International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Identification of Web Site Reliability through Data Scrapping at Web Crawler's Navigation

S. Ponmaniraj, Tapas Kumar, Amit Kumar Goel

Abstract: ***Searching a specified content on the web site is like epistle a single character in bunch of pages. When the user enters their keyword into any search engines, it takes that in to web server mining process for collecting the entire terms related to that entered key phrase. Few pages gives legal and authenticated matter for the user, which they really wanted to access. Whereas many other pages are bringing them some unwanted and malicious codes of pages or virus activity pages to harm user's activities and the system's functions. Generally a web page attacks the targeted system by faulty instructions and malevolent programs through some sort of intrusion methodologies are called as phishing. In this attacking method user is set to access unknown or illegal sites by the way of accessing some unidentified websites link imbedding with legal site contents. Once victim's system performance got compromised then hackers started to do attack. To avoid this kind of molestations, user needs to understand reliability of web page's contents before started to continue browsing. This research paper is going to present web crawler architecture, design complexities and implementation for scrapping web contents from visited web pages for indentifying their reliability and freshness.***

Keywords: ***Intrusion Detection System, Parser, Scanner, Search Engine Optimization, Semantic Web, Unstructured Information Management Architecture, Web crawler, Web Robot.***

I. INTRODUCTION

In this internet era every consumer wants to access millions of web pages for a single query passed on search engines. Search engines using more optimization techniques to bring the exact content to the victims still by analyzing the keyword phrase it is in the state of producing more output pages to the user. For an example if a client machine passing the term "Crawler" in a google search engine then it brings about more than 2,57,00,00,000 sites with searched content.

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
S. Ponmaniraj *, Research Scholar, School of Computing Science and Engineering, Galgotias University, Uttar Pradesh, India, ponmaniraj@gmail.com
Prof. Dr. Tapas Kumar , Doctorate, Professor, School of Computing Science and Engineering, Galgotias University, Uttar Pradesh, India, tapas.kumar@galgotiasuniversity.edu.in
Prof. Dr. Amit Kumar Goel, Doctorate, Professor, School of Computing Science and Engineering, Galgotias University, Uttar Pradesh, India, amit.goel@galgotiasuniversity.edu.in

Same keyword passed in to yahoo search engine then it bringing more than 19,400,000 numbers of web sites and if that keyword is searched at duckduckgo search engine then it produced the result pages in terms of 34 billion web site count. From these enormous counting of web sites few sites only legally authorizing the original data to hold and others are simply holding the key terms and coming into the place to increase site counts[1] [18].

Some of these pages are maintaining by the hackers to violet the victims systems information and functions of the systems. On the internet environment, any user can access legal content on authorized web sites. Several sites are tied up with some other corporate for advertising their business. So that legal sites allow them to embed their trade mark logo or advertising image along with their company URL. Unknowingly once this logo or image got click action then automatically user is directed to the targeted web sites which user doesn't want to look for. If the directed page is a legal then there will not be any issue for the victim's data otherwise that link will find the loop hole to make users sensible data or the systems to get compromise with their security where hackers can play well their attacking process. In general any unauthorized or unwanted activities happening at the legal sites by illegal links in the form of any interruptions like asking users to click on some buttons, to follow some links unnecessarily, make users to accept something or by posting some spam like images and videos are called Intrusions. Following chapter contents will extract the process taken on web searching and crawling.

II. RELATED WORKS

Andas Amrin et.al, presented their views on Fish algorithms to identify the web contents by the way of accessing its score values. At the time of visiting every URL it update the first link and then the next linked URLs are processed by ranking their value, for relevant (1) and non-relevant (0). Depth of the related URL (Child) assigned with predetermined value in the list otherwise URL will be dropped. Their implementations on searching works like a browsing with optimized stratagems. It is faster than other algorithm to set parental and child URLs. In this model downloading web documents from WWW is consuming more time. During the progress it creates high traffic due to accessing network resources and the hidden web crawling is impossible [2]. Herseovici M et.al [3], and Lei Luo et.al [4], developed an efficient algorithm named (Adaptive) shark searching algorithm to give remedial actions against fish searching algorithm.

