International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 8, October - 2012

A Heuristic Based AGE Algorithm For Search Engine

Harshita Bhardwaj1
Computer Science and Information Technology Deptt.
Krishna Institute of Management and Technology, Moradabad, Uttar Pradesh, India

Gaurav Agarwal2
Computer Science and Information Technology Deptt.
Krishna Institute of Management and Technology,Moradabad, Uttar Pradesh, India

Chanchal Dhaka3
Computer Science and Information Technology Deptt.
Krishna Institute of Management and Technology,Moradabad, Uttar Pradesh, India

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.

1. Introduction

Many kinds of web search engines, such as Yahoo!, AltaVista, g00, fresheye, Google and so many others, have been developed. However, as the volume of data on the web is increasing rapidly. There is a mammoth data on the Internet. So now we have problem is how accordant data can be found in mammoth data on Internet. Sometimes this task seems to be as finding a pearl in an ocean. So to sort out this problem we enrooted an algorithm, A Heuristic Based AGE Algorithm for Search Engine, by which we can easily search accordant data. There are three on range variable on the basis of these three variables our algorithm is working i.e. Density of keywords, number of successors to the nodes and the age of the web page

2. WORKING OF SEARCH ENGINE

A search engine works on following these steps:

- Web crawling by crawler
- Indexing by Indexer
- 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.

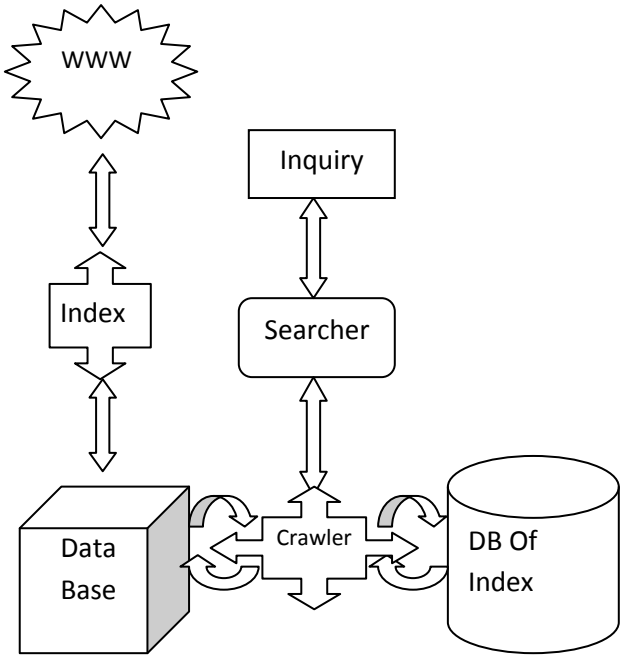

Figure1: Working of Search Engine

3. TYPES OF SEARCH ENGINE

3.1 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.

3.2 Human-Powered Directories: This is similar to Directory .Human is responsible for its listing.

3.2 Hybrid Search Engine: Hybrid searches Engine are combination of Crawler based and Human Powered Directories.

4. PROBLEM FORMULATION

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).

- 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.

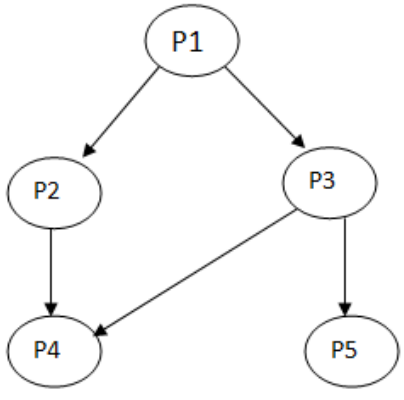

Fig.3: Directional Web Graph

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 Nj can only be reached from Page Ni (see Fig. 1).

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.

Density of keyword, level,Age)

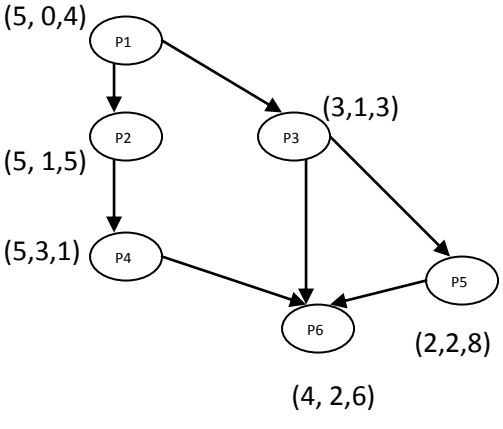

Fig.2: Shows Page Graph with Density value of keyword Level of Page 1 and Age.

Here each page is indicated by a pair of two 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.

5. PROPOSED WORK

Algorithm:
(A Heuristic Based AGE Algorithm for Search Engine)

1. Set the Density Value 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 ranku for all the pages for the first searched page by search engine in database 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, pi from the list of pages.
7. For each page Tk in the page-set pi do
8. Compute ranku
9. End While
10. Sort the pages in a list by non increasing order of ranku values.

