IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 10, October 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023. 1210049 |

# Implementation to Find Navigational Patterns from Log Files Using Hadoop Techniques

**Sumit Kumar Chaturvedi, Prof. Rajesh Nigam Sir**

M.Tech Scholar, Department of Artificial Intelligence & ML Engineering, TIT Excellence College, Bhopal [M.P.], India

Associate Professor, Department of Artificial Intelligence & ML Engineering, TIT Excellence College, Bhopal[M.P.], India

**ABSTRACT:** Web using mining is concerned with finding user navigational patterns on the world wide web by extracting knowledge from web usage logs. The log files which in turn give way to an effective mining and the tools used to process the log files. A web log contains lot of information so it is preprocessed before modeling. The web log file is preprocessed and converted into sequence of user web navigation sessions. The web navigation session is the sequence of web page navigated by the user during a window. The user navigation session is finally modeled through a model. Once the user for finding the interesting pattern. Modeling of web log is the essential task in web usage mining. The prediction accuracy can be achieved through a modeling the web log with an accurate model to improve the performance of the servers. Caching is a technique in which frequently visited pages are saved in proxy server caches. Pre-fetching of web pages is a new study field that, when combined with caching, dramatically improves speed. In this paper, a better algorithm for predicting the web pages is proposed. Clustering is done and then eachcluster is mined using FP Growth algorithm to find the association rules and predict the pages to be pre-fetched for storing in cache.

**KEYWORDS**: Web Usage Mining, Semantic Web, Domain, Sequential Pattern Mining, Recommender Systems and Markov Model, Prediction, Server Web Log files, Map Reduce, Hadoop

## I. INTRODUCTION

### Web Usage Mining:

In recent times, Web Usage Mining has emerged as popular approach in providing web personalization. Web usage mining is concerned with finding user navigational patterns on the world wide web by extracting knowledge from web usage logs.(we will refer to them as web logs). The assumption is that a web user can physically access only one web page at any given point in time, that represents on item.

The process of Web Usage Mining goes through the following three phases are:

**Processing Phase:**The major aim here is to clean up the web log by deleting irrelevant and noisy data. Users are also recognized in this step, and their visited web sites are sorted sequentially into sessions based on their access time and recorded in a sequence database.

**Pattern Discovery Phase**: The core of the mining process is in this phase. Usually, Sequential Pattern Matching (SPM) is used against the cleaned web log to mine all the frequent sequential patterns.

**Recommendation/Prediction Phase**: Mixed Patterns.

Web Usage mining is the field of web mining which deals with finding the interesting usage pattern from the logging information. The logging information is stored in a file known as web log file. Web log file contains lot of information like IP address, data, time, web page requested etc.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023. 1210049 |

### Domain Knowledge:

The pattern association or relationship among all this data can provided information. Information can be converted into knowledge about historical patterns and future. Domain knowledge consist of information about the data that is already available either through some other discovery or form a domain expert. Domain knowledge classify into three classes are Hierarchical Generalization Tree (HG Tree), Attribute Relationship Rule (AR-Rule) and Environment Based Constraints(EBC).

### Ontology:

Popular definition of "ontology is the specification of conceptualization". Ontology compartmentalizes the variable needed for some set of computations and established the relationship between them. Ontology types are domain, upper and hybrid. In experiment we used domain ontology of electronic item. A Domain ontology show concept which belong to part of the world.

Domain ontology is a representation by a set of concepts C within the domain and the relations R among them. For example the set C contains Product, Purchase, Supplier and Warehouse as some of the concepts of the domain ontology considered in our Model.

| Seq No. | Date | Source | Thread... | Severity | Event Id | Text |
|---|---|---|---|---|---|---|
| 8 | 5/09/2022 12:09:25... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 10 | 5/09/2022 12:09:28... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 12 | 5/09/2022 12:12:40... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 14 | 5/09/2022 12:14:55... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 16 | 5/09/2022 12:19:08... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 18 | 5/09/2022 12:19:10... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 20 | 5/09/2022 12:25:54... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 22 | 5/09/2022 12:28:10... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 24 | 5/09/2022 12:28:13... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 26 | 5/09/2022 12:35:04... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 28 | 5/09/2022 12:35:06... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 30 | 5/09/2022 12:37:48... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 32 | 5/09/2022 12:37:51... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 34 | 5/09/2022 12:59:12... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 36 | 5/09/2022( 12:59:13... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |
| 38 | 5/09/2022 13:19:09... | WinGate NAT | 248 | Info | 2 | Authorisation failure: NAT STATUS: fire |

**Fig-1 A Sample of Server Side Web log**

## II. RELATED WORK

Sneha Y.S at al in this paper has used OWL technology to add semantics to the existing navigational paths[3]. This research they present a framework for integrating semantic information along with the navigational patterns. This research evaluated the framework and it illustrates promising results in terms quality recommendation of products.

