A Comparative study of Algorithms in SEO & approach for Optimizing the search engine results using Hybrid of Query Recommendation and Document clustering, Genetic algorithm

Ashish Kumar Kushwaha[1] , Prof. Nitin Chopde[2]

[1]ME (CSE), GHRCEM, Amravati
[2]HOD (CSE), GHRCEM, Amravati

Abstract: Promoting a website in search engine result is a major task for a webmaster and SEO Engineer in website development. This become bottleneck for Webmaster or search engine optimization (SEO) engineer to drive visitors to their site by learning technique and algorithm. In this paper, we present a comparative study of various algorithms use by search engine like Page Rank Algorithm, Weighted Page Rank Algorithm, HITS Algorithm, Query Independent Algorithm and proposed an hybrid approach for optimizing the search engine results using artificial intelligence techniques such as document clustering, genetic algorithm and Query Recommendation to provide the user with the most relevant pages to the search query.

Keywords
Document Clustering; Genetic Algorithm, search engine, Query Recommendation

1. INTRODUCTION

Due to the tremendous growth of the Internet in recent years, huge amount of data is added to the World Wide Web, search engines have to perform complex task of sorting billions of pages and displaying only the most convenient and relevant pages for the submitted search query. With this huge amount of data over on web lead to difficulty in managing and displaying data according to end user perspective and become bottleneck for SEO Engineer and Webmaster. It becomes very essential to promote a website in search engine result in website development. Webmaster or search engine optimization engineer have to be actively learning the techniques and algorithms that drive visitors to their site. For this purpose some ordering of webpage is in result list became important. Most relevant page should be place on the top of list and least relevant page should be at bottom according to user query. For this purpose ranking of web page is needed for arranging of page according to user demand dynamically. Page ranking is assigning a value (rank) to the web page among the similar type of page to decide its importance. In this we present some algorithm used in page ranking and their comparison and work proposed aims to optimize the results of a search engine by displaying the more relevant and most user relevant pages on the top of search result list. For this we propose a Hybrid of Query Recommendation and Document clustering, Genetic algorithm. This approach starts with finding most popular query by pre-mining the query logs to fetch the potential clusters of queries and from this all clusters we get most popular queries. Every cluster entries are again mined to obtain sequential patterns of pages accessed by the users. After both mining process, output of both mining is combined to get relevant pages to users with recommendation of popular historical queries. After this document clustering and genetic algorithm is applied resultant output. Document clustering is applied to output to group all similar pages together in one cluster (partition) after genetic algorithm is applied on results to optimize the result and Select the best pages which have highest score depending on other features like number of keywords. At last list of web pages are chosen from different regions of information which are the result of genetic algorithm. This give a optimize list of WebPages for user demand query in a short time.

2. RELATED WORKS

2.1 Page Rank Algorithm

This algorithm developed by SurgeyBrin and Larry Page and is used by Google. This algorithm is based on the link structure of the web. It divides the page rank of a page evenly among its outgoing links. According to this algorithm the page rank of a page can be given by:

$$\mathrm{PR(u)} = (1 - \mathrm{d}) + \mathrm{d} \sum_{\mathrm{v} \in \mathrm{B(u)}} \mathrm{PR(v)}/\mathrm{N_v}$$

Where

PR(u) : page rank of page u, PR(v) : page rank of page v, N(v) : number of outgoing links of page v, B(u) : set of pages that points to u, D : damping factor(the probability of following direct link, usually taken 0.85).

2.2 Weighted PageRank Algorithm

This algorithm is proposed extension to Page Rank algorithm by Xing and Ali Ghorbani[1]. This is also a link based algorithm but it does not divide the page rank evenly. It assigns more page rank to more popular pages. It assigns page rank on the basis of incoming and outgoing links to the page. According to this algorithm page rank of a page is given by:

$$\mathrm{WPR(u)} = (1 - \mathrm{d}) + \mathrm{d} \sum_{\mathrm{v} \in \mathrm{B(u)}} \mathrm{WPR(v)} \mathrm{W}^{\mathrm{in}}_{(\mathrm{v,u})} \mathrm{W}^{\mathrm{out}}_{(\mathrm{v,u})}$$

$$\mathrm{W}^{\mathrm{in}}_{(\mathrm{v,u})} = \frac{\mathrm{I_u}}{\sum_{\mathrm{p} \in \mathrm{R(v)}} \mathrm{I_p}} \mathrm{W}^{\mathrm{out}}_{(\mathrm{v,u})} = \frac{\mathrm{O_u}}{\sum_{\mathrm{p} \in \mathrm{R(v)}} \mathrm{O_p}}$$

Where:$\mathrm{I_u}$ and $\mathrm{I_p}$: number of incoming links to page u and p, $\mathrm{O_u}$ and $\mathrm{O_p}$: number of outgoing links of page u and p.