Where

Pi = Index pages

Tk = Pages under the index page(Pi)

ranku (Entry Page) =(Density value of keyword on Target Page /100)+1/Age and

ranku (Any other Page) = Density of keyword on this page/100 + 1/Age value of this page + 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

6. Result and Analysis: There is little Result analysis

6.1 Results of Google's Page Rank Algorithm

Page rank algorithm runs on 10 different problems with Problem Identification Number (PIN) 1 to 10 for each problem to note the ranks of pages. (See Table 1) We set following parameters for Google's Page Rank Algorithm:

Damping Factor=0.85
No. of iterations=10

6.2 Result of Proposed Algorithm (A Heuristic Based AGE Algorithm for Search Engine)

The proposed Algorithm, A Heuristic Based AGE Algorithm for Search Engine, discussed in previous section was implemented and evaluated on the same set of problems used to evaluate Google's Page ranking.(see Table2)

6.2 COMPARISON OF GOOGLE'S PAGE RANK AND HEURISTIC BASED AGE ALGORITHM FOR SEARCH ENGINE (HBAASE)

Results obtained from experiments are analyzed for following factors:

6.2.1 Comparison of Ranks

Comparison of ranks calculated by Google's page rank algorithm and HBAASE for PIN (1) is given in fig. 4.10. Similar results can be obtained for other Problems

6.2.2 Total Ranks Assigned

Comparison of total ranks assigned by Google's page rank and HBAASE for each and every PIN (1-10) is shown in Table 5.3 and in fig. 5.8

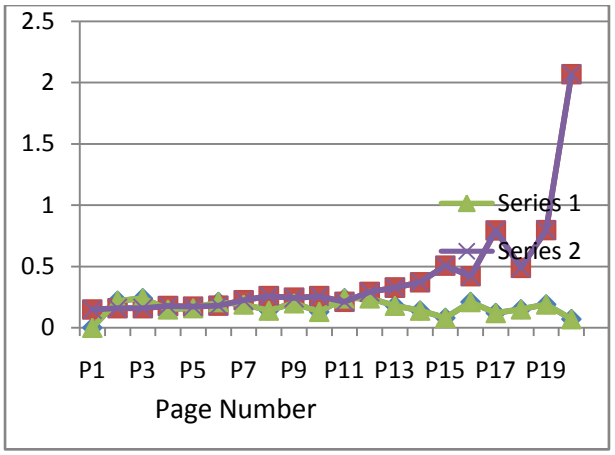

Fig. 4: Comparison of Google's Page Rank and HBAASE for PIN1

Table 3 Comparison of Total ranks assigned to PIN (1-10) by Page Rank and HBAASE.

PIN (1-10)	Google's Page Ranking Algorithm (Total Rank)	HBAASE
1	8.273778	3.26
2	8.371763	5.24
3	7.544906	6.17
4	6.743834	7.91
5	7.345614	10.21
6	7.65544	11.02
7	6.834018	13.94
8	7.874847	16.32
9	7.255741	17.53
10	6.634371	18.41

6.2.3 Google Page Rank Vs HBAASE

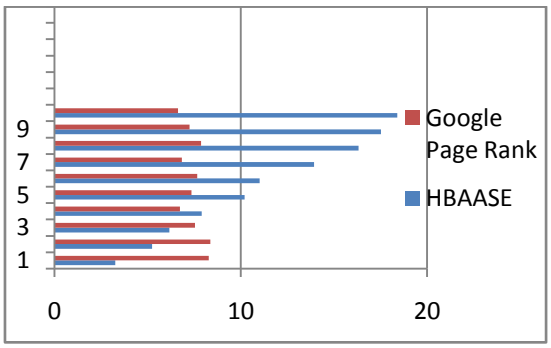

Fig. 5: Comparison of Total Rank assigned by Google's Page Rank and HBAASE for PIN (1-10).

For the 10 PIN, Results comparisons are shown in fig 5. From the Fig.5 It is clear that 70% of the results of HBAASE are better than Google.

Conclusion

In this paper we are representing 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.

REFRENCES

[1]. Sean A. Golliher, "Search Engine Ranking Variables and Algorithms", SEMJ.Org Volume 1, Supplemental Issue, August 2008.

[2]. Christine Churchil, "Search Engine Algorithms and Research" Search Engine Watch, April 2005.

[3]. Junghoo Cho, Hector Garcia-Molina, "The Evolution of the Web and Implications of a Web Crawler".

[4]. Ao-Jan Su, "How to Improve Your Google Ranking:Myths and Reality".

[5]. TIAN Chong, "A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine"

[6]. T.B. Rajashekar, "Effective Web Searching". November 2000.

