Search Engine Optimization on Big Data

J.Satish Babu, Jakkala Pranay Reddy, Dantu Sri Vishnu Sai Lohit, Ravanala Chandu, G. Krishna Mohan

Abstract*: The term big data was invented to capture the means of the rising trend in the volume of knowledge conjointly exhibits the improved characteristics as compared with ancient knowledge management and analytics of big data is important for achieving scientific and engineering breakthroughs, mining for timely and relevant info. Two potential solutions are to style a replacement real-time operation model or a knowledge analysis mechanism. During this paper we tend to primarily specialize in data transmission, data acquisition, data storage, and data analytics. We tend to aiming to solve these issues in search engine optimization (SEO). Search engine Optimization (SEO) is characterized as a gathering of techniques and practices that permit a site to get more activity from internet searchers also, it is still one of the biggest challenges in search engines of Semantic webs. This paper proposes another sort of web page search which depends on the competitive intelligence. It use link based ranking evolutionary scheme to suit clients' preferences.

***Index Terms*: Big data, Data transmission, Data acquisition, Data Storage, Data analytics, Search engine optimization (SEO).**

I. INTRODUCTION

Worldwide data is very complex to segregate and to manage this data we will be following the new approach called big data. In this paper we will be undergoing through various techniques to optimize the searching procedure in the searching procedure. This search engine has various process and procedure to crawling, indexing, processing, calculate relevancy, recovering results. Even through semantic web and various searching strategies. Case in point, an IDC report predicts that, from 2005 to2020, the worldwide information volume will develop by a component of 300, from 130 Exabyte's to 40,000 Exabyte's, speaking to a twofold development at regular intervals. The term of ``big data" was begat to catch the significant which means of this information blast pattern and without a doubt the information has been touted as the new oil, which is required to change our general public.

For instance, a Mckinsey report states that the potential estimation of worldwide individual area information is assessed to be $100 billion in income to administration suppliers throughout the following ten years and be as much as$700 billion in worth to customer and business end clients. The enormous potential connected with big data has prompted a developing research that has immediately pulled in great enthusiasm from various areas, for instance, industry, government and examination group. The expansive hobby is exampled by scope on both mechanical reports and open media (e.g. The Economist the New York Times, and the National Public Radio (NPR), .Government has additionally assumed a noteworthy part in making new projects to quicken the advancement of handling the big data challenges. At last, nature and Science magazines have distributed exceptional issues to talk about the big information marvel and its difficulties, extending its effect past innovative [1].

The development of datasets of monstrous size, differing qualities and rates, termed "Big Data", is quickened by high-throughput experimental instruments, and portable and online sensors installed in our day by day lives. Administration and examination of big data is basic for accomplishing investigative and building leaps forward, digging for opportune and apropos data, and choice making. The capability of big Data can be interpreted into reality just through improvement of novel calculations, compelling programming stages to explore information, and imaginative utilization of equipment foundation to scale them. Big Data applications should be bolstered on HPC frameworks as well as on developing digital foundation, for example, Cloud stages, and quickening agents like GPGPUs, FPGAs and numerous centre processors. The union of Big Data programming stages and quickened digital framework is crucial for transformative examination. For this uncommon issue, we welcome articles on imaginative exploration to address Big Data difficulties utilizing novel calculations, programming architectures, rising processing stages, and one of kind methodologies. Entries that relate to Big Data examination in any field are pertinent to this exceptional issue[1].

II. LITERATURE SURVEY

Yonggang Wen has published a paper named "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial" – His thought was to give a gist for the readers who are non experts and indulge a spirit for advanced developers to modify their own big-data results. Initially he presents the brief on the big data and challenges around it.Further he presented the big data framework called the storage, generation, analytics, acquisition.He also presented the Hadoop Frame work for solving the challenges in big data [1].

Revised Manuscript Received on May 07, 2019.

J.Satish Babu, Department of Computer Science and Engineering,KoneruLakshmaiah Education Foundation, Vaddeswaram,A.P.,India.

Jakkala Pranay Reddy, Department of Computer Science and Engineering,KoneruLakshmaiah Education Foundation, Vaddeswaram,A.P.,India.

Dantu Sri Vishnu Sai Lohit, Department of Computer Science and Engineering,KoneruLakshmaiah Education Foundation, Vaddeswaram,A.P.,India.

Ravanala Chandu, Department of Computer Science and Engineering,KoneruLakshmaiah Education Foundation, Vaddeswaram,A.P.,India.