2.3 HITS Algorithm

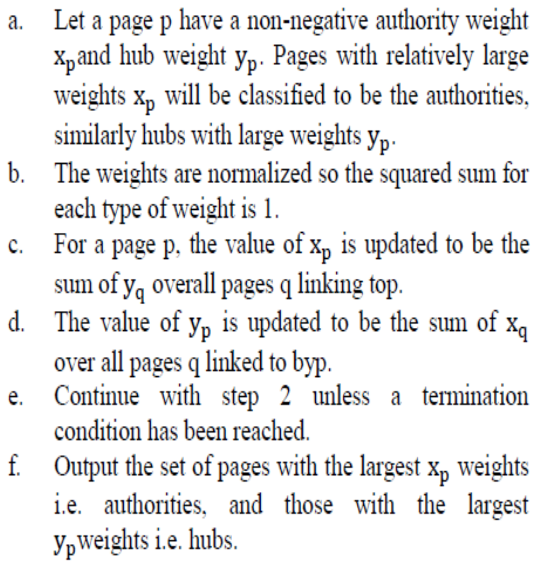
It is called Hyper Induced Topic Search [2]. It is both link based and content based. It considers two types of pages authorities and hub. The former one is a set of pages that are popular and relevant to the query and the later one contains links to useful sites including link to authorities. This algorithm works in two steps: Sampling Step: In this step the page relevant to the user query are collected and a sub graph of the pages is formed. From this sub graph a root set R is taken and algorithm is applies on this root set for expanding it into a base set S by using the algorithm:

Input: Root set R; Output: Base set S Let S =R

a) For each page p E S, do Steps 3 to 5
b) Let T be the set of all pages S points to.
c) Let F be the set of all pages that point to S.
d) Let S = S + T + some or all of F.
e) Delete all links with the same domain name.
f) Return S

I. Iterative Step: Using the output of sampling step that is the base set hub and authorities are identified using the algorithm:

Input: Base set S, Output: A set of hubs and a set of authorities.



a. Let a page p have a non-negative authority weight $x_p$and hub weight y_p. Pages with relatively large weights x_p will be classified to be the authorities, similarly hubs with large weights y_p.

b. The weights are normalized so the squared sum for each type of weight is 1.

c. For a page p, the value of x_p is updated to be the sum of y_q overall pages q linking top.

d. The value of y_p is updated to be the sum of x_q over all pages q linked to byp.

e. Continue with step 2 unless a termination condition has been reached.

f. Output the set of pages with the largest x_p weights i.e. authorities, and those with the largest $y_p$weights i.e. hubs.

2.4 Query Independent Algorithm

It is query and content independent algorithm that assigns a value to every document independent of the query [3]. It is concerned with the static quality of web page. It computes page rank using web graph.

In this algorithm N is the number of documents in the collection, m represents the probability that the random surfer will get bored and restarts from some another random document, “prob” represents the probability transition matrix which is a N*N matrix considering total N pages, “adj” is adjacency matrix and x is probability vector all entities of which are in the interval [0,1]. The algorithm is:

Create a Web Graph

Initialize the probability transition matrix for all I. j€1 to N

Compute a no of out links from a particular node say counter

a) If node having no out link then equally distribute probability otherwise distribute it according to out links
b) For all i,j if(counter==0) then
c) Prob[i][j] =1/N else
d) If(prob[i][j]==1) then
e) Prob[i][j]=1.0/counter
f) Multiply the resulting matrix by 1-m
g) Add m/N to every entity of the resulting matrix, to obtain probability transition matrix
h) For all I,j do prob[i][j]=(prob[i][j]*(1-m))+(m/N)
i) Randomly select a node from 0 to start a walk say s_int.
j) Initialize a random surfer and itr to determine no of iterations required to 0.
k) Try to reach at steady state within 200 iterations otherwise toggling occurs.
l) Multiply probability transition matrix probability vector to get steady state.
m) Check either system enter in steady state or not.
n) Print the ranks stored in probability vector and exit.