[7]. Sadegh Kharazmi, Ali Farahmand Nejad, Hassan Abolhassani "Freshness of Web Search Engines: Improving Performance of Web Search Engines Using Data Mining Technique"

[8]. Animesh Tripathy "A Web Mining Architectural Model of Distributed Crawler for Internet Searches Using PageRank Algorithm" 2008

[9]. David Hawking, "Web Search Engines: Part1"

[10]. Mohan Ponnada, Nalin Sharda, "Model of a Semantic Web Search Engine for Multimedia Content Retrieval".

[11]. Animesh Tripathy,Prashanta K Patra "A Web Mining Architectural Model of Distributed Crawler for Internet Searches Using PageRank Algorithm"

Table1: Result of Google's Page Rank Algorithm

Page No.	Problem Identification Number →									
	1	2	3	4	5	6	7	8	9	10
P1	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
P2	0.16	0.15	0.15	0.15	0.15	0.16	0.16	0.16	0.16	0.16
P3	0.16	0.16	0.15	0.16	0.16	0.17	0.17	0.17	0.16	0.15
P4	0.18	0.18	0.15	0.19	0.20	0.18	0.16	0.15	0.17	0.17
P5	0.17	0.20	0.15	0.19	0.19	0.17	0.17	0.19	0.20	0.18
P6	0.18	0.22	0.20	0.15	0.19	0.19	0.17	0.17	0.16	0.20
P7	0.22	0.27	0.17	0.20	0.20	0.18	0.18	0.19	0.21	0.20
P8	0.26	0.23	0.21	0.21	0.19	0.18	0.22	0.21	0.25	0.20
P9	0.24	0.25	0.26	0.17	0.23	0.26	0.24	0.27	0.19	0.23
P10	0.26	0.25	0.26	0.20	0.26	0.23	0.28	0.28	0.22	0.25
P11	0.21	0.25	0.31	0.29	0.26	0.24	0.20	0.21	0.27	0.23
P12	0.29	0.31	0.34	0.29	0.30	0.40	0.23	0.32	0.28	0.32
P13	0.33	0.37	0.42	0.37	0.46	0.31	0.31	0.36	0.25	0.34
P15	0.51	0.42	0.30	0.33	0.33	0.38	0.51	0.48	0.37	0.63
P16	0.42	0.64	0.56	0.42	0.51	0.27	0.52	0.49	0.34	0.45
P17	0.80	0.79	0.45	0.60	0.63	0.75	0.56	0.66	0.67	0.81
P18	0.49	0.50	0.74	0.37	0.91	0.87	0.55	0.75	0.63	0.53
P19	0.80	0.65	1.36	0.92	1.17	1.33	0.86	0.73	0.54	0.53
P20	2.07	2.05	0.84	1.05	0.44	0.85	0.91	1.41	1.62	0.49

Table 2: Result of Proposed Algorithm (A Heuristic Based AGE Algorithm for Search Engine)

Page No. ↓	**Problem Identification Number** →									
	1	2	3	4	5	6	7	8	9	10
P1	0.24	0.41	0.43	0.69	0.54	0.79	0.9	1.18	1.19	1.33
P2	0.22	0.24	0.44	0.7	0.94	1.05	1.18	1.15	1.42	1.44
P3	0.24	0.18	0.12	0.18	0.2	0.36	0.4	0.52	0.59	0.85
P4	0.15	0.18	0.34	0.51	0.48	0.11	0.6	0.79	0.37	0.82
P5	0.16	0.12	0.24	0.32	0.56	0.74	0.84	0.89	0.83	0.94
P6	0.21	0.44	0.56	0.76	0.81	0.81	0.94	1.04	1.2	1.04
P7	0.19	0.36	0.54	0.27	0.42	0.63	0.63	0.87	0.81	0.96
P8	0.14	0.27	0.51	0.09	0.72	0.41	0.64	0.91	1.17	1.29
P9	0.2	0.34	0.32	0.42	0.45	0.48	0.71	0.99	0.62	0.74
P10	0.13	0.22	0.11	0.46	0.59	0.77	0.59	0.85	0.99	1.19
P11	0.24	0.35	0.12	0.52	0.44	0.62	0.79	0.87	0.94	1.08
P12	0.24	0.46	0.45	0.7	0.85	0.7	1.06	1.01	1.07	1.34
P13	0.18	0.24	0.33	0.46	0.22	0.33	0.41	0.57	0.6	0.38
P14	0.14	0.36	0.45	0.67	0.56	0.75	0.85	0.84	0.97	1.07
P15	0.08	0.15	0.23	0.3	0.36	0.29	0.54	0.54	0.68	0.87
P16	0.21	0.32	0.06	0.44	0.31	0.41	0.71	0.57	0.76	0.67
P17	0.12	0.31	0.38	0.23	0.53	0.66	0.68	0.91	1.14	1.19
P18	0.15	0.3	0.37	0.15	0.61	0.69	0.79	0.91	1.17	1.11
P19	0.19	0.22	0.34	0.56	0.73	0.79	0.95	1.16	1.15	0.91
P20	0.07	0.18	0.26	0.17	0.43	0.42	0.63	0.93	1.05	0.52