ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 5, May 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

An Innovative Outlier Detection Scheme to Identify the Web Page Usage Strategies

S. Vasuki[1]
Assistant professor
Department of Computer Applications
J.J. College of Arts and Science, (Autonomous)
Pudukkottai
Tamil Nadu – India

Dr. K. Subramanian[2]
Head, Assistant Professor
Department of Computer Science
Government Arts College
Pulankuruchi.
Tamil Nadu – India

Abstract: ***Today in this modern world Internet is the fast growing innovative medium to work with and provide lots and lots of features to its users. Usual operation of internet requires a browser, which interconnect with the defined server and creates a log file into the server according to the usage of the user. The main motto of this paper is to analyze the web server log files from start to end and create an advanced outlier detection schema, which gathers all the records from log file and removes the outliers presented in it. Finally the resulting strategies produce the summary of web pages usage strategies, through that the web page administrator can analyze the usage of the page like how many hits the page gathers and which page contains low hit rate and which page contains highest hit rates and the date wise summary of the web page usage strategies. To achieving this web page hit detection process we introduce a new algorithm called "Web Page Hit Detection Algorithm". This new methodology is operated on the server side and checks for the web page selection and surfing strategies simultaneously at the time of user search for the required needs in we medium. Finally it makes a summary of which page is getting how much number of hits and which page is getting more hits and which one receives low number of hits. So that the web page management team can again redesign the site and posted it into the server again for getting more response from users. The outlier detection and elimination scheme is applied into this approach for eliminating the unwanted redundant data into the server end, for that an advanced outlier detection methodology is applied called "Schematic Outlier Elimination Scheme". Experimental results shows that our scheme produce better results compare to the past researches and attain the accuracy better than all other proposed earlier.***

Keywords: Clustering, Data Mining, Outlier Detection.

I. INTRODUCTION

Web based data access mechanisms are the most popular search medium now-a-days and all kinds of people using the internet to search for the required things like temples, hospitals, theatres and many more. Web access mining is an efficient method for inspecting a customer's scrutinizing conduct in Clickstream data. The discovered scrutinizing outlines not simply be used to perceive how a customer investigates through a site, however in like manner give a prevalent backing of the customer, to make adaptable site and site personalization. Consequent to webpage diagram is the most basic accomplishment variable for a webpage, especially in E-business, the examination of perusing conduct to find regular and bizarre skimming conduct assumes a urgent part in site plan. Majority of the exploration business related to web use mining strategies to discover client's searching conduct depends on the immediate technique which forms the weblog information specifically to discover either an intriguing example or uninteresting examples. Still diverse examples of client's searching conduct found can have particular implications in various sites. An intriguing example in a site may not be fascinating in another. In this framework, we propose another web use mining way to deal with identify uncommon perusing practices of the client to unravel this glitch. The proposed

methodology is helpful for site fashioners to gage how a client scans their site, particularly for those creators why should sharp overhaul their site.

The method of reasoning behind this methodology of web use mining is that the fashioner of the webpage must be in a position to characterize examples of common scanning conduct, and afterward by utilizing this example, the planner ought to have the capacity to find any surprising deviations. The thought behind this is the general configuration idea of the site is best comprehended by the architect of the site and the originator is the best individual to characterize a typical navigational example. Utilizing this predefined designs and the proposed DASPAT [3] calculation, skimming designs that don't coordinate the predefined are distinguished as examples of unordinary perusing conduct. The site architect can then utilize these found information to find the weaker segment of the site with low hits [1] and act as needs be.

II. WEB PAGE HIT DETECTION ALGORITHM

Purpose of Detection

For each and every user there is a value for web page administrators, once they enter into the site lots of benefits arises to the page owner via advertisement providers like AdSense, click sense, ad words and many more. Not only web page administrators or owners concentrating only on advertisements, they highly concentrate on user satisfaction, means once the user enter into the page they have to feel the following things in it like without scrolling they can view all details, good color contrasting, meaningful visuals and so on. So that the administrators need to know how many users like the particular page and continuously visit in it again and again. For all these kind of purpose we need to implement the Web Page Hit Detection Algorithm, so that it is more beneficiary to both web page administrators as well as web page users.