G Krishna Mohan, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram,A.P.,India.

Ma Xiao ling has published a paper named "For research of search engine optimization (SEO) " he has presented SEO analysis that has increase web site ranking and therefore to fetch the web site traffic. Hence this paper is often used as basis for SEO engineers [2].

Graf, Cousin has published a paper named"SEO TOOLSET" Here what the paper presents is the SEO toolset which guides the web developer to understand and analyze the page for searching thekeywords and related contents of page that will contribute to SEO [3].

III. DESCRIPTION

A. SEARCH ENGINE OPTIMIZATION

Webpage improvement (SEO) is the technique of extending site's dimension of introduction to individuals when all is said in done. It suggests that it encourages spectators to find the site. Web optimization is a method in which webpage achieves solid, high rankings of pages of Search motor for most of the crucial significant words. This should be conceivable by completing a couple of changes which makes the site look benevolent. In case the website appears at top position of the web searcher, the more visitors it gets.

Web optimization basically redesigns the page code, the site page content by executing its own specific relevant calculation and as such it helps the web index to bring the page. So SEO is one of the techniques for bringing the web action. Anyway there are also a couple of various advances to meet the target like Pay per Click (PPC) model in which supports pay their advertiser exactly when their ads get clicked by the visitors. So supports offer on the particular significant word look.

B. WORKING OF SEARCH ENGINE

The is searching uses the indexing such that the specified string is searched through the crawler language , so that the result is achieved in an efficient manner. Initially the string is searched in the database. In this technique we need a programming which are known as robots or arachnids. These are customized to discover either the new record or refreshed reports[8].later on this produces the new reports. It does the searching in the database of web search tool[7].these words are taken in the crept websites. It is a area where they are recorded and their coding. Hence the clients enters the web search tools and the crawlers search in the database or the recent cache by examining its rundown. Web indexes play out a few exercises so as to convey query items.

- Crawling – crawler is all about the process of bringing pages to a site. this crawling uses crawler language. This is also known as arachnid.
- Indexing – For a specific catch phrases, doling out the page is done. This portrays the best paging. Indexing is all about bringing all site pages and keep them into a major databases. Such that these pages can be recovered later.
- Processing – The search tool uses pursuit string demand with the listed pages such that the inquiry demands comes ,these listed pages in the database.
- Calculating Relevancy –when a string is searched , the searching starts in an ascending order. So many pages are likely to get more pages.th significance of pages in its list to the inquiry string.
- Recovering Results–These crawler gets the efficient result by recovering the best coordinated outcomes. Hence it is all about the showing the, in program.

C. IMPORTANCE OF SEO

The primary significance of SEO is in web promoting on the grounds that large portions of the general population don't go past 20-30 website pages while seeking on the internet searcher. On the off chance that the client has aim to purchase a few items on the web page and if your site is not recorded on top of the web index comes about then its high likelihood that u would lose your business. They essentially visit some other site. So SEO serves to accomplish the positioning and expands the shots of onlookers going to your site.[13]

D. SEARCH ENGINE OPTIMIZATION IN SEMANTIC WEB (SEO)

This SEO Web is a worthy searching process of the World Wide Web to make a metadata web of benefits that can predict themselves not simply by how they should be appeared or semantically ,also by the centrality of the metadata. The important reason for the SEO is enabling so as to make the improvement of the current Web users to find, go halves in, and gather data with at most efficiency. A search engine can't perform searching assignments without human involvement, in light of the fact that site pages are proposed to be examined by people, not machines. The semantic web is a fantasy of information that can be expeditiously interpreted by machines, so machines can play out a more prominent measure of the dull work incorporated into finding, merging, and following up on information on the web. Search engine optimization [9], it is a main concept in Semantic Webs and it suggests to the amassing of methodology and practices that allows a site to get greater development from web searching tools. The web crawlers generally are adequately quick to concede you that rank as is normally done. Web indexing have ended up being more standard on the web, nearly anyone endeavoring to get seen on the web can benefit by a little SEO adoring. Folksonomy[10] is the techniques of Web 2.0 which is expanded one to Semantic Web page positioning is also an important technique in web seeking methodologies.

