International Journal of Advanced Science and Technology
Vol. 29, No. 9s, (2020), pp. 3102-3107

# An Efficient Big Data Analytics based Cloud System for Optimizing Web Page Discovery Techniques

S.Dhanasekaran[1], P.Vijayakarthik[2], A.Sivanesh kumar[3], B.S.Murugan[4], V.Vasudevan[5]
*[1,4,5] Department of Computer Science and Engineering*
*Kalasalingam Academy of Research and Education*
*(Deemed To Be University)*
*Srivilliputtur, Tamilnadu, India*
*[2]Dept of Information Science and Engineering*
*Sir.M.Visvesvaraya Institute of Technology, Bangalore*
*[3]Department of Computer science and Engineering*
*Saveetha School of Engineering,*
*Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India.*
*srividhans@gmail.com[1], vijaykarthik_is@sirmvit.edu[2],*
*kas.sivanesh@gmail.com[3],muruganbs@gmail.com[4],vasudevan@yahoo.co.in[5]*

***Abstract***

*In recent days Internet plays vital role for many purpose such as sharing of information, Academic related activities, searching about meticulous topics and also for the customer entertainment. This system is planned to develop a search engine to assist the clients for discovering relevant web page from large amount of database. This web page searching system provides most relevant top ten results to the clients. By using this type of searching techniques user have to easily find the suitable information that they are looking for. In this research work big data analytics based navigation strategy is introduced in addition to cloud based computing. With the help of cloud computing technique the search engine can be able collect more information and at the same time information can be able to share to any part of the world. The big data analytics concept will be used to store infinite information in database. This will assist the client to discover most related search item.*

***Keywords****: Cloud Computing, Crawler, Indexing, Ranking, Search engine, Information Retrieval.*

## 1. Introduction

Cloud computing techniques mostly used to develop Internet based Applications and also utilize the computing technology for dynamic scalability and virtualization. Here virtual resources are distributed as services through World Wide Web consortium. People who use the cloud computing technique will not depend on their own computer alone. They depend on third party computer who consisting of mass storage and data available in the system. Cloud users generally have to search for appropriate Cloud based service manually.

Basically cloud based search engine system includes crawler, indexer and ranking mechanism. These are works to gather to find the relevant documents, images, data , information and videos and so on. The web pages are displayed on the screen as per the ranking principles of cloud system. The highest priority webpage are displayed in the top of the search results. Efficient and Effective cloud based search engines are able to extract suitable cloud information's in the internet.

## 2. Related Work

### Google

The Google is the very popular search engine known to all of us. The Estimated unique monthly visitors to this search engine are 1.6 billion. The Alexa Rank is 1. Google is the search


Copyright ⓒ 2020 SERSC

engine that is being used by most of the people in our society. Google is being able to be done by using Webpage rank algorithm. It consisting of vast amount of keywords so it can be able to do fulfil the user needs. In this Google search engine it also consisting of Google map, Gmail, gplus..,

**Bing**

The Bing is traditional searching system This search is being known by us. The Estimated unique monthly visitor to this search engine is 400 million. The Alexa rank is 22. Bing is very efficient for the video files to the process in it. While comparing to the other search engine to search the videos it is very efficient. It is often gives the auto complete suggestion while user do the browsing.

**Yahoo**

The Yahoo is the power full searching system. The Estimated unique monthly visitors to this search engine are 300 million. The Alexa Rank is n/a. Yahoo is a search engine that is being able to be operate independently. Yahoo news,finance and sports platform being added to AOL's media assets.

**Ask.com**

Its based on the questions and answering system. This search engine most probably works on the basis only to answering to the users. It also consisting of general search functionality but the result returned lack quality.

**Baidu**

This Baidu search engine is being more familiar in china. Baidu is serving billion of search queries per month. The rank position of Baidu is 4; while these searches do answering it can be able to be replied only on the basis of the user being asked to them.

## 3. Issues Identified in Existing System

**Lack of links**

Lack of links are the major issue that is being happened in the most of the search engine. While the user doing the search in one site simultaneously continue link building is failed in most of the search engines.Many people still don'tunderstand the importance of the continual building.

**Repetitive title tags**

Some time some of the pages are repeated. This is another common issue that is being happened in the search engines. While using search engines we must be clear to check that every Webpage that is being available in our site should consist of unique title tag.

**Too many 404 errors**

The customer has to verify the errors. We can be able to spider our Webpage and check to see if you have errors in our Webpage. A shortcut to perform in Google for site: Webpage.com and it will result all our Webpages.

**Too many 301's**

This scenario can be incurred during when the user needed to redesign the site and initial SEO audits;making it increasingly importance to make sure and do proper SEO and the keyword is being reached initially, to avoid having to redo this later.