Web Mining Approach

Web mining [2] is the utilization of information mining procedures to naturally find and remove data from Web reports and benefits. There are three general classes of data that can be found by web mining: Web action, from server logs and Web program movement following. Web mining is the utilization of information mining to the web information and follows client's meeting practices and separates their hobbies utilizing designs. Since this region is relevant in e-trade and Web examination straightforwardly, web mining has gotten to be one of the imperative zones in software engineering. Web Usage Mining utilizes mining systems as a part of log information to remove the conduct of clients which is utilized as a part of different applications such as customized administrations, versatile sites, client profiling, making appealing site.

Algorithm

The algorithm and the required calculations for the Web Page Hit Detection Algorithm are described below:

Algorithm-1: Hit Detection Algorithm

Input Summarizing Web Pages

Summary.Clear();

get TransSummary();

Get No. of Transactions

getntrasactions()

declare PageIntialHits(i);

declare ListItem (PageIntialHits);

for(i = 0,i++,i<=nItems - 1)


```
                PageIntialHits = PageIntialHits + " " + Items.Item(i)
        return (PageIntialHits)
```

Check Surfing Data

```
declare 4 integer values (i , j, n, Surfstatus);
assign i and j value is 1;
declare string called unistring;
add item(0) to uniqueitms;
            for i = 1 to nitems - 1
                Surfstatus = 0
                for j = 0 to n - 1
                    if uniqueitms.item(j) = items.item(i) then
                        Surfstatus = 1
                    end if
                if Surfstatus = 0 then
                add item(i) to uniqueitms;
                    increment n by 1;
                close loop[if]
            close loop[1]
                Surfstatus (n) to surfdata;
for i = 0 to n - 1 unistring=unistring and  uniqueSurfItems(i)
        Set setuniqueitmset(Surfstatus) to uniqsurrfSts;
```

Child Page Analysis

```
        declare 3 different integer variable (i , totpages, t = 1);
        declare string variable called merge and disp;
        totpages = getsubpages() - 1
        for i = 0 to totpages
            add subpg.unipages(i) to subtotpages;
            add subpage1.unipagelist(i) to ftotpages;
            increment totpages  by 1;
        close loop[1]
        for i = 0 to totpages findsubPages()
        return subtotpages
```