E. Page Rank - PR (E)

Page Rank is a process of calculating and assigning ranks to sites in the results of the search engine. It works on the basis of number and also the nature of the connection with a particular page to know how important the site is. Widely used sites get more associations from various sites. The weight that it allots to some random segment E is implied as the Page Rank of E and implied by PR (E). The Page Rank is the probability of getting in contact at that page after innumerable. This will be equal to t-l where t is the quantity of pictures. One basic problem of Page Rank is it gives more importance to particular pages. Other page site, though having a very good quality, don't have many associations unless it is a part in the ongoing main site [2].

Page Rank α 1/ (The Number of clicks)

The Page Rank has a probability appropriation in and around 0 and 1 as the prior quality for all the pages. Page Rank characteristics for each page v maintained in the set Bu (the set containing all connected pages of u), separated by the count L(v) of associations from a page v. In the end, the Page Rank is equal to the stores and notes down the own Page Rank score separated by the number of outbound associations L [2].

$$PR(u) = \sum VEBu \frac{PR(v)}{L(v)} \text{ --- (1)}$$

Damping variable is described as the probability, that the separate individual will move on is a damping element d. The damping element is maintained in and around 0.85. It is taken out from 1 and result is added to the result of the damping element and the complete of the upcoming Page Rank scores. The result is then separated by the number of records (N):

$$PR(U)=((1-d)/n)+ d\sum_{vebu}(PR(v))/L(v) – (2)$$

Simplified Page Rank of the current web site I is taken as:

$$PR_i= ((1-d)/n) + d\sum_{j=1}^{n} (PR_j)/(L_{ji}) \text{ --(3)}$$

Here Lj, is the quantity of outbound associations from page, when page is connected to page i. We have noted that we have to take the past behavior into account of every user. We have calculated Lji by divide the count of the times the page was accessed to page i when is connected to I [2].

$$C_{ji}= \sum_{u\ e\ U} C_{ji}\ (u) \text{ --- (4)}$$

$$L_{ji}= (\sum_{j=1}^{n} C_{ji})/ C_{ji} \text{ --- (5)}$$

In which Cji(u) is the count on times the page was accessed by a user u from page to i, while U is group of the users. To have an account on the User's choice to support personalized search, it is must to the count on times the page was accessed that have associated to page i on the count on times the page was accessed that have connected to all the pages. Equation (6) says its definition.

$$PE_i(u)= (\sum_{j=1}^{n} C_{ji}(u))/(\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ji}(u)) \text{ ---(6)}$$

Finally, Equation (7) is my improved Page Rank algorithm for the proposed system.

$$PR_i= PR(u_i)=(1-d)PE_i+d\sum_{j=1}^{n}(PR_j)/(L_{ji}) \text{ ---(7)}$$

F. SEMANTIC WEB FOLKSONOMY STRATEGY

Folksonomy is another order procedure which cooperatively makes and oversees labels to sort substance. Folksonomy may hold the way to building up a Semantic Web, in which each Web page contains machine-intelligible metadata that portrays its substance. Such metadata would significantly enhance the accuracy (the rate of important reports) in web search tool recovery records. It join labels or marks to every website page to suffice the practice and system for sorting substance. "Labels" are watchwords that dispensed by clients to every page uninhibitedly and subjectively, taking into account their importance. Anybody can pick any word as label and can put different labels to one page[3].

Fig2 demonstrates a case of labeling site pages by diverse clients. We characterize the "recurrence" of a tag for a page as the quantity of clients who utilized the tag for the page, and decide the class of a page utilizing the recurrence of every label doled out to the page: the tag with the biggest recurrence turns into the classification of the page. The most essential point of preference of folksonomy is that clients can rapidly hunt and effortlessly order related website pages. It is surely understood that folksonomy gives a level, non-various leveled and shared phrasing. Fig3.1 demonstrates the folksonomy labeling chart.

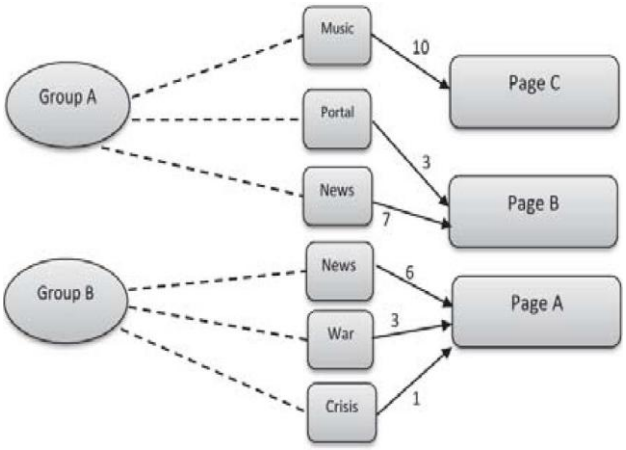

Fig. 1 Folksonomy Tagging graph

G. COMPETITIVEINTELLIGENCE

Imperialist Competitive Algorithm (ICA) is widely accepted web searching methodology which was brought to the world for handling with many of the optimization works[12]. This algorithm starts with an initial population where each individual of the population is named as country. Country is a set that says with the same idea of chromosomes in the Genetic algorithm methodology and is described in equation 8:

$$Country= (p1, p2, p3,\ldots, p_{Nvar}) \text{ ---(8)}$$

The Price of a country is examined by calculating of the function f which is the similar to metric function in genetic algorithm, and is described by equation 9:

$$Cost=f(Country)=f(p1,p2, p3,\ldots, p_{Nvar}) \text{ ---(9)}$$

Small amount of the countries are selected to be the states and the remaining will get into the streets of these imperialistic states. All the states of beginning countries are separated from the mentioned imperialistic states depending on their power. The imperialist states adding on with their streets form kingdoms (empires). After forming basic kingdoms, the streets in every of them start moving to their imperialist state. Here Policy is designed by navigating all the respective streets to the imperialist states. The entire strength of a kingdom will stand on both the strength of the country and the strength of its streets.

This idea is shaped by defining the total strength of a kingdom as the strength of country plus a percentage of average strength of its streets.

Opposition can be said as a drastic difference in socio-political nature of a street that is, rather than being acclimatized by a settler, the street changes its place in the socio-political axis. Moving to the settler, a word is made and that may reach to the place with low cost than settler. At that place the evaluation will proceed by the settler in the new place and the states will be accumulated by the settler in its place. In the improvement of places toward the world of less issues a few settlers may navigate to respective places. The chance that the partition between two settlers will change out to be no limit partition, they combine and change into other domain. Each of the understanding of two new areas change into the areas of the new kingdom and the new colonist will be in the place of one of the two settlers. Imperialistic competition is a process that gets the force of weak areas and all kingdoms to take the rights of settlements of many kingdoms and control them. It is shown by taking the percentage of weak agreement of weak kingdom and makes an opposite colonist in rest of all areas to have these imperialistic states[3].

H. PROPOSED ARCHITECTURE AND REDEFINED ICA

SEO suggest to the collection of systems and processes that allow a page to get much motion from web searching apparatus. To increase the SEO use ICA[11]; at prior we require in quantity of web mining evaluation to say the stay a reason the grounds that we can't think about each extraordinary word as a catch, again a catch is made of other words for the term Internet Consortium can be separated into two more terms Internet and Consortium in fact of the Interne tis a known term.

I. OVERALL ARCHITECTURE

Identifying and sketching are the known and vital pages to requestors is our needed concern. We thought of using other page searching algorithm in shade of ICA joins folksonomy and connection association based positioning technique.

- Realm Instatement Layer: It has folksonomy data storage and Site ranking database. They have similar information come from management layer.
- Information transport layer: Show bulk space of web page information.
- Application layer: It has a search engine and Quality Assurance engine. Quality Assurance Engine is program that can give solutions from an unstructured set of natural language data records. This layer is incharge of preparing the queries of clients and giving back the query output.
- Management layer: It has ICA head, ICA head is used to understand and distribute bulk data, all parts of ICA (Except for beginning of realm; which is done in realm instatement layer) should be done in this layer.

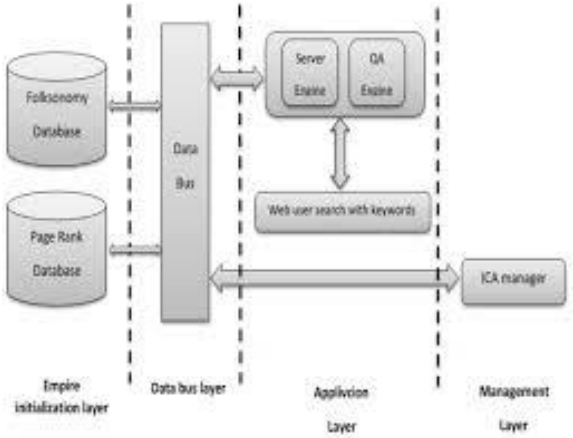

Fig. 2 ICA_based Search Engine architecture.

