ISSN 2320 - 2602
Volume 10 No.9, September 2021
International Journal of Advances in Computer Science and Technology
Available Online at http://www.warse.org/IJACST/static/pdf/file/ijacst011092021.pdf
https://doi.org/10.30534/ijacst/2021/011092021

A Conceptual Framework for Detecting and Analysing Website Performance Anomalies

Alao O.D.[1], Joshua J.V.[2], Ajufo C.[3], Onanuga G. A.[4]
[13]Babcock University, Department of Computer Science, Ilishan-Remo, Nigeria, jimialao@hotmail.com
[2]Babcock University, Department of Software Engineering, Ilishan-Remo, Nigeria, joshuaj@babcock.edu.ng
[4]Ogun State College of Health Technology, Department of Computer Science, Ilese, Nigeria, onanugagboyega@gmail.com

Received Date : August 02, 2021 Accepted Date : August 25, 2021 Published Date : September 07, 2021

ABSTRACT
Anomalies in website performance are very common. Most of the time they are short and only affect a small portion of the users. However, in e-commerce an anomaly is very expensive. Just one minute with an underperforming site means a big loss for a big e- commerce retailer. E-commerce web site operations are heavily transactional and prone to small, short time failures. Anomalies are sometimes small, and as such, they are not caught by the retailer web operations. However, the customers do perceive these anomalies. This paper highlights the major websites anomalies and formulates a conceptual framework that analyses them.

Key words: Anomaly, e-Commerce, Websites, World Wide Web (WWW).

1. INTRODUCTION

The internet in general has been more productive and effective; it has affected and immensely improved our everyday lives and activity sectors which include the economy, education, and entertainment industry in all the countries of the world. The internet currently inhabits not less than a billion websites and has many frequent daily users [1]. There are rapid uses of mobile devices with so many useful functionalities and with the introduction of the "Internet of Things" ("IoT"), the amount has increased to more than 50 billion devices overtime [4].

The increasing establishment and use of web technologies towards the WWW, has made the Internet users more cautious about pre-existing/primitive users' needs when checking the web. This automatically changes the constant interaction and alliance [3]. More than half of the contents of the internet is being launched on websites that are basically Hyper Text Markup Language (HTML) webpages and conjunction with JavaScript and Cascading Style Sheet (CSS). The modern versions in html protocol and the browser stacks allow webpages to be seen on any device that runs browser software and with Internet connectivity.

With the increase in technology and websites, various errors/mistakes are made with website design which has led to vulnerabilities and user dissatisfaction [16]. Most websites have a common set of anomalies, which includes but not limited to vulnerability, mobile incompatibility, lack of content, unclear target audience, bad navigation, broken links and long load time which is another annoying quality of a poorly designed webpages.

Broken links (404s) and the problem of ascertaining a user's position is another basic problem a user might face. It is clearly stated that there is useful information hidden or embedded in a link, but when that link is clicked or pressed upon, an error page displays instead of displaying the data wanted by the user. Unfortunately, broken links can stack up very quickly, which often time happens [2].

Uncompressed images contained in the websites also makes them to be digitally slow i.e., it takes forever to load and oftentimes people are impatient so they leave the site almost immediately. This paper highlights the major websites anomalies and formulates a conceptual framework that detects and analyses them.

1.1 Statement of the problem

The World Wide Web over the past decade has grown exponentially in all facets. With this growth come vulnerabilities of websites, incompatibility with devices, lack of contents, etc. Website anomalies can cause breaches which often lead directly or indirectly to fraudulent activities such as stealing ones' identity online, regulatory fines, destroyer of brands, heavy crimes against the law, downtime, virus breeding, and shortage of users. [9].

According to a white paper on security [13], 86% of websites have more than one severe vulnerability and the probability of information leakage is about 56%. Also, 30% of traffic and 15% of all sales that are generated online are done from mobile devices [9]. Thus, website abnormalities greatly affect users experience and functionality.

The main goal of this research is to carry out a review of website anomalies and their effects on users and come up with a conceptual anomalies detection/analysis framework.

1.2 Objective of the study

The main objective of this study is to develop a conceptual framework for analyzing website performance anomalies. The specific objectives are:

i. A review of different types of website performance anomalies
ii. Develop an anomaly behavior analysis framework of websites that catches not expected patterns in the data on real time. The output of this model is a label that states whether a specific instance is an anomaly or not.