Copyright ⓒ 2020 SERSC

**Bad links to your home Webpage**

The unwanted sites are being displayed due to in correct prediction. In some browser in the home Webpage of it will be displayed with unwanted sites and unwanted advertisements. Some type of thing will be able to be irritating the users.

## 4. Proposed Work

This research work have included the crawler, index and ranking techniques. As we have analysed in the most of the search engine the browser will not give the user expectation alone. Thinking that everything is being provided the data that is being unrelevent to also be given the user who the searching. By downloading the search engine to the Systems the efficiency of software which means ram .We consider these two issue and we can be overcome it. By using our browser the ram consumption will be less and it will be able to show only relevant data that the user being asked to it.

**Crawling Techniques**: The crawler is acting as bot or spider in search engine. that travels all around the web looking out for new WebPages ready to be indexed.

**Indexing Process**: After the Search Engines crawls the web and comes across the new WebPages, it then indexes or stores the information in its giant database categorically, to be retrieved later when any search query related to it comes up.

**Providing information**: Its provide relevant information or acting as answering machines.

Google searching system uses automated program called spider or crawler. In Google search engine consist a collection of keywords which we can use for searching data. Since it consisting of more amount keyword it can be visible to the user search that the data needed and irrelevant data also visible in such type of search engine.

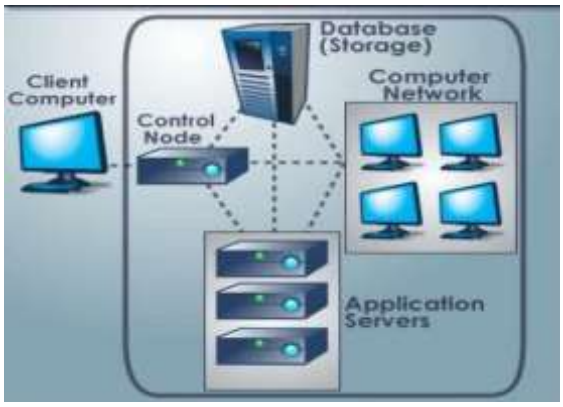


Fig: 1.1


Copyright ⓒ 2020 SERSC

**Algorithm for Webpage Rank:**

1: $AB_0^G(u) := 1-\alpha/n$

2: $layer_0 = \{u\}$

3: $info(u, u) := 1$

4: **for** $p = 1, \ldots, q$ **do**

5: $Layer_p :=$ Get all in-neighbours of nodes in $layer_{p-1}$

6: **for** each $v \in layer_p$ **do**

7: $inf_p(v, u) := 1/outdeg(v) \sum_{w \in layer_{p-1}, v->w} inf_{p-1}(w,u)$

8: **end for**

9: $AB_t^G(u) := AB_{p-1}^G(u) + 1-\alpha/n \sum_{v \in layer_p} \alpha^p inf_p(v, u)$

10: **end for**

11: **return** $AB_q^G(u)$

Here we have to use Webpage Rank algorithm. A small universe of four web WebPages: P, Q, R and S. If all those WebPages link to P, then the Webpage Rank of Webpage P be the sum of the AB of WebPages Q, R and S.

*AB(P) = AB (Q)+ AB(CR)+ AB (S)*

But then suppose Webpage Q also has a link to Webpage R, and Webpage has links to all three WebPages. One cannot vote twice, and for that reason it is considered that Webpage Q has given half a vote to each. In the same logic, only one third of S's vote is counted for P's Web Page Rank.

*AB(P) = AB(Q)/2 + AB(R)/1 + AB(S)/3*

Divide the PR by the total number of links that come from the Webpage.

*AB(P) = AB(Q)/L(q) + AB(R)/L(R) + AB(S)/L(S)*

The Web Page Rank value of a Webpage reflects the frequency of hits on that Webpage by the random surfer. So the equation it follows:

$PageRank(u_i) = v + (1-v) \sum PageRank(u_j)$

$u_j \in L(u_j)$


Copyright ⓒ 2020 SERSC

## 5. Navigation Techniques

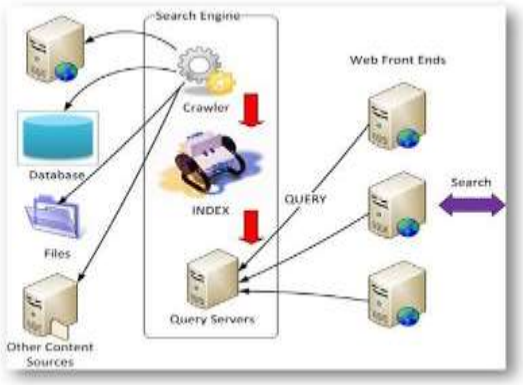


Fig1.2

The crawler is connected with large database and the files that are being needed are connected to the crawler. The crawler will store the data that is available in it. After collecting the data's the next process will be able to be sending it to the index and then index will send those things to the query server. The user query will be able to them through the web front end. For the use the web frontend alone visible and other the process cannot be able to be viewed by the users.