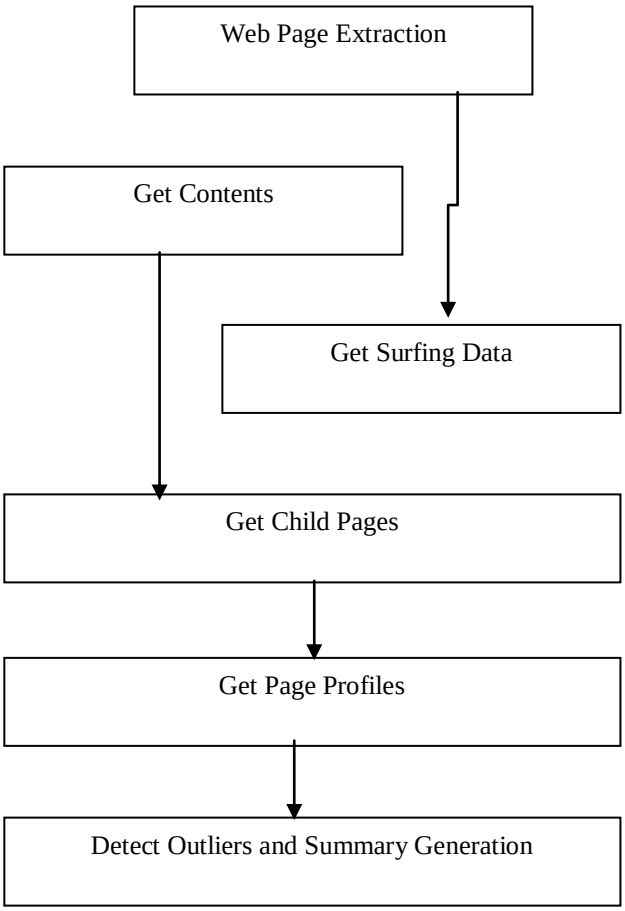

Fig.1. Page Extraction and Outliers Detection Scheme

III. SCHEMATIC OUTLIER ELIMINATION SCHEME

Grouping content into predefined classifications is a central errand in data recovery (DR). DR and web mining strategies have been connected to sort website pages to empower clients to oversee and utilize the tremendous measure of data accessible on the web. Accordingly, creating easy to understand and mechanized devices for overseeing web data has been on a higher interest in web mining and data recovery groups. Content order, data directing, ID of garbage materials, theme ID and organized pursuit are a portion of the problem areas in web data administration. A lot of methods exists for ordering web archives into classifications. Interestingly, none of the current calculations consider records having 'fluctuating substance's from whatever is left of the reports taken from the same area (classification) called web content outliers. In this methodology, we exploit the HTML structure of web and n-gram procedure for incomplete coordinating of strings and propose a schematic outlier detection scheme based calculation for mining web content anomalies. To decrease the handling time, the upgraded calculation utilizes just information caught as a part of <Meta> and <Title> labels. This approach results in utilizing planted themes demonstrate the proposed calculation is equipped for finding web content outliers. Furthermore, utilizing writings caught as a part of <Meta> and <Title> labels gave the same results as utilizing content installed as a part of <Meta>, <Title>, and <Body> labels.

Outlier Sample

{"S_id": 97,

"Skind": "BlogPost",

"Stitle": "MySQL Schema Design",

"Scontent": "Pattern adaptability ...",

"Screator": "Sway",

"Sdate": "2016-02-14",

"Sremarks": [

{ "ScommentContent": "Not sure...",

"ScommentDate": "2014-02-24" },

{ "ScommentContent": "Let me notice ...",

"ScommentDate": "2014-02-26" }] }

Algorithm-2: Schematic Outlier Elimination

Input: WebPage, Content DI

Outputs: Outlier Present in Web Pages

Used Variable: weights[w[Tk]], penalties[p[Tk]]

1. Read the subject and details of the WebPage[Di] and related Contents

2. Generate n-ranges frequency profile for content learning

3. Generate n-range frequency profile for content

4. For [int i =0; m< No-Of-Doc i ++] {

5. For[int n =0; n< No-Of-Outliers ; n++{

6. IF [N-range exists in content]{

7. =[∗ ∗] i k n j e W i p Nj w Tk F N j Tk Dieight [ ] [ ] [, ,] Else

8. = [∗]iknj eW i w Tk F N j Tk Dieight [ ] [, ,]End IF

9.} // end of loop

10. DIm =Weightage/ NoofOutliers

IV. Dataset and Methodology

The dataset comprise of 200 words (pages from 200 sites giving words administrations) and 10 home change (pages demonstrating home change administrations). The substance encased in the labels <Body>, <Title>, and <Meta> were recovered. To diminish excess, we just utilized information gathered as a part of the "Depiction" of <Meta> tag. In the wake of preprocessing, a lexicon words for the words class is created utilizing 50 word pages. The staying 150 words pages numbered P1 to P150 and 10 home change pages numbered P151 to P160 constitute our test information.

The 10 home change pages constitute the planted themes which we need to distinguish. Initially, we created 4-territory recurrence profiles for the lexicon and our test information independently and processed the uniqueness between every page and the word reference. The calculation effectively identified 9 home change pages and 1 words page among the main 10 outliers. Each of the 10 home change pages were among the main 10 anomalies when the examination was ran utilizing 5-territories as appeared as a part of Table 1. The test is rehashed utilizing <Meta> and <Title> labels just. We made another word reference from the same 50 words pages utilizing content encased as a part of <Meta> and <Title>. The data contained in the <Body> tag was disregarded. The content encased in <Meta> and <Title > labels of the staying 150 words pages and 10 home changes pages were separated and preprocessed. The 4-territory and 5-territory recurrence profiles were created for the word reference and test information independently. Difference measures were processed and the main 10 outliers decided. In both cases, every one of

the 10 home change pages was recorded among the main 10 outliers. The outline of trial results is appeared in Table 1.