Searching improvement is based on conventional contents. In this progress they used the metadata from Fish searching algorithm and it brings more relevant pages by relevance classification system. More quality of information is retrieved and Operation time is lesser than its ancestor process. This method is working well on a single host or a less number of hosts. Ethan Burns et.al, [5] given their idea about heuristic search for multi core machines based on best first search.

It uses the artificial intelligence algorithms to classify the best relevancy pages by its best score. This algorithm evaluated by heuristics estimation function. In their implementation hopeful nodes are meant to explore the graphs. This process is fully working on open and closed data sets and also It depends on the promise node specification value n. By using their model only the small amount of problems will be able to solve.

Kuber Mohan et.al, [6] done the page ranking method to access the relevant URL items. In their process, they have assigned certain values for the pages on the basis of number of times visited and back linked to the source URL. It gives the relevance score for the visiting pages and also it increases the number of counts on searching links in terms of relevancy. This algorithm is fully based on queries and it doesn't depend on content of a given URL. It brings the static quality on searched page and outbound links of a source page also added advantage of searching in this method. This process is predicting client's behaviors and ranking will be calculated when it is in offline by utilizing web graphs functions.

F. U. Ogban et.al, [7] developed path-ascending crawling design for enhanced search results, which is simply started to crawl from its URL first link to last link supplies. It helps to cover all the isolated link items of a page will be crawled and maintained easily. This uniqueness makes the search engines to perform well in deep web search to all the separated source links from the ascending links items. It increases the recall and function on the query which is similar. This Uniqueness expansion is used in precision for result pages. This implementation does effective operations in identifying the isolated resources. Though this model brings lots of advantages still few degrades are there to rectify. For an example it works on simple shortest paths. Presence of self loops also one of the issues for linking reach ability. Micarelli, A et.al, [8] reviewed Focused Crawler algorithm and they found that this algorithm bring us all the web documents wherever entered search query got traverses. This algorithm focused on the content retrieval model. They also found that it has the learning ability to responsive for all the possible alterative solution of the user's query. Learning process of this algorithm ends before search starts. They concluded that it is suitable for similar interest communities.

Rashmi et.al [9] presented their views on Intrusion Detection Systems (IDS) on search engines to find the unwanted interruptions over the legal web sites. Simply it looks for URL content and the linked pages present in the specified URL. Once IDS fetches the URL ID then it matches the précised page contents with spam database to identify the stimulated desires. If any spam contents are presented then IDS warning the user about site information. Other than detections IDS cannot perform preventive actions or it cannot suggest the ways how intrusions are happening in the internet access to the administrator of a system. Based on the relevancy of the entered key terms only search engines looks for the data resides on the internet. Every search engines are working on different pattern to satisfy the user's requests. Search engines are automating the identification of relevant information from World Wide Web through URL. Semantic links are working along with search engines to bring up their expected contents and due to this search engines are also called as web crawler. Web crawler is a piece of programs to identify relevant information over WWW by using searching engines. Further this paper is going to present the chapters of web search engine architecture, issues, web crawler approaches on implementation of content analysis based on scanner and parser of natural language processor (NLP).

III. WEB SEARCH ENGINE AND CRAWLER ARCHITECTURE FOR FRESHNESS CALCULATION

Search engines are designed for the purpose of getting relevancy items from internet sources. Crawlers are just an updated instruction for the search engines to put up best throughput, response time and topical or freshness of the web sites. Throughput is the process of how many queries and related sites are brought into the picture as output for the particular time period. Response time is the time delay to process the query at user side, server side as well. Topical or the freshness of the web page is meant to be the "age" of the page visited or the number of time visited for a page. These are the technical matters to deal with designing a crawler for a search engine. Architecture of a search engines consists of Unstructured Information Management Architecture (UMIA) principle. In general search engines are processing with text documents and other unstructured data contents. To dealing with these kinds of unstructured data many software components are linked with the search engine architecture. Architecture contains the software modules and components along with interfaces to connecting them to deliver efficient progress. Two primary goals of a good search engines are effectiveness and efficiency [10].