2.5 A Relative Comparisons

Table 1. Comparison of algorithms in SEO

	Page Rank	Weighted Page Rank	HITS	Query Independent Algorithm
Description	Divides page rank equally among outgoing pages	Unequal distribution of page rank among outgoing pages based on popularity	Results in highly relevant and important pages	Assigns value to each web page independent of the user query
Based On	Link structure of web	Link structure of web	Link structure and content	Trust level and link structure
Input Parameters	Back links	Back and forward links	Back link, forward link, content	Graph of links
Advantages	Simple, easy to understand, takes O(logn) time	Less complexity than Page Rank i.e.<O(logn)	Gives importance to both structure and content, complexity is< O(logn)	Prioritize the documents on the web independent of the query
Disadvantages	Ignores relative importance, theme shift problem, stresses on old pages	Do not give importance to relevancy	Less efficiency, topic drift problem	Do not consider query during ranking

3. PROPOSED WORK

The proposed Query Recommendation system in a paper [6] learning from historical query logs. This proposed system calculate user's information requirements in a better way by performing query clustering to find the similarities between the two queries, which is based on user query keywords and clicked URLs. After that Generalized Sequential Patterns algorithm is used to generate the frequent sequential pattern of web pages visited by user in each cluster then previously assigned rank score of the web page are modified to re-rank the search result list by using the discovered sequential patterns. The relevancy of the web pages based on its access history is enhanced by rank updation.

After that, the frequent sequential patterns of web pages visited by the users in each cluster are generated with the help of Generalized Sequential Patterns algorithm . The final approach is to re-rank the search result list by modifying the previously assigned rank score of the web pages using the discovered sequential patterns. The rank updation enhances the relevancy of the web pages based on its access history. By this method, the time user spends looking for the required information from search result list can be reduced and the more relevant Web pages can be obtained.

The proposed architecture of Query Recommendation system in paper which consists of following functional components:

- Query Log
- Query Similarity
- Query Clustering Tool
- Sequential Pattern Generator
- Rank Updater

When user writes a query on the interface of search engine, query terms are matched with the index repository of the search engine by query processor and produce a list of matched document. Result optimization system performs its task of gathering user intentions from the query logs in reverse order. Query similarity module continuously analyse the user browsing behaviour as well as the submitted queries and clicked URLs get stored in the logs. The output of which is forwarded to the Query Clustering Tool to create potential groups of queries based on their similarities. Sequential patterns of web pages in every cluster are discovered by Pattern Generator module. Matched documents retrieved by query processor are input to Pattern Generator module. Sequential patterns improve the rank of page which contains search context and the user preference. This improved ranked list is feed to Intelligent Search Engine described in paper[5]. In this first step is Page vectorization in which list from sequential pattern is used to create vector of characteristics for each page. Then this vectorized pages are clustered into similar page called cluster this step is known as page clustering. Finally in third step optimizing is done by applying genetic algorithm on structure identified by the cluster and the score of pages for selecting the best sets of page from each cluster to get most relevant result for user demand query.

4. CONCLUSION AND FUTURE WORK

In this paper, search engine result depend on the various algorithm based on this algorithms web pages are displayed according to their rank which is calculated by using factor like content , number of outgoing link etc. Relative comparison of these algorithms is shown above and proposed hybrid approach of optimizing using query recommendation and Document clustering, genetic algorithms can be useful for search engine to optimize the displaying result and able to display the most relevant WebPages with recommendation to user query so user not have to search through list of displayed page.

In future, query clustering and page clustering will be combined for as a single phase so the time for both clusters will be minimizes and we will able to provide the most relevant in least time.

REFERENCES

[1]. Xing,Ali Ghobrani,"Weighted PageRank Algorithm".IEEE, 2004. [2]. C.Ding, X. He, P. Husbands, H. Zha, H. Simon, "Link Analysis: Hubs And Authorities On The Web" ,2001

[2]. TIANG Chong, "A Kind Of Algorithm For Page Ranking Based On Classified Tree In Search Engine ".IEEE.

[3]. Harmunish Taneja and Richa Gupta." Web Information Retrieval Using Query Independent Page Rank Algorithm". IEEE,978-0-7695-4058- 0,2010.

[4] N.N. Das ,Ela Kumar, Sheetal, "Approaches of Page Ranking Algorithms: Review" 2013

[5] H.M.Zahera, G. F. El Haddy ,A.E. Keshk, "Optimizing Search Engine Result using an Intelligent Model"2012

[6] N.Taneja, R Chaudhary, "Query Recommendation for Optimizing the Search Engine Results" 2012.

[7] Srikant R.and Aggarwal R. Mining Sequential Pattern: Generalizations and Performance Improvements .Proc of 5th International extending database technology, France March, 1996