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="5544" NEWID="1">
<DATE>26-FEB-1987 15:01:01.79</DATE>
<TOPICS><D>cocoa</D></TOPICS>
<PLACES><D>el-salvador</D><D>usa</D><D>uruguay</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;C T
&#22;&#22;&#1;f0704&#31;reute
u f BC-BAHIA-COCOA-REVIEW    02-26 0105</UNKNOWN>
<TEXT>&#2;
<TITLE>BAHIA COCOA REVIEW</TITLE>
<DATELINE>     SALVADOR, Feb 26 - </DATELINE><BODY>Showers continued
throughout the week in
the Bahia cocoa zone, alleviating the drought since early
January and improving prospects for the coming temporao,
although normal humidity levels have not been restored,
Comissaria Smith said in its weekly review.
     The dry period means the temporao will be late this year.
     Arrivals for the week ended February 22 were 155,221 bags
of 60 kilos making a cumulative total for the season of 5.93
mln against 5.81 at the same stage last year. Again it seems
that cocoa delivered earlier on consignment was included in the
arrivals figures.
     Comissaria Smith said there is still some doubt as to how
much old crop cocoa is still available as harvesting has
practically come to an end. With total Bahia crop estimates
around 6.4 mln bags and sales standing at almost 6.2 mln there
are a few hundred thousand bags still in the hands of farmers,
```

Fig.2. Dataset Sample

V. WEB DATA LOG SHEET SAMPLE

The web data log [4] sheet helps to the server administrator to find out the logs which are manipulated by the clients at the time of surfing the required data through web interface. These logs are extremely helpful to identify the working and usage of client and the required web pages. For instance, the following web data log sheet provides a clear vision.

Table I. Web Pages Planted Design

Data	Web Content Outliers	4-Region	5-Region
<Meta>, <Title>, <Body>	{"S_id": 97, "Skind": "BlogPost", "Stitle": "MySQL Schema Design", "Scontent": "Pattern adaptability ...", "Screator": "Sway"}	DI pages 9 Resume pages 1	DI pages 10 Resume pages 0
<Meta>, <Title>	"ScommentDate": "2014-02-24" }, { "ScommentContent": "Let me notice ...", "ScommentDate": "2014-02-26" }	DI pages 10 Resume pages 0	DI pages 10 Resume pages 0

Table II. Sample Web Data Log Sheet

Surfing IP Address	Action	Protocol	Browser	Browsed On
192.168.1.2	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.3	GET	HTTP 1.1	Opera	11/2/2015
192.168.1.4	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.6	GET	HTTP 1.1	Internet Explorer	11/2/2015

192.168.1.5	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.7	POST	HTTP 1.1	Internet Explorer	11/2/2015
192.168.1.4	POST	HTTP 1.1	Opera	11/2/2015
192.168.1.6	GET	HTTP 1.1	Firefox	11/2/2015
192.168.1.5	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.7	POST	HTTP 1.1	Internet Explorer	11/2/2015
192.168.1.4	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.6	GET	HTTP 1.1	Opera	11/2/2015
192.168.1.5	POST	HTTP 1.1	Google Chrome	11/2/2015
192.168.1.7	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.4	POST	HTTP 1.1	Firefox	11/2/2015
192.168.1.6	GET	HTTP 1.1	Firefox	11/2/2015
192.168.1.5	POST	HTTP 1.1	Opera	11/2/2015
192.168.1.7	POST	HTTP 1.1	Google Chrome	11/2/2015
192.168.1.4	POST	HTTP 1.1	Opera	11/2/2015
…	…	…	…	…

VI. PROBLEMS IN DETECTION OUTLINERS

Uniquely talking outliers [5] are examples that go amiss from expected typical conduct, which in its least difficult structure could be spoken to by a locale and imagine every single ordinary perception to fit in with this typical area and consider the rest as anomalies This methodology looks basic yet is profoundly testing because of taking after reasons. It is exceptionally hard to characterize the typical conduct or an ordinary locale. The challenges are as under.

- ✓ Encompassing of each conceivable ordinary conduct in the area.