Basic and foremost building blocks for search engine are Indexing and Query processing. Indexing is the substance to create structure for the websites and Query processing is the course of action to ranking the documents based on the structure created for user's query. Text acquisition, Transformation, Index substitutions are done through some of the important factors such as Merging of commands and algorithms to access text with databases, storing the index values, maintaining the scalability of the index created, searching capacity and providing reliable services. Text acquisition is the progress where all the input text will be indexed with searching word and traits. In general searching word is called as "Terms" and the traits are called as "Features". The collections of "Terms" and "Features" in index storage are known as "Index Vocabulary".

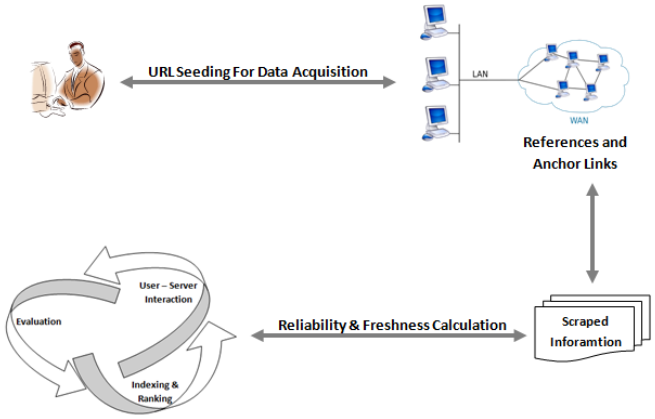


Fig. 1. Over all model architecture of Web Search Engines.

The index vocabulary simplifies the searching engine functions by means of fetching related items when user forgets about the exact terms to be searched. Indexing must be efficient in terms of space and time when larger amounts of data to be searched on the internet. It should be able to update the newly visited sites information to satisfy the scalability property. Inverted index also works in progress to optimize the search function of searching engines by the way of bringing the direct opposite documents for the entered search terms [11]. Fig. 1 shows that the overall model architecture for a search engine.

Steps involved in indexing for the search engine optimizations by the crawler;

1. Extracting URLs of both internal and external links
2. Removal of HTML tags, reference characters and styles
3. Recognizing input languages
4. Tokenization for the web contents
5. Document Parsing and Syntactic Analysis
6. Lemmatization or Stemming for natural language processing
7. Normalization or Automated translation of languages

Query processing possesses three main features such as user interactions with web server, ranking the searched documents and evaluating log data. Ranking the documents is based on the structure and relevancy about the searched keyword in indexing databases. In the other hand how frequently the sites are responded to the given keyword by the user on query. Based on the number of related content and the returning sites for the request through index values is assigning a ranking for that specified search [12].

IV. SEARCH ENGINE DESIGN COMPLEXITIES

Web crawlers are working based on set of policies on the basis of how to select, load and void the web pages when users made a request. Those policies are Assortment policies, Re – Visiting the pages policies, Civility policies and Parallelization Policies.

Policies are making the web crawlers to coordinate the search engines for their performance improvement. All these policies are well suited for the different kinds of search engines based on Web crawling, Directory accessing methods, Meta data, Hybrid level access and Perceptional activities [13]. Despite the fact that many policies to be govern by searching procedures, still crawling of web pages facing few issues at the time of implementing them with search engines.

TABLE - I Big (O) Complexity for Time and Space in web page access

Data Structure	Time Complexity								Space Complexity
	Average				Worst				Worst
	Access	Search	Insertion	Deletion	Access	Search	Insertion	Deletion	
Array	Θ(1)	Θ(n)	Θ(n)	Θ(n)	O(1)	O(n)	O(n)	O(n)	O(n)
Stack	Θ(n)	Θ(n)	Θ(1)	Θ(1)	O(n)	O(n)	O(1)	O(1)	O(n)
Queue	Θ(n)	Θ(n)	Θ(1)	Θ(1)	O(n)	O(n)	O(1)	O(1)	O(n)
Singly & Doubled Linked List	Θ(n)	Θ(n)	Θ(1)	Θ(1)	O(n)	O(n)	O(1)	O(1)	O(n)
Hash Table	N/A	Θ(1)	Θ(1)	Θ(1)	N/A	O(n)	O(n)	O(n)	O(n)
Cartesian Tree	N/A	Θ(log(n))	Θ(log(n))	Θ(log(n))	N/A	O(n)	O(n)	O(n)	O(n)
B-Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(n)
Red-Black Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(n)
Splay Tree	N/A	Θ(log(n))	Θ(log(n))	Θ(log(n))	NA	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(n)
AVL Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(n)
KD Tree	Θ(log(n))	Θ(log(n))	Θ(log(n))	Θ(log(n))	O(n)	O(n)	O(n)	O(n)	O(n)

Two main complexity issues are found at the web engines crawling, Big (O) Time and Space complexities. The table 4.1 [14] depicts the complexity analysis on the data structure concepts carried out on web page searching. Other than the above said data structure algorithm

implementation aspects, Web crawler raises four different issues for the civics and individual browsing activities in terms of privacy, cost, denial of services and copy right issues [15] [16].