2. THE START OF THE WEB AND WEB DESIGN

What came to be known as the World Wide Web in the early 90s was proposed by Tim Berners-Lee in 1989, whilst working at CERN. At this time, the concern was with text files and how to view them on the browser (Berners-Lee, 2012). It was however in 1993 that Marc Andressen and Eric Bina built the mosaic browser. There were many UNIX based text heavy browsers. At this time, there was no concerted design effort geared towards graphics. The W3C was created in October 1994" to lead the World Wide Web to its full potential by developing common protocols that promote its evolution and ensure its interoperability [17]. This discouraged any one firm from monopolizing the World Wide Web or any programming language used for its development – this would have altered the way the web is perceived. The World Wide Web Consortium (W3C) sets "the standards for the web as is seen with the use of JavaScript. Andressen in 1994 founded communication corps which later became Netscape. Netscape took the prerogative to create special (Hyper Text Markup Language) HTML tags without regard to the standards. For example, Netscape 1.1 included tags for changing background colors and formatting text with tables on web pages. Between 1996 through 1999, Microsoft and Netscape fought over browser dominance. During the time of the browser wars, many great technology were born, including JavaScript, Dynamic HTML, and CSS. In general, the browser wars era gave birth to many positive and huge leaps for web design [10].

I.

2.1 What is a Website Performance Anomaly?

A website performance anomaly is a sustained spike in website performance that is outside normal expected variation for that page. In other words, it is a sort of unusual delay in page performance.

2.2 Causes of Websites Performance Anomalies?

According to [12], Modern e-commerce websites are complex. A typical e-commerce page includes hundreds of elements such as high resolution images, CSS files, database calls, 3rd party JavaScript, etc. It's not uncommon for a page to have more than three hundred (300) calls to over one hundred (100) servers that must all return content correctly in order for the page to load. Anomalies occur when something interrupts the loading of an HTML page. Some of the most common causes of websites anomalies are third party JavaScript errors, incorrect image placement and sizing, traffic errors, response codes, bots and security threats. These are just the most common causes. In truth, there are an endless number of events that can cause a performance anomaly.[12] further highlighted other causes of website performance anomalies as follows:

2.2.1 Un-optimized Images

Un-optimized images are images that could be further compressed in size, without any noticeable changes. Optimized images are usually very identical when compared to the original file, these images lose the meta-data that helps describe them. These meta-data are often very useful to the designer and not end-users.

2.2.2 Content Served Without Compression

Web contents are usually large where multimedia files are included. But enabling Hyper Text Transfer Protocol (HTTP) compression on the webserver increases drastically the speed of page loading. The downloaded page is therefore reduced drastically.

2.2.3 Combinable CSS Images

Browsers usually make single request for every CSS background image specified on the page. It is therefore, significantly meaningful to combine CSS images into single sprites to enable the server make one call to the images.

2.2.4 Images without Caching Information

HTTP caching, helps the browser to save a copy of the images to the local computer so that the next time that page loads again, it will not have to make requests to the server for that same image file but only load from the cache, it is necessary to update webserver configuration to provide expiry time to the header of image requests. For images such as icons, which do not change often the expires header attribute should be such that it's far into the future, typically about 6 months to a year out from the current date.

2.2.5 Domain Shading Not Implemented

Browsers usually allow 2-4 simultaneous downloads of static files from every hostname. Therefore, if any web page is downloading too many static resources from one hostname, the browser finds it difficult to download the contents. This

can be rectified by dividing the task over various hostnames. While it is difficult to physically move files to new hosts, host names can be mapped to trick the browser into downloading numerous concurrent static resources without any bottle necks [7].

3. ANOMALY DETECTION

Anomaly Detection is the art of finding patterns that do not conform to certain predefined normal characteristics. Detecting such nonconformities from predictable activities in temporal data is essential for certifying the normal processes of systems across several domains such as economics, biology, computing, finance, ecology. It is important to characterize what is usual, what is different or irregular and how noteworthy the anomaly is. This classification is forthright for systems where the performance can be defined using simple mathematical models – for example, the output of a Gaussian distribution with known mean and standard deviation". However, real life systems have complex behaviors over time. It is essential to classify the normal state of the system by noticing data about the system over a period of time when the system is assumed normal by monitoring users of that system and to use this characterization as a reference point to identify anomalous behavior.