IV. EXPERIMENTAL RESULTS

Based on the space of the page connected to one particular central node page and also by the observations what we have observed are the page ranks for sample pages of P1 to P20.These Calculations are made by considering the central page ranks as a normal numbers varying between 0.1 to 0.2 and for the first 20 pages we considered a link relationship with next 10 pages, here thereby the only pages from 1 to 20 are being involved. We get the page ranks scores of pages P1 to P20 using the Page Rank algorithm. Table 1 shows the resulted calculations of page ranks.

Table. 1 Resulted Page ranks for pages from 1 to 20

Page	Page Rank	Page	Page Rank
P1	0.08599	P11	0.08991
P2	0.09998	P12	0.09590
P3	0.09321	P13	0.09498
P4	0.09775	P14	0.06693
P5	0.08898	P15	0.09597
P6	0.07598	P16	0.07798
P7	0.09004	P17	0.09596
P8	0.09590	P18	0.08898
P9	0.09693	P19	0.09594
P10	0.09448	P20	0.09397

Here the above figure 4 indicates the sample space where we each single point in the graph represents the page rank of that particular page.

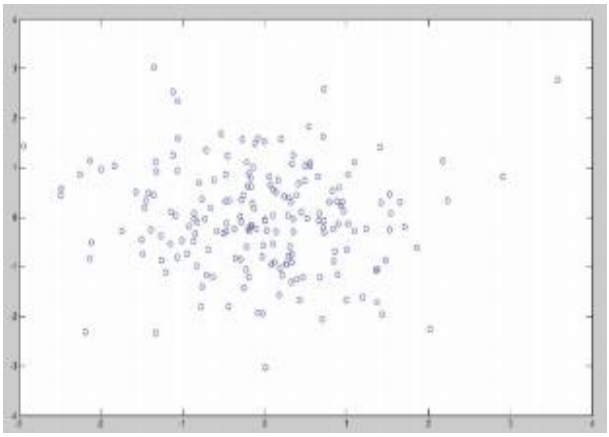

Fig. 4 Sample Space of random page ranks.

Here the above figures 5 and 6 indicates the page rank of pages at each iteration where at each iteration we the page chooses the central page which is closest at its distance and finally it reaches the best closest central page.

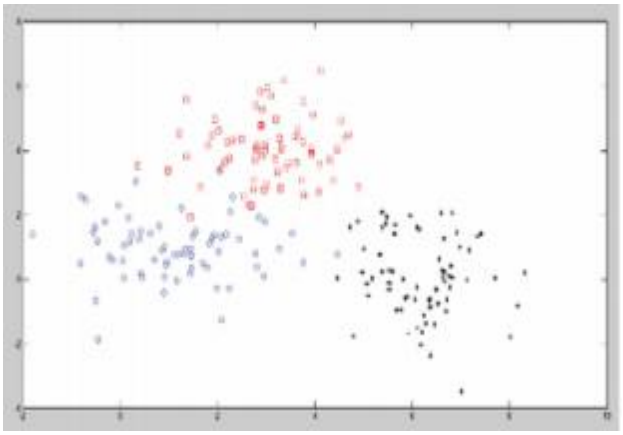

Fig. 5 Page ranks for different iterations

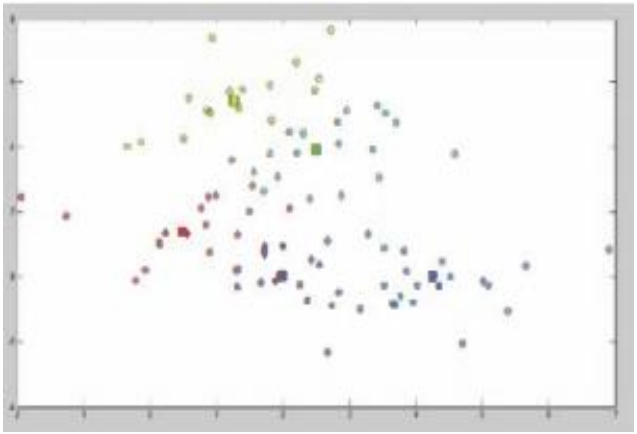

Fig. 6 Group Of page ranks per different iterations

The above figure 7 is a graph that represents the cost of the each page for each iteration.