- ✓ Imprecise limit in the middle of ordinary and exception conduct following now and again anomaly perception lying near the limit could really be typical, and the other way around.
- ✓ Adaptation of noxious foes to mention the anomaly objective facts seem like typical when outliers result from malignant activities.
- ✓ In numerous areas ordinary conduct continues advancing and may not be present to be a delegate later on.
- ✓ Differing thought of anomalies in various application spaces makes it hard to apply system created in one area to another. For instance, in the restorative space a little deviation from ordinary body temperature may be an anomaly, while comparable worth deviation in money markets area may be considered as would be expected. Indeed, even inside of same area say wrongdoing discovery there could be circumstances where utilization of remote make weapons might be viewed as ordinary in violations submitted in metro urban communities however an anomaly for homicides of normal people in tribal districts.
- ✓ Availability of named information for preparing/acceptance of models utilized by anomaly discovery strategies.
- ✓ Noise in the information which has a tendency to be like the genuine anomalies and consequently hard to recognize and uproot.

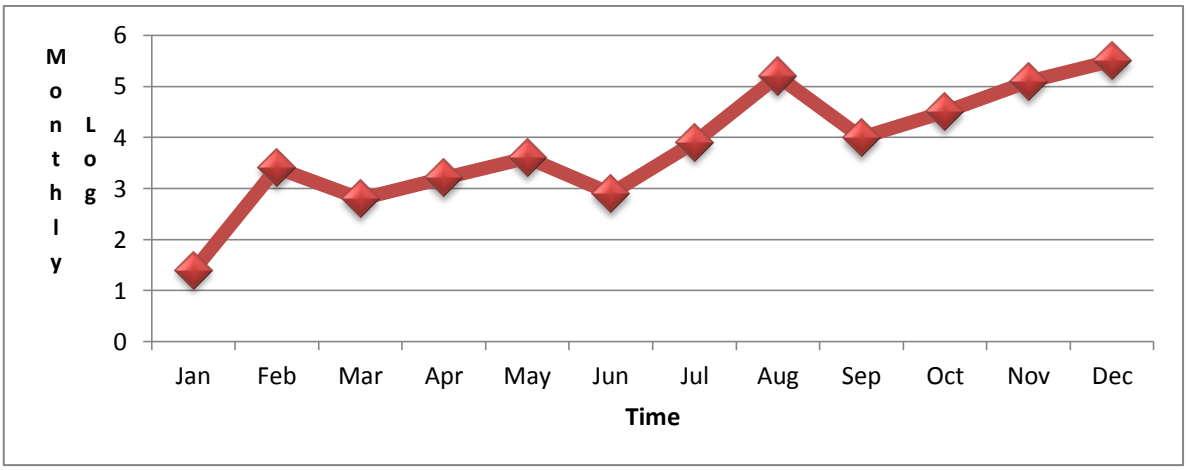

Fig.3. Logical outlier in a respective dataset time arrangement. Temperature at time t1 is same as that at time t2 however happens in an alternate setting and consequently is not considered as an outlier

Table III: Differences in Nature of Host Based and Network Outlier Frameworks

FEATURE	HOST BASED	NETWORK BASED
Outliers-In	Operating System Calls	Network Information
Interpret To	Outlier Code, Unusual Nature, System Violations	DoS Service in Network
Scenery of Data Analysis	Linear Analysis	Linear Data Collection
Outlining/Granularity	Programming Code / User Programming Nature	Packet Level and Network Flows

Table IV: Methods for Identifying Web Outliers and Fradualants

METHODS	BY-OPERATION	BY OWNER
Web Context	Location / Place	User
Price	Price is expensive , even work with long term or large data processing	Price is expensive , even work with long term or large data processing

Fig.4. Comparison of Existing Hit Detection and Proposed Method.

VII. Conclusion and Feature Work

Web mining is a developing examination territory in the mining group in view of the colossal support the web keeps on getting a charge out of. The accomplishment of electronic business on the web has pulled in mechanical support for web mining research (ex: analysis of web page usages, hit detection and so on). Arranging site pages into predefined classes to help data seek, hits identification and web page recovery from faults is exceptionally regular errand. In any case, filtering through officially ordered records, searching for site pages with shifting substance has not gotten any consideration in the mining group. This paper proposes an n-region-based calculation for mining web content exceptions.

Exploratory results utilizing planted themes demonstrates the proposed calculation is fit for distinguishing web content exceptions in web information. The outcomes additionally affirms that for officially ordered pages, utilizing information catches as a part of metadata [<Meta> & <Titles>] labels give the same results as utilizing the real substance of the pages. Utilizing metadata as a part of the trial decreased the quantity of conceivable n-regions correlations radical ly yet we mean performing more tries different things with extensive datasets to build up this. Regions of future exploration incorporate exploratory assessment of full word match calculations and n-region-based calculations regarding exactness, review and reaction time. We should likewise analyze the aftereffects of our uniqueness measure with standard similitude grids. At long last, benchmark information should be built up for assessing the execution of web exception mining calculations.

