Optimized Search Engine Recommendation System

Pruthvi Mandaliya
BE-IT
Mob:+91-9967300204

Suraj Khot
BE-IT
Mob:+91-8369662897

Swapnil Mohite
BE-IT
Mob:+91-8451815070

Rakesh Varma
BE-IT
Mob:+91-9987882841

***Abstract*—This system is designed to provide optimal search result to the user.Our idea is to build search engine which considers certain factors to provide user search result. We introduce an algorithm to index the web pages using two concept based namely web page hyperlink analysis using page ranking and weightage to terms in page for classification.Our system also provide recommandation to the user based on his area of interest and past history of the user.This system ba-sically build to provide search result based on query entered by user in a search box.This system satisfy user requirement by providing optimal search result.**

Keywords—Search Engine, Hadoop, Recommendation, Big-Data, PageRank, BM-25, TF-IDF, K-nearest neighbour.

I. Introduction

This system is basically design to provide optimized search result to the user which will satisfy user requirement.We are using pagerank and classification algorithm to get optimal search result according to user query.This system is developed to provide search result which has higher accuracy by considering factors which necessary in calculating page rank of the web pages.Our system also provide context based search and handle link spam to provide more accu-rate search result.

A. PageRank:

PageRank is a vote, by all the other pages on the Web, about how im-portant a page is. A link to a page counts as a vote of support.
Formula: $PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$

B. Term Frequency - Inverse Document Frequency :

Term Frequency: Which measures how frequently a term occurs in a document.
Formula: $TF(t)$ = (Number of times term t appears in a document) / (Total number of terms in the document).

Inverse Document Frequency: Which measures how important a term is.
Formula: $IDF(t) = \log e$(Total number of documents / Number of documents with term t in it).

C. BM25 Algorithm:

BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document (e.g., their relative proximity)

D. K Nearest Neighbors:

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance func-tions).

II. Conclusion

This search engine system consists of page rank algorithm,classification and indexing of the web pages.It uses KNN classification algorithm to classify the web pages and then page rank of the pages related to the user query is calculated so that it will provide search result according to page rank value of the web pages.We are implementing the page rank algorithm in a such way that it will provide optimal search result to the user.The context based search result can also be provided to enhance the accuracy of search result.This system is made to provide search result to the user which will satisfy the requirement of the user.

Acknowledgment

I would like to express my special gratitude and thanks to our guide for giv-ing me such attention and time. My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities

References

[1] Pooja gupta Mrs. Kalpana Johari,IMPLEMENTATION OF WEB. CRAWLER, in Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09.

[2] TIAN Chong, A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine, International Conference on Computer Application and System Modeling(ICCASM 2010).

[3] Debashis Hati,Biswajit Sahoo,Amritesh Kumar, Adaptive Focused Crawling Based on Link Analysis, 2nd International Conference on Education Technology and Computer(ICETC), 2010.

[4] Fan Chung, A Brief Survey of PageRank Algorithms, IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, 2014.

[5] Wang Hui-chang,Ruan Shu-hua,Tang Qijie, The Implementation of a Web Crawler URL Filter Algorithm Based on Caching, Second International Workshop on Computer Science and Engineering, 2009.

[6] Chia-Chen Yen,Jih-Shih Hsu, Pagerank Algorithm Improvement by Page Relevance Measurement, FUZZ-IEEE 2009.

[7] Radha Shankarmani,M.Vijayalakshmi, Big data Analytics.