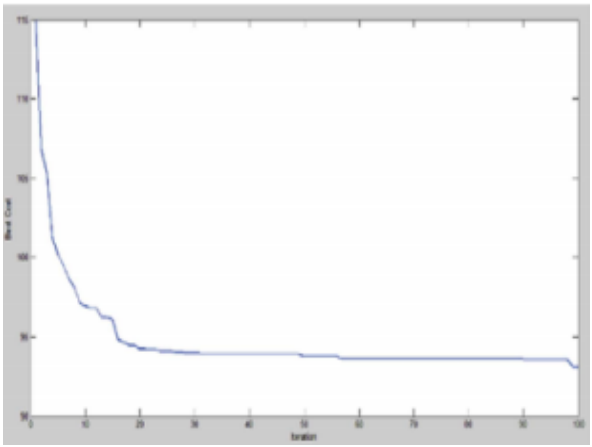

Fig. 7 Cost of page per iteration.

V. CONCLUSION

We can conclude that drawback presented in the page ranking algorithm we will add some more algorithms to overcome the problem so web gets the required results dynamically. For future individuals can take every necessary step in site design improvement utilizing pictures, connect, structure based should be possible to create the outcomes quick and identified with the web surfer. The proposed work is added with ICA algorithm and associated connection used positioning. Our goal is to redistribute the absorption and revolt with the objective that they can be impeccable with broad scale look for in semantic systems."Improving the running process" of searching and analysing and applying the found information on big scale information, combinational enhancements is our other goal in this paper. Future work is much more needed to make the current work fit to present World Wide Web.

REFERENCES

1. Han Hu, Yonggang Wen, Tat-Seng Chua and Xuelong Li," Toward Scalable Systems for Big Data Analytics: A Technology Tutorial" 2014 IEEE.
2. Pagerankalogorithmhttp://www.linksandlaw.com/technicalbackground-pagerank.htm.
3. "A new Competitive Intelligence-based strategy for Web Page Search"ImanRasekh, PhD Candidate Institute of computer science, University of Philippines at Los-Banos, Los-Banos, Laguna, Philippinesiman.rasekh@gmail.com
4. Monika Yadav, Mr. Pradeep Mittal," Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.
5. Search Engine Optimization: An Hour A Day- 3Rd Ed by Jennifer Grappone and GradivaCouzin.(text book)
6. SEO 2015 &Beyond :: Search engine optimization will never be the same again (Webmaster Series) [Kindle Edition] by Dr. Andy Williams.(text book)
7. Google. Google's Search Engine Optimization Starter Guide.PDF, November 2008.URLhttp://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf
8. Albert Bifet and Carlos Castillo. An Analysis of Factors Usedin Search Engine Ranking. In Proceedings of the 14thInternational World Wide Web Conference (WWW2005), First International Workshop on Adversarial Information Retrieval onthe Web AIRWeB'05), 2005.http://airweb.cse.lehigh.edu/2005/bifet.pdf.

9. Jennifer Grappone and GradivaCouzin. Search EngineOptimization: An Hour a Day. John Wiley and Sons, 2nd edition,2008.
10. Ma Xiao ling, Wu Yong he." For research of search engine optimization (SEO)". Information Retrieval. 2005,12(1):6~10
11. Graf, Cousin. Search Engine Optimization.Beijing:Qinghua University Press,2007.
12. GrapponeJ,Couzin G. Search Engine Optimization:An Hour a Day. Yang Mingjuntranslation. Beijing:Tsrnghua University Press, 2007.
13. Fei Wei, Huang Ruhua. "Based on user behavior analysis of search engine optimization strategy". Library and Information Service, 2005

AUTHORS PROFILE

Jakkala Pranay Reddy is a bachelors student at Koneru Lakshmaiah Education Foundation (Deemed to be University), in the research field of big data. He is undergoing the bachelors in technology in stream of computer science and engineering in Koneru Lakshmaiah Education Foundation,Vijayawada,AP,India. His Research interests include Data Mining, Cloud Computing, and Distributed Databases.

Dantu Sri Vishnu Sai Lohit is a bachelor's student in the Department of Computer Science at Koneru Lakshmaiah Education Foundation (Deemed to be University), India.Hisinterestsfocuson Data Mining, Cloud Computing. Currently he works in the Search Engine Optimization research project. He is a part of research going in the field of big data since 2 years.

Ravanala Chandu is a bachelor's student in the Department of Electronics and Communication Engineering at Koneru Lakshmaiah Education Foundation (Deemed to be University), India. His interests focus on Data Mining, Micro Processing, and Speech Translation. Currently he works in the Search Engine Optimization research project and also as a research student in bid data field.