## 6. Result and Performance Evaluation:

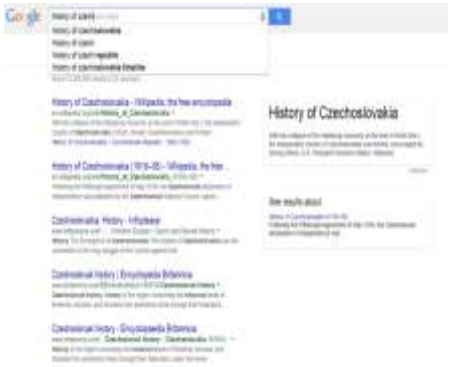


Fig: 1.3

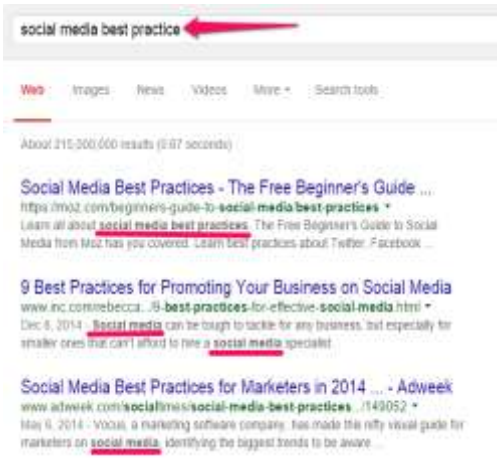


Fig: 1.4


Copyright ⓒ 2020 SERSC

Hence the system have to planned for the browser and the browser only needed to give the thing that is being asked to it. In such a way it can be able to provide only the relevant data needed to the user.

## 7. Conclusion and Future work

This research work has found the relevant web page based on the user intention and requirements. By using this system, the cloud search engine has added with home page for helping the people to choose most relevant web pages instantly. We concluded with the search engine that is being created and in future the system will be updated with Access token to enforce the authentication. This enhanced technique will give second layer of authentication for added protection against hackers. The client account will be enabled by two step verification.

## 8. References

1. K. M. Sim, "Towards complex concession for cloud economy," *in Proc.5th Int. Conf. Grid Pervasive Comput., vol. 6104, LCNS, pp. 395-406.*
2. K. M. Sim, "Agent-based cloud commerce," in Proc. *IEEE Int. Conf.Ind.Eng. Eng. Manage, Hong Kong, 2009, pp. 717-721.*
3. S.Dhanasekaran, Dr.V.Vasudevan, "A Dynamic Multi-Intelligent Agent System for Enhancing the Cloud Service Negotiation", *International Journal of Applied Engineering Research,* vol. 10, no. 43, pp. 30469-30473, 2015.
4. Dhanasekaran.S and Vasudevan.V., A Smart Logical Multi agent System for Consolidating Suitable Cloud Services, *International Journal of Computer Science and Information Security,***14** (9) (2016), 517-522.
5. K. M. Sim, "Complex and concurrent Negotiation for multiple interrelated e-markets (Position paper)," in *Proc. Group Decision Negotiation Conf. Del*, The Netherlands, pp. 253–264.
6. Dhanasekaran.S and Vasudevan.V.Rational Agent Based Multiple Concurrent and Complex Concession for Service Composition and Discovery, *IEEE Xplore Digital Library,* (2016), 2797-2801.
7. Y.-S. Chang, T.-Y. Juang, C.-H. Chang and J.-S. Yen, "Integrating intelligent agent and ontology for services discovery on cloud environment," In: Systems, Man, and Cybernetics (SMC), IEEE International Conference on, 2012, pp. 3215-3220.
8. Dhanasekaran.S and Vasudevan.V., A Cognizant agent system for optimizing cloud service searching strategy, *The Journal of Networks, Software Tools and Applications: Cluster Computing, Springer,* **20** (78) (2018) , ISSN: 1386-7857 (Print) 1573-7543 (Online).
9. Dhanasekaran.S and Vasudevan.V. Multiple Intelligent Agent Coordination Strategy for Categorizing and Searching Appropriate Cloud Services, IEEE Xplore Digital Library, (2018), 387-391.
10. J. Kang and K. M. Sim, "A cloud portal with a cloud service search engine," in International Conference on Information and Intelligent Computing (ICIIC), Hong Kong, China, Nov. 2011


Copyright ⓒ 2020 SERSC