3.1 Anomaly Detection Methods:

[5] identified the following three categories of data mining approaches used to detect anomalies from online social networks:

3.1.1 **Supervised Anomaly Detection Techniques:** Supervised anomaly techniques are used to model both normal and abnormal behaviors. These techniques require pre-labelled data for anomaly detection classified as normal or abnormal. Different training models are used to identify the normal or abnormal data from dataset. Supervised techniques work on two approaches:

3.1.1.1 Training model is compared with dataset to find analogues data from data set that is classified as normal data.

3.1.1.2 In opposite to above method some anomalous data is compared against training model to find abnormal data from dataset.

3.1.2 **Unsupervised Anomaly Detection Techniques:** Unsupervised techniques work on clustering mechanism. These techniques have no pre-labelled data normal or abnormal. These techniques find clusters of nodes whose behavior is similar to group. Sometimes this assumption becomes wrong as many anomalies also make clusters with similar pattern; therefore, unsupervised techniques are inefficient to find accurate results.

3.1.3 **Semi Supervised Anomaly Detection Techniques:** In semi supervised techniques, data set is only labeled with one label as normal. Training model detect abnormal class by itself from dataset [11].

3.2 Anomaly Detection Operator

An anomaly detection operator is used to detect different types of anomalies in event streams. For example, a slow decrease in free memory over a long time can be indicative of a memory leak, or the number of web service requests that are stable in a range might dramatically increase or decrease.

According to [6], anomaly detection operator detects three types of anomalies:

(i) **Bi-directional Level Change**: A sustained increase or decrease in the level of values, both upward and downward. This value is different from spikes and dips, which are instantaneous or short-lived changes.

(ii) **Slow Positive Trend**: A slow increase in the trend over time.

(iii) **Slow Negative Trend**: A slow decrease in the trend over time.

When using the Anomaly Detection operator, you must specify the limit duration clause. This clause specifies the time interval that should be considered when detecting anomalies.

4. REVIEW OF RELATED WORKS

Cheng et al (2021) in their paper titled An Improved Feature Extraction Approach for Web Anomaly Detection Based on Semantic Structure proposed an improved feature extraction approach which leveraged the advantage of the semantic structure of request URLs. Semantic structure is an inherent interpretative property of the URL that identifies the function and vulnerability of each part in the URL. Their evaluation showed that feature extraction method has better performance than conventional feature extraction routine by more than average dramatic 5% improvement in accuracy.

[14] collectively wrote an article called the real-time anomaly detection for streaming analysis. This research paper made emphasis on how data can be streamed and the abnormalities that come with passing out useful information, the detection abnormalities while streaming data entities in real life and also learning to make simultaneous predictions. The methodology procured was a novel abnormality detection technique based on an online sequence algorithm which was created which is hieratical temporal memory (HTM). The result gotten was a live application that detects abnormalities in financial metrics which is done in real time.

[15] developed a framework of analysis techniques for abnormal behavior in mobile applications, these abnormal behaviors in mobile phones can cause a lot of harm and side effects like insufficient implementation of application life

cycle, memory issues and virus related problems which can cause problems like crashing, bad usability and even data loss in mobile applications. The framework developed established a way of tracking abnormal behavioral traits in mobile applications. In their work, static and dynamic analysis techniques were discussed and also implemented in android applications for identifying the causes of abnormal behaviors.

[8] in his article titled Web Performance Anomaly Detection with Google Analytics highlighted two steps to detecting web performance anomalies. Step one monitor all the things. Step two, dedicate 90% of your analytics time and resources to analyzing data, deriving insights, and **iterating** on what metrics are being monitored and are being optimized. However, there is one small problem. **Chances are, the amount of data produced by the instrumentation outpaces your ability to analyze, monitor, and correlate all the variations of the variables at play.** Google Analytics samples page load time performance data for browsers that support the W3C Navigation Timing API's, which includes: redirect and DNS times, TCP establishment, server response times, as well as DOM-level metrics such as the on-load time. There are over half a dozen metrics in total, each recorded from a real user accessing your site - in other words, this is Real User Measurement (RUM), not synthetic data.

5. CONCEPTUAL FRAMEWORK DESIGN

Figure 1 shows the general architecture for anomaly detection in website while figure 2 shows the stages involved in the anomaly detection architecture.