At the moment when user enters their request on the search engines, then the input and retrieval content must be authenticate and reliable to access with the minimum cost benefits. Since the web crawler does the page return by mining the content from World Wide Web (WWW) through semantic analysis and searching methods, it has to meet up many intermediate servers, agent servers and gateways to passing request or taking information inside the boundary meanwhile more requests might be raised parallel to those path connectors and this causes the overloading of responding to all the users requests from various places.

V. IMPLEMENTATION FOR FRESHNESS CALCULATION

To evaluate pages and its links there should be parser to analyze the key text and anchor text. This parser reads those key phrases and separates the required contents to filter from the whole downloaded web document. The following table 5.1 shows that the vocabularies used for content separation through the lexical grammar defined by regular expression.

Web crawler does the search operations based on the input queries and the methodology adopted by search engines. After analyzing the key terms from the parser crawling of any pages has been done through with the following important steps [17].

PROCEDURE FOR WEB DOCUMENT FRESHNESS CALCULATION

- ★ Read URL Seed (S)
- ★ Look for HTML page (P)
- ★ Scanning / Retrieving HTML pages for anchor texts (Page(L_i))

$$\boldsymbol{Page(S) = Page(L1) + Page(L2) + \cdots + Page(L_{n-1})} \quad (1)$$

- ★ Parse HTML code to extract other URL links (Page(L_i))

$$\boldsymbol{Page(S) = \frac{Page(L1)}{OL(L1)} + \frac{Page(L2)}{OL(L2)} + \cdots + \frac{Page(L_{n-1})}{OL(L_{n-1})}} \quad (2)$$

- ★ Check for the freshness to assign score / page rank value,

$$\boldsymbol{Page\ Rank\ Value = IRV + \sum_{PC=0}^{N} \frac{PR}{No.of\ OL}} \quad (3)$$

Where,

- IRV = Initial Rank Value,
- PC = Page Count,
- PR = Page Rank for Web Pages,
- OL = Outbound Links

- ★ In general terms,

$$\boldsymbol{Page\ Rank(S) = \sum_{RV \in OL} \frac{PR}{No.of\ OL}} \quad (4)$$

Where,

- S = Source Page
- RV = Rank Value for Web Pages,
- OL = Outbound Links
- PR = Page Rank Score

- ★ If (Visiting pages==existed) then return to source URL
- ★ If not update new URL with indexing values
- ★ For each visited pages confirms that URL to agree for updating the checking process.

The above said mechanism is general working procedure for all the search engines to crawl the web contents. Most of the implemented algorithms are working on finding parent page or node and from that it elaborates their children nodes by accessing score values to assign its page ranking values. Typically those parent nodes are fetching from the key terms entered to search the web links by the user. Then it forms other links by the means of navigating towards the key items by its score values over internet pages.

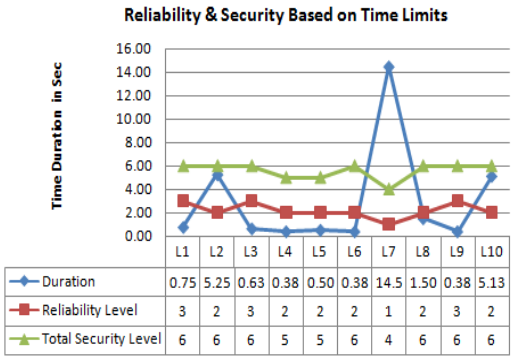
VI. RESULT AND DISCUSSION

Ten different web URL seeds were applied into the above said algorithm for finding security levels, reliability and freshness of them. This algorithm provides more than 90% accuracy of identifying page's reliability.