Based on principles from bioinformatics and information retrieval, Amit Bose et al established a paradigm for customizing that combines use information with domain knowledge[9]. Unlike our model, these works do not integrate the domain knowledge of the web application in all phases of web usage mining.

Vellingiri is present at all. Examination of user behaviors in web site communication can provide information that lead to modification and personalization of a user's web practice[5]. As a consequence of this, web usage mining is of

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023. 1210049 |

extreme attention the e-marketing and ecommerce professionals. Web use mining is divided into three stages: preprocessing, pattern detection, and pattern analysis. There are different techniques available for web usage mining with its own advantages and dis-advantages. This paper provides some discussion about some of the techniques available for web usage mining.

The studies presented in, extracted domain level objects from user sessions and created a user profile for each user by aggregating these objects according to their weights and a merge function. It is assumed that there already exists a domain level ontology for the website and merge functions have already been defined on every attribute of objects.

Seidenberg developed a methodology for extraction of related concepts from GALEN based on one or more classes given as input by the user Pirolli and Pitkow's research in addition to Sarukkai in lead to the use of higher order markov models for link predication. The order of a Markov Model Corresponds to the number of prior events used in predicting a future event. So a kth order Markov model predicts the probability of the next event by looking at the past k events.

In this section, we presen some applications that study of markov chain originated with a and rei and reyevich markov(1856-1922). For a sketch of Markov's life, see gives a more detailed account. One of Markov's earliest works on the topic was presented in a lectrure in St. Petersburg about a century ago.

Markov was interested in exploring sequences of random variables where this assumpition of independence did not hold.

He took some russian text from web work and omitted spaces and punctuation marks, discarded two letters as consonants or vowels.

Searching the world wide web is part of everyday life in 2012. In the early days of the WWW there were several competing search engines. But then Google emerged as the leader. One reason for the sucess of Google is the algortihm that underpins it's search engine. The Algorithm is known as PageRank and it is an application of linear algebra and markov chains.

An account of the history behind this development and some technical details can be found in Langxille and Meyer.

Hadoop Architecutre

At it's core, Hadoop has two major layers namely:

- Processing/Computation layer(Map Reduce)
- Storage layer(Hadoop Distributed File System)

IT organizations analyze server logs to answer questions about security and compliance. A server log is a basic text file that records server activities. Computer-generated logs that record data about network activities. Use full for managing network operations, especially for security and regulatory compliance. There are several sorts of server logs, but website owners are most interested in access logs, which record hits and associated data. These logs are in large amount thus resulting collection of large amount of data Big data. Big data means really a big data, it is a collection of large datasets that can't be processed using traditional computing techniques. It includes huge volume, high velocity and extensible variety of data. This data can be in structured, semi-structured or in un-structured form.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023. 1210049 |

## III.PROPOSED SYSTEM

We propose the use of domain ontology, and its integration in a complete web usage mining system, targeting web recommendation and web usage mining.

Proposed architecture divides into three phases:

In the first phase is preprocessor, a clean server-side web log is provided for preprocessing. This log contains all information about server. Web log maintain all the incoming request by users page views and items pi in the web log are mapped to their ontology concepts turning the web log into a sequence database D of semantic thus enriching the web log with semantic information.

The second phase of proposed architecture is pattern discovery. The rich sequence database D, is fed into this process. We proposed two algorithms S_P_M and Join_Apriori for sequence database in which the semantic information from the preprocessor phase is used to prune candidate sequence and sequence less than supporting count. These proposed algorithms show with experimental results, that laptop base domain ontology used in web uses mining core any algorithm S_P_M and it improved its effectiveness and performance, without compromising the quality of the frequent patterns under few conditions.

