easily accessible than ever before [1, 3, 4]. Although the general Web search today is still performed and delivered predominantly through search algorithms, e.g., Google's Page Rank, the interests in helping Web users effectively get their desired Web pages have been growing over the recent years.

In proposed work, we study two strategies which are widely used to enhance the retrieval quality of Web search. One is personalized re-ranking. To satisfy the diverse needs of users, we improve the global notion of importance in ranking search results by creating personalized view of importance [3, 4]. A set of efficient mechanisms are devised to re-rank search results by personalized ranking criteria which are typically derived from the modeling of users' search behavior and interests. The other is keyword suggestion. Our work focuses on finding semantically

related search queries (though probably dissimilar in their terms). We propose

using additional sources to enrich short Web queries and compare different feature spaces to better represent the meanings of a query than the keywords in defining it. Thus, the relatedness between two queries can be estimated based on their enriched representations. The suggested keywords can assist Web users in rephrasing their query formulation. We show that personalized re-ranking and related keyword suggestion are effective strategies to enhance Web search.

II. WORKING OF SEARCH ENGINE

The search engine [2, 9, and 10] is one that helps us to find exactly what we are looking for. Early search engines held an index of a few hundred thousand pages, and received one or two thousand user queries each day. Nowadays, an efficient search engine indexes hundreds of millions of pages, and respond to tens of millions of queries each day. In this section, we explain how a search engine works, from the indexing of the pages to the selection of relevant pages regarding to a query.

Search engines work in two major steps:

- Create a database of indexed pages.
- Search in this database the relevant pages regarding to a user query.

The first step is known as Web crawling, and the second step uses efficient ranking [4] and filtering.

Crawling [6] is a software program which follows links available on the web to different websites, and gathering the contents like web pages, images, documents and other files of these websites for storage in the search engines page repositories. Crawling can be start with a well-liked website containing lots of links or from existing older indexes of websites. A software program that does crawling is known as web crawler or a web spider. Once a page has been crawled, its contents can be indexed in a giant database. A crawler works in following steps:

- Start from stored web pages,
- take out their links,
- track these links to locate new pages,
- append them in the index,
- Goto first step.

After crawling, a huge volume of pages are indexed in a page repository or database. Now the main aim of a search engine is to sort and filter them in order to extract only relevant pages [5]. A Web page is relevant to a query, if it contained high density of keywords related to subjects contained in the query. For finding these important pages ranking is done. Ranking algorithm does not rank web sites as a whole, but is determined for each page individually. After ranking pages are arranged in order of their increasing or decreasing rank values. After sorting, pages which have high relevancy to the user query send back to the end user.

Figure1: Working of Search Engine

III. RELATED WORK

In literature, three approaches to calculate the rank of the pages of web graph are:

A. Ranking Based on Human Efforts

Human editors have been used by companies such as yahoo to manually correlate a set of keywords with a subset of documents on the web [7,8]. These are then matched against the user's query to return

relative matches. The problem with this method is that: (a) it is slow and can only be applied to a small number of pages, and (b) often the keywords assigned by the human judges are incomplete.

B. Ranking Based on practice of Information

Some services such as Direct Hit collect information on the queries individual users submit to search engines and the pages they look at subsequently and the time spent on each page. This information is used to return pages that most users visit after set up the given query [9]. A huge amount of space is required to collect this necessary information for each query. Thus, on a small range of queries this technique can be applied.

C. Ranking Based on Hyperlinking

This approach works on number of links between the pages on the web on the assumption that pages on the topic link to each other[9]. Today's most of the search engine works on this technique.

IV. TYPES OF SEARCH ENGINE

- CRAWLER-BASED SEARCH ENGINE:** In this type of Search Engine files are arranged automatically. If a change is occurred in any page, search engine knows these changes.
- HUMAN-POWER DIRECTORIES:** This is similar to Directory .Human is responsible for its listing.
- HYBRID SEARCH ENGINE:** Hybrid searches Engine are combination of Crawler based and Human Powered Directories.

V. PROBLEM DESCRIPTION

A Page Graph System model consists of an application (Page graph), a target computing environment (no. of successors), and a performance criteria for scheduling (page rank must be high):

Fig.2: Directional Web Graph

- Directional Web Graph:** Web graph is an acyclic graph, in which nodes are representing the web pages and the edges are representing the links between the web pages. A page graph is represented by a directed acyclic graph (V, E) Where V is the set of v linked pages and E is the set of e Links between the Pages. Each Link (i,j) represents the precedence constraint such that Page N_j can only be reached from Page N_i (see Fig.2) In a given Page graph, a Page without any

parent is called an entry Page and a Page without any child is called a target Page.

B. In a given Page graph, a Page without any parent is called an entry Page and a Page without any child is called a target Page. In Fig 3. Here each page is indicated by a pair of three numbers where first Number shows the density value of the keyword on that page and the second Number shows the level of the page. Density is considered as the number of appearance of keyword on the page. The Numbers on the links indicated the number of successors of any page using which a target page can be reached; The AGE shows the freshness value of every page The Performance criterion of calculating the ranks of each and every page works in such a manner that ranks assigned to the pages should be high.

Fig.3: Shows Page Graph with Density value of keyword Level of Page 1 and Age

VI. PROPOSED RANKING ALGORITHM

To search any information available on the web, end user submits the relevant query to the search engine. Search engine looks into the page repository for the relevant pages already indexed by the web crawler. For sorting these pages according to their relevancy to the user query search engine can use proposed AGE algorithm (see Fig. 4). Proposed, AGE algorithm is highly influenced by HEFT algorithm. HEFT algorithm is divided into two phases:

1. Task Prioritizing Phase

This phase requires the priority of each task of a Task Graph to be set with the upward or downward rank value which is based on mean computation cost (Execution time). The task list is generated by sorting the tasks by decreasing or increasing order of $rank_u$. Tie-breaking is done randomly. It can be easily shown that the decreasing order of $rank_u$ values provides a topological order of tasks, which is a linear order that preserve the precedence constraints.