TABLE – II. Web document reliability and freshness

Input Phrase	Domain	IP Address	Duration	Reliability Measurements	Total Security Level		Domain Level Security		Transport Layer Security	
					Rating	Passed	Rating	Passed	Rating	Passed
https://alwaysjudgeabookbyitscover.com	alwaysjudgeabookbyitscover.com	104.248.60.43	0.75	B	B	6/6	A	1/1	A	6/6
http://beesbeesbees.com	beesbeesbees.com	217.70.184.38	5.25	C	B	6/6	A	1/1	C	0/6
https://chrismckenzie.com	chrismckenzie.com	64.13.192.155	0.625	B	B	6/6	A	1/1	A	6/6
http://eelslap.com	eelslap.com	64.13.192.209	0.375	C	C	5/6	A	1/1	C	0/6
http://endless.horse	endless.horse	104.236.181.76	0.5	C	C	5/6	A	1/1	C	0/6
http://hasthelargehadroncolliderdestroyedtheworldyet.com	hasthelargehadroncolliderdestroyedtheworldyet.com	216.92.96.71	0.375	C	B	6/6	A	1/1	C	0/6
https://heeeeeeee.com	heeeeeeee.com	118.24.246.126	14.5	D	D	4/6	A	1/1	C	4/6
http://ihasbucket.com	ihasbucket.com	103.224.182.245	1.5	C	B	6/6	A	1/1	C	5/6
https://theuselessweb.com	theuselessweb.com	159.65.2160.232	0.375	B	B	6/6	A	1/1	A	6/6
http://tinytuba.com	tinytuba.com	52.217.15.179	5.125	C	B	6/6	A	1/1	C	0/6

Table – II shows the URL seeds and reliability level. Security level of the given URL seed passes three different states such as overall rating, domain level security and transport layer security for intermediate services. Fig. 2 shows that the reliability and security level of passed input key phrases on web crawlers. It is clearly showing that when time duration going high then there will be more chance for lacking of security issues and reliability problems. Due to authenticity verification and validation of legal contents of a visiting web pages has to be consider for reliability of that particular documents and number of visiting of the same pages from various users also be mandatory for authenticating a web page as legal. In the following Fig.2, L1-L10 is web links for ten different web sites. In duration field time is measured with unit second and when throughput takes much time to discover it means that intrusion happens in middle of fetching the targeted web site.



	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10
Duration	0.75	5.25	0.63	0.38	0.50	0.38	14.5	1.50	0.38	5.13
Reliability Level	3	2	3	2	2	2	1	2	3	2
Total Security Level	6	6	6	5	5	6	4	6	6	6

Fig. 2. Time based reliability and security measurement

VII. CONCLUSION

Web search engines are used to make relationship between other pages through key terms and updating the visited links in its page rank value. Semantic web content identification plays the vital role in search engine for optimizing its crawler functions. Though many number of efficient algorithms are being developed to improve web search still some more secure functions needed to be implement along with those algorithms to secure user's and sensitive data on transactions over internet milieu. Due to insecurity options easily hackers can deploy their phony web sites with most hunting key terms.

Therefore search engines and crawlers must be implemented with some intrusion detection, prevention mechanisms and other security functions to proliferate the safe browsing.