Third Phase of Proposed architecture that uses them to generate semantics association rules and n top recommendations. This phase includes techniques that allow further reporting and filtering of results. These results of this phase represents them as a set of recommended items, to be used in web site restructuring or marketing campaigns or to the active user or as decision making recommendation the administrator.

## IV. SEQUENTIAL PATTERN MINING

A proposed algorithm S_P_M a variation of Apriori algorithm, improved by semantic information for generating frequent sequences and Join_Apriori() for generate candidate is described S_P_M generates. A proposed Algorithm S_P_M taken inputs are sequence database (SQ_Database), Sequence matrix Mat[][] and minimum support and out in form of frequent sequences for rich semantic_algorithm: S_P_M (SQ_Database, Mat[][],support) is shown in figure2.

**Algorithim: Join Apriori**($S_{k-1}$, Mat[][],..)

M1 = 0

for all P,Q,__$S_{k-1}$

with P = { i1,i2,....,ik-2,ik-1}

and Q = {i1,i2,...ik-2,i'k-1}

and D =(ik-1,ik-1) /* D(ik-1,i'k-1) defines the semantic distance between ik-1,i'k-1 */

/* Maximum semantic distance maximum allowed semantic distance between any two semantic objects and can be determined as specified in D(ik-1,i'k-1) is derived

**Fig.-2Algorithm**: S_P_M(SQ_Database,Mat[][],support)

Another Proposed algorithm is Join_Apriori which take two input candidate set and calculate sequence matrix and distance of two objects Algorithm: Apriori(Lk-1,Mat[][]…) is shown in figure4.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023. 1210049 |

**Algorithm:Apriori**($L_{k-1}$,Mat[][],…)

M1=()

For all P,Q…..$L_{K-1}$

WITH P = {i1,i2,……i(K-1)}

And Q = {i1,i2,……i(K-2)i'(K-1))

And L (ik-1,l'k-1)

/* L(ik-1,l'k-1) defines the semantic distance between ik-1,l'k-1 & L(ik-1,l'k-1) */

**Fig – 3Apriori Algorithm($L_{k-1}$,Mat[][])**

## V. NEXT PAGE REQUEST PREDICTION

The integration of semantic information directly in the transition probability matrix of lower order Markov models, was presented as a solution to this tradeoff problem. We propose using semantic information as a criterion for pruning states in higher order (where k>2) selective markov models, and compare the accuracy and model size of this idea with semantic-rich markov models and traditional markov models.

Markov Model as a proposed solution to proved semantically meaningful and accurate predictions without using complicated all $K^{th}$ order. The semantic distance matrix weight martrix W and TransitionMatirx P is directly used in markov model.

## VI. EXPERIMENTAL RESULTS

All the experiments performed on system Intel B 960 processor 4GB RAM windows XP Professional OS: programs coded on Java Platform with mysql as database.

In Experiment we used web server's log file. In this file contain information is describe in figure-1. Log file size 7MB approximately. The ontology model of the domain includes the concepts Laptop (price,model,company, screensize).

The Algorithm S_P_M was run with support count(Minimum) = 0.01 and Semantic Distance(Maximum) =10.

**Table 1.1 No. hits by particular host**

| Host URL | Hits |
|---|---|
| www.icicilombard | 8685 |
| 172.168.10.170 | 34 |
| www.gmail.com | 5649 |
| www.cricinfo.co/cricket | 2340 |

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023. 1210049 |

Above table 1.1 shows that how many time particular host is access by users. For example www.icicilmbard hits 8685 which is maximum hits and minimum hits 34 on host 172.168.10.170 We need two servers 172.16.2.167 and 172.16. 19.167.Load of each server is shown in Figure-05