Figure 1: **The General Architecture for Anomaly Detection**

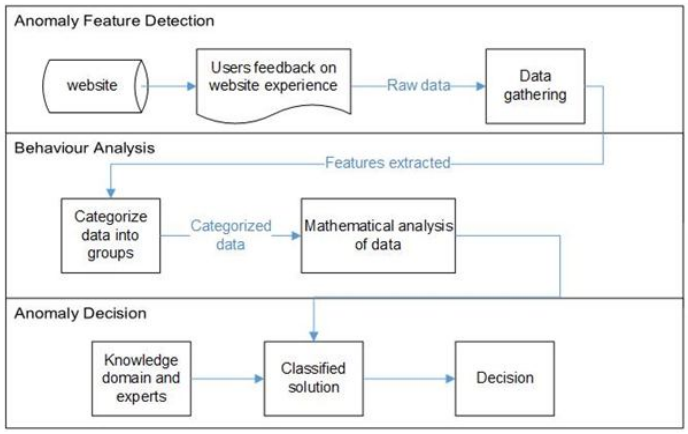

Figure 2: Anomaly detection framework showing the stages involved

5.1 Components of the Anomaly Detection Framework

This section explains the components of the anomaly feature detection framework.

5.1.1Anomaly Feature Detection Stage

This is the stage where the anomaly features were extracted from websites user's feedback.

5.1.2 User Feedback on website experience

The websites listed in figure 3 features are extracted based on the abnormality and documented as user feedback. Here the data generated form user feedback is cleaned for feature extraction in the Behavior analysis stage

5.1.3 Behavior Analysis

This is a term or discipline which is concerned in applying basic techniques used to change social behavior significance. It is also known as behavior modification. The data generated from anomaly feature detection is passed into the behavior analysis for feature extraction and classification. Mathematical analysis is done on the data for classification and clustering. The end result of the behavior analysis stage is the clustered data.

5.1.4 Categorize data into groups

The data generated from the anomaly feature detection group is progressed into the behavior modification stage. The data is categorized as Low, medium and high. The main challenge of the categorization is the identification of the grouping criteria. The classification is based on the impact and the traffic of the site. The traffic on the website considered is key to grouping the website. A website with high traffic will be categorized under high anomaly.

5.1.5 Mathematical analysis of data

Machine learning algorithms is used on the classified data. The existing dataset and feature extracted is used to generate a predictive model for use in classification of future anomaly and recommender system. Supervised anomaly detection algorithms are used to analyze the dataset

5.1.6 Anomaly Decision

The clustered data from the behavior analysis is passed into the anomaly detection phase, the information generated from the Domain Expert is used to classify the clustered data for the recommender system. A recommender system gives advice on solutions for the website anomaly

5.2 Data Sampling Method

Stratified random sampling method will be used to select websites under the identified classification. Stratified random sampling ensures that each subgroup of a given population is adequately represented within the whole sample population of a research study.

5.2.1 Selected websites

The section identifies all the websites to be used in this research. The classification of websites under this research are entertainment, business /economy, education, financial services, government /legal agencies and the health sector as shown in table 1

Table 1:– List of all website to be considered in the research

Entertainment	Business / Economy	Education	Financial Services	Government/Legal	Health
Linda Ikeji blogspot	NigeriaWorld -- Business & Economy	Federal Ministry of Education	Rosabon Financial Services: Home	Law Nigeria – Nigerian Legal Resources	Nigeria Health Watch — Information, Insight and Intelligence on the ...
Naij	BusinessDay	Welcome to Babcock University	Libra Reliance Finance: Financial Services in Lagos, Nigeria	Legal Aid Council of Nigeria	HealthNewsNG - HealthNewsNG
TooXclusive	Nigeria's No1 Economy and Financial Information Hub	Bowen University - Home	Investment One Financial Services	Foward Thinking Investment ...	The State House, Abuja | statehouse.gov.ng	Nigerian Medical Association – National – The Largest Medical
Nairaland	BizWatchNigeria.Ng | All Your Industry News At a Click	Covenant University: Home	GTBank	Nigerian Government Websites	Welcome to Healthcare Federation of Nigeria
Olodo Nation	The Nigerian Economic Summit Group	OAU - Obafemi Awolowo University	Central Bank of Nigeria | Commercial