REFERENCES

1. Rashmi Janbandhu, Prashant Dahiwale, M.M.Raghuwanshi, Analysis of Web Crawling Algorithms, International Journal on Recent and Innovation Trends in Computing and Communication, ISSN: 2321- 8169, Volume: 2 Issue: 3, March 2014
2. Andas Amrin, Chunlei Xia, Shuguang Dai, Focused Web Crawling Algorithms , Journal of Computers, May 21, 2015, doi: 10.17706/jcp.10.4.245-251
3. Herseovici, M., Jacov, M., Maarek, Y.S.: The Shark-Search Algorithm-An Application: Tailored Web Site Mapping. Computer Networks and ISDN Systems 30, 317–326 (1998)
4. Lei Luo, Rong-bo Wang, Xiao – Xi Huang, Zhi – Qun Chen, A Novel Shark-Search Algorithm for Theme Crawler, Springer, WISM 2012, LNCS 7529, pp. 603–609, 2012.
5. Ethan Burns, Sofia Lemons, Wheeler Ruml, Best-First Heuristic Search for Multicore Machines, Journal of Artificial Intelligence Research 39 (2010) 689–743.
6. Kuber Mohan, Jitendra Kurmi, Technique to Improved Page Rank Algorithm in perspective to Optimized Normalization Technique, IJARCS, ISSN No. 0976-5697, April 2017.
7. F. U. Ogban, P O Asagba, Olumide Owalobi, On a cohesive focused and path-ascending crawling scheme for improved search results, IJNAS, VOL. 8, NOS.1& 2 (2013)
8. Micarelli, A., & Gasparetti, F. (n.d.). Adaptive Focused Crawling. Lecture Notes in Computer Science, 231–262. doi:10.1007/978-3-540-72079-9_7
9. R. Rashmi, M. Vivek Anand, IDS Based Network Security Architecture with TCP/IP Parameters using Machine Learning, Electronic ISBN: 978-1-5386-4491-1, (PoD) ISBN: 978-1-5386-4492-8
10. W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines:Information Retrieval in Practice, Pearson Education, Inc, 2015
11. Hussein Al-Bahadili, Saif Al-Saab, A Web Search Engine Model Based on Index-Query Bit-Level Compression, SWSA'10, June 14–16, 2010,DOI: 10.1145/1874590.1874597
12. MarkLevene, An Introduction to SearchEngines and Web Navigation, DOI: 10.1002/9780470874233, WileyISBN: 978-0-470-52684-2, ch5.
13. Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks and ISDN Systems 30 (1998) 107- 117.
14. https://www.bigocheatsheet.com
15. Leng, A. G. K., Kumar P, R., Singh, A. K., & Dash, R. K. (2011). PyBot: An Algorithm for Web Crawling. 2011 International Conference on Nanoscience, Technology and Societal Implications. doi:10.1109/nstsi.2011.6111993
16. Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. Journal of the American Society for Information Science and Technology, 57(13), 1771–1779. doi:10.1002/asi.20388
17. https://mkyong.com/java/jsoup-basic-web-crawler-example/
18. Apoorv Vikram Singh , Vikas , Achyut Mishra, A Review of Web Crawler Algorithms, International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6689-6691, ISSN:0975- 9646

AUTHORS PROFILE

S. Ponmaniraj, is a research scholar in School of Computing Science and Engineering at Galgotias University, Greater Noida. He got his Master Degree from CSE at Sri Muthukumaran Institute of Technology, affiliated to Anna University, Chennai. Currently he is perusing Ph.D. in the area of "Network Security" for the concept of "Intrusion Detection Mechanism Based on Web Images" under the guidance of Prof. Dr. Tapas Kumar. He has 12 years of academic experiences from various educational institutions.

Prof. Dr. Tapas Kumar, is a Head of the Department, Computer Science and Engineering, Galgotias University, Greater Noida. He was awarded a Ph.D. degree with Dept. of Computer Sc. & Engineering, Birla Institute of Technology, Mesra, Ranchi (Jharkhand) under the guidance Prof. G. Sahoo and Prof. I.M.S.Lamba on An Application of Cellular Automata Paradigm in Image Processing in the Year 2013, May 27. He has received his Master Degree from Guru Jambeshwar University of Science and Technology, India and B.E. from Amravati University, Amravati. He has 106 research articles to his credit as found through Google Scholar with 542 citations has got the h-index of 12 and i10 index of 13. It has been found that article authored with his guide Dr. G Sahoo, Prof. and Dean, Dept. Of Computer Science & Engg, Birla institute of Technology, Mesra, Ranchi entitled published in the year 2010 has received highest citations of 48 with an average citation of 1.31. It is has been found that article authored with his scholar in the year 2014 received highest citations of 29 with an average citation of 1.02.

Prof. Dr. Amit Kumar Goel, is a professor and OBE coordinator for computer science and engineering department, IQAC lead member as well at Galgotias University, Greater Noida. He posses around 22 years of teaching experience from various academic institutions. He was awarded a Ph.D. degree in "Multiagent System" from Birla Institute of Technology, Misra. He has published more than 30 research articles and 2 patents. has been a member of Editorial Board of IIJC and IJAIAT national journals.