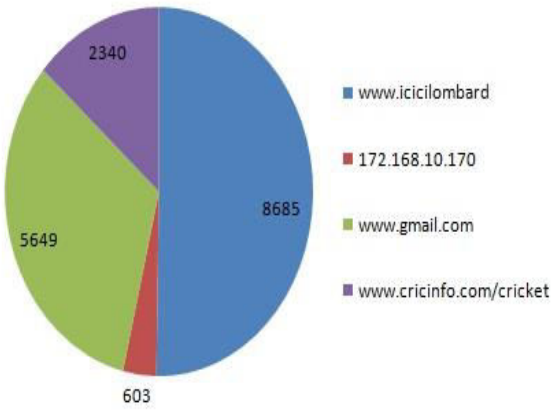


Figure- 04

**Figure-04** show that maximum and minimum hits of user.

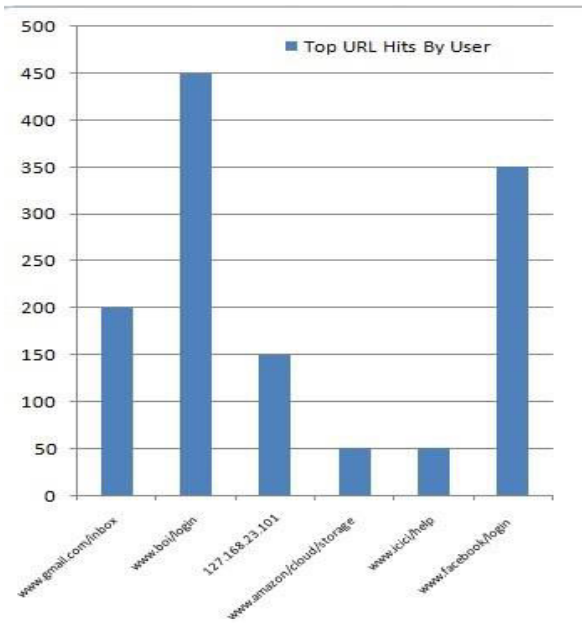


**Figure – 05**

Figure-05 shows that frequency of Top URL Hits by User.For Example www.boi/login hits 450 times by user and www.amazon/cloud/storage hits just 50 times by same user.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023. 1210049 |

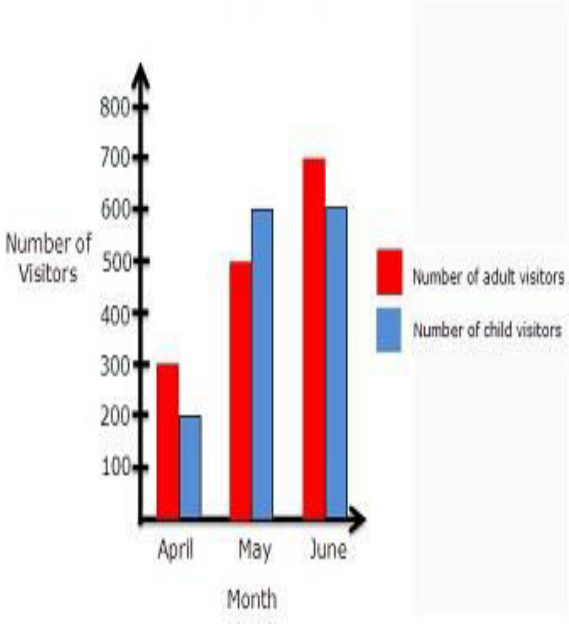


Figure-06

**Figure-06** shows that monthly record of server 127.168.23.201 uses by categorizing adult and child visitors on sites.

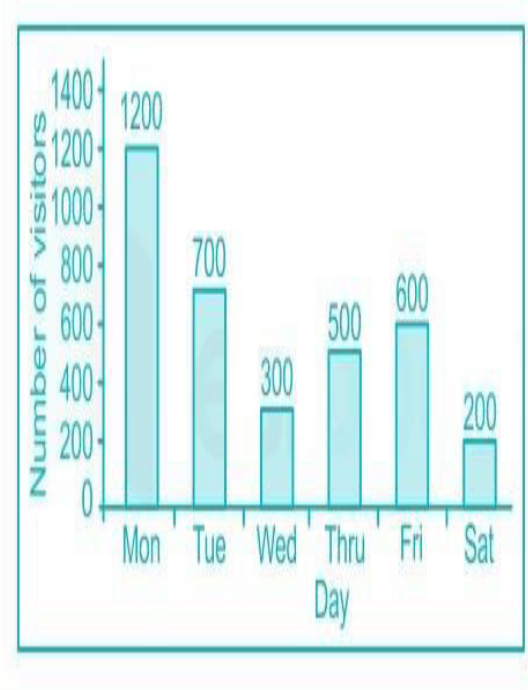


**Figure -07**

**Figure-07** shows that daily record of 127.168.23.201 uses by same user.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 10, October 2023 ||