References

1. S. Vasuki, Dr. K. Subramanian. “A new outlier detection approach to discover low hit web pages using sequential frequent pattern mining to improve website’s design, IJCSI, Volume 12, Issue 6, November 2015
2. Robert Cooley, Bamshad Mobasher, Jaideep Srivastava , “Web Mining: information and Pattern Discovery on the WWW”
3. DASPAT “A new outlier detection approach to discover low hit web pages using sequential frequent pattern mining to improve website’s design, IJCSI, Volume 12, Issue 6, November 2015
4. Agyemang, M., Barker, K., & Alhajj R. Framework for Mining Web Content Outliers. Proceeding of the 19th Annual ACM Symposium on Applied Computing (ACMSAC), Nicosia, Cyprus, 2004, pp 590-594
5. Agyemang, M. and Ezeife, C.I. LSC-Mine: Algorithm for Mining Local Outliers. Proceedings of the 15th Information Resource Management Association (IRMA) International Conference, New Orleans, 2004, pp 5-8
6. Barnett, V. and Lewis, T. Outliers in Statistical Data. John Willey, 1994
7. Breunig, M.M., Kriegel, H-P., Ng R.T., and Sander, J. LOF: Identifying Outliers in Large Dataset. Proc. of ACM SIGMOD 2000, Dallas, TX 2000
8. Mannila, H., & Toivonen, H. (1996). Discovering generalized episodes using minimal occurences. In 2nd Intl.Conf. Knowledge Discovery and Data Mining.
9. Mannila, H., Toivonen, H., & Verkamo, I. (1995). Discovering frequent episodes in sequences. In 1st Intl. Conf.Knowledge Discovery and Data Mining.
10. O. R. Zaiane, M. Xin, and J. Han, “Discovering web access patterns and trends by applying olap and data mining technology on web logs,” in ADL ’98: Proceedings of the Advances in Digital Libraries Conference. Washington, DC, USA: IEEE Computer Society, 1998, pp. 1-19
11. Albanese, M., Picariello, A., Sansone, C. & Sansone, L. (2004), ‘Web personalization based on static information and dynamic user behavior’, WIDM’04, USA pp. 80–87.
12. Liu, B., Hsu, W. & Ma, Y. (1999), ‘Mining association rules with multiple minimum support’, KDD, San Diego pp. 337–341.
13. Kousalya, Suguna, Saravanan “Improving the Efficiency of Web Usage Mining Using K-Apriori and FP-Growth Algorithm “,March-2013.
14. R. Cooley, B. Mobasher, and J. Srivastava, “Data preparation for mining world wide web browsing patterns,” Knowledge and Information Systems, Vol. 1, No. 1, pp. 5-32, 1999.
15. Jung, J.J., & Jo, G-S. Semantic Outlier Analysis for Sessionizing Web Logs. Proceeding of 14th European Conference on Machine Learning/7th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD), Cavtat – Dubrovnik, 2004, pp 13-25

AUTHOR(S) PROFILE

S. Vasuki, the educational qualification of author is M.Phil. In computer science done in AlgappaUniversity, Karaikudi, Tamilnadu, India. In the year of April 2008.P.G degree M.S (IT&M) in Ayya Nadar Janaki Ammal College, Sivakasi,Tamilnadu, India in the year of April 2003. The author's major area of interest is data mining. She presented and participated in various colleges International and national conferences. Currently she is doing Ph.D. in Bharathidasan University under part time mode and working as an Assistant professor in J.J college of Arts and Science (Autonomous), pudukkottai, Tamilnadu, India.

Dr. K. Subramanin earned his Ph.D. degree in Alagappa University on 2012.Now he is guiding 8 research scholars in Bbharathidasan University Tiruchirappalli, Tamilnadu, India. He is having more than 18 years teaching experience. Currently he is working as a Head, assistant professor in Government Arts College, Tamilnadu, India.