2. Processor Selection Phase

When rank for each task is estimated then the best suited processor is selected in such a way that after assignment of the task on that processor scheduling time remains minimum.

Proposed algorithm follows the first phase of HEFT algorithm. This technique requires that the priorities of each page to be set with the downward rank value, $rank_u$, calculated in the page graph, which is based on Density of the Key word on that page and the Number of out going links from that page. Priorities are assigned according to the highest Rank value. Tie-breaking is done randomly.

Fig. 4 Search Engine Architecture

Proposed AGE Algorithm for the ranking is as follows:

1. Set the Density of keyword on pages as initial rank of pages.
2. Set the number of outgoing links for every page.
3. Set the AGE of every page.
4. Compute $rank_u$ for all the pages by traversing graph downward, starting from the Entry page to the Target page.
5. **While** there are unranked searched pages in the queue **do**
6. Select the first page, p_i from the list of pages.
7. **For each** page T_k in the page-set p_i **do**
8. Compute $rank_u$
9. **End While**
10. Sort the pages in a list by non increasing order of $rank_u$ values.

Where

P_i = Index pages

T_k = Pages under the index page(P_i)

$rank_u(\text{Entry Page}) = (\text{Density value of keyword on Target Page} / 100) + 1/\text{Age and}$

$rank_u(\text{Any other Page}) = \text{Density of keyword on this page} / 100 + 1/\text{Age value of this page} + \text{Max Successor Rank Value.}$

A Heuristic Based AGE Algorithm for Search Engine is basically the technique of calculating the rank of the page. Each page contains the calculated rank. The pages are arranged in ascending/Descending order such that the web page having the best priority.

VII. EXPERIMENTS AND RESULTS

In this work, implementation of proposed algorithm is done for solution of web page ranking problem. For performance evaluation of proposed algorithm we generated some problems of varying sizes and solved them by proposed algorithms. Details of experimental setup and results obtained by proposed algorithm are as follows:

A. Experimental Setup

A system is implemented to automatically generate the ranking problems of required sizes. This is done to avoid biasing in giving values of different parameters required for the problems. This system fits random values to these parameters in appropriate ranges. Few problems have been generated for experiments with the following characteristics:

- There are 15 pages return by the search engine for any search.
- There is no limit on the number of successors (outgoing links) of each page.
- The density of keyword within each page is a random number between 5 and 25.

B. Results of Google's Page Rank Algorithm

Page Rank is Google's way of deciding a page's importance. It matters because it is one of the factors that determine a page's ranking in the search results. Page Rank is a numeric value that represents how important a page is on the web.

We run Page rank algorithm on ten different problems with Problem Identification Numbers (PIN) 0 to 9 for each problem to note the ranks of pages (see Table 1). In Table 1, symbol 'A' represents ranks return by Page Rank algorithm. We set following parameters for Google's Page Rank Algorithm:

- Damping Factor=0.85
- No. of iterations=20

C. Results of Proposed Algorithm

The proposed web graph based ranking algorithm discussed in previous section was implemented and evaluated on the same set of problems used to evaluate Google's Page Rank Algorithm. Results obtained are shown in Table 2. In table 1 symbol 'B' represents ranks calculated by AGE algorithm.

Comparison of Google's page rank and AGE algorithm

Fig 5. Comparison of Google's Page Rank and AGE algorithm for PIN 1.

Fig 6 Comparison of Total Rank assigned by Google's Page Rank and AGE algorithm for PIN (1-10).

Fig 7. Comparison of Google's Page Rank and AGE Algorithm for PIN (10) with respect to density of keywords.

CONCLUSION

A Heuristic Based AGE Algorithm for Search Engine. in which we used three on range variable its means the density of keywords, number of successors to the nodes and the age of the web page. This research has solved the problem of how accordant data can be found in mammoth data on Internet. Sometimes this task seems to be as finding a pearl in an ocean. we compared this algorithm with Google Page rank Algorithm. After Comparing the results of our proposed Algorithm (A Heuristic Based AGE Algorithm for Search Engine) we found 70% results are better than Google. Results comparisons are shown in fig 5.

ACKNOWLEDGMENT

We would like to thank the anonymous referees for their helpful comments and suggestions that have improved the quality of this manuscript. I would also like to express sincere gratitude to Krishna Group for the sponsorship.

REFERENCES

- [1] Bela Gipp, Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study). Proceedings of the 3rd International Conference RCIS'09, 22-24 April 2009, pp. 439-446.
- [2] Christine Churchil, Search Engine Algorithms And Research Search Engine Watch, April 2005.
- [3] H. Topcuoglu, M.Y. Wu. Performance-effective and low-complexity task scheduling for heterogeneous computing. IEEE Transactions on parallel and distributed System, Vol. 13, pp.260-274, 3, 2002.
- [4] Johan Oudinet, Search Engine Ranking. Technical Report no0617, July 2007, revision 1126.
- [5] Jonathan I. Gross, Jay Yellen. Handbook of Graph Theory. CRC Press.
- [6] Junghoo Cho, Hector Garcia-Molina, The Evolution of the Web and Implications of a Web Crawler.
- [7] Martin Charles Golumbic. Algorithmic Graph Theory and Perfect Graphs. Second Edition, Elsevier, 2004.
- [8] Sean A. Golliher, Search Engine Ranking Variables and Algorithms. Semj.org Volume 1, Supplement Issue, August 2008.
- [9] Sergey Brin, Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine.
- [10] Sunny Lam, The Overview of Web Search Engines.