| DOI:10.15680/IJIRSET.2023. 1210049 |

## VII. CONCLUSION

Web usage mining model is kind of mining to server logs. Web usage mining used for the improvement of improving the requirement of the system performance, the customers relation and realizing enhancing the usability of the website design. In this paper we suggest offline recommender system using markov mode for next page prediction. We proposed new framework that integrates semantic information into all the phases of web usage mining, second phase pattern discovery phase that calculate semantic distance matrix and pattern mining algorithm to prune and support counting. Semantic annotation in information extraction on web in a batter and efficient way. We build A 1st order Markov model during the mining process and enrich with semantic information, to be used for subsequently page request prediction, as a solution to ambiguous predictions problem and providing an informed lower order Markov model without the need for complex hybrid order markov models.

In future work can be

• Enhanced to live log analysis as currently this analysis is of off line analysis.

• Also it can be further enhanced to greater performance if we use parallel tasking or multi threading concept in programming.

## REFERENCES

[1] Samneet Singh and Yan Liu,"A cloud Service Architecture for Analyzing Big Monitoring Data,"ISSN1107-02141115/1011pp55-70 Volume 21,Number 1, Feburary 2016

[2] JOSPEH A. ISSA, "Performance Evaluation and Estimation Model using Regression Method for Hadoop WordCount",Received November 19,2015, accepted December 12, 2015,date of publication December 18,2015, data of current version December 29,2015.

[3] Sneha Y.S, G. Mahadevan: Semantic information and web based product recommendation system- A Novel Approach", International Journal of Computer Application(0975-8887) Volume 55- No.9,October-2012

[4]ZhuoyaoZhng Ludmila Cherkasova, "Benchmarking approach for designing a MapReduce performance model",ICPE12 April 21-24,2013

[5]J. Vellingiri, S. ChenthurPandian, "A Survey on Web Usage Mining", Global Journal of Computer Science and Technology Volume 11 Issue 4 Version 1.0 March 2011.

[6]NikadBabaiiRizvandi Javid Taheri 1, Reza Moraveji, Albert Y. Zomaya, "on Modelling and Prediction of Total CPU Usage for Applications in MapReduce Enviornment",2011

[7]Extracting Weblog for Siam University for learning user BehaviouronMapReduc 2012 4th International Conference on Intelligent and Advanced Systems(ICIAS2012)

[8]Mining of Web Server Logs in a Distributed Cluster Using Big data Technologies-(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 5, No.1 ,2014.

[9]Amit Bose, KalyanBeemanapalli, JaideepSrivastava and Sigalashar, (2006) "Incorporating Concept hierarchies into usage mining based recommendations", Processing of WEBKDD'06, PennsyIvania.

[10]Hadoop MapReduce Change log, Releease 0.22.1 – Unreleased. http://hadoop.apache.org./mapreduce/docs/r0.22.0/changes.html>,Accepted02012012

[11]Web log analysis for security compliance using big data- volume 5, Issue 3, March 2015 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[12] M.D. Kunder, World Wide Web Size- daily estimated size of the world web. http://www.worldwidewebsize.com/2011.Last Visit:2011 November

[13]H.Liu and V. Ke_delj. Combined mining of werserverlogs and web contents for classifying user navigation patterns and predicting user's future request. Data knowl. Eng. 61:304{330, May 2007.

[14] B. Mobasher, H. Dai, T. Luio, and M. Nakagawa. E_ective personalization based on association rule discovery from web usage data. In Processdings of the 3rd international workshop on Web information and data management, WIDM '01,pages 9{15,New York, NY, USA,2001.

[15] L. Page, S. Brin, R. Motwani and T. Winograd, The Pagerrank citation raning: Bringing order to the web. Techniacal Report 1999-66, StandfordInfoLab,November 1999.

[16]Miki Nakagawa and BamshadMobasher, (2009)" Impact of site characteristics on Recommendation Models Based on Association Rules and Sequential Patterns", Proceedings of the IJCAI'03 Workshop on Intelligent Techniques for Web Personalization, Acapulco, Mexico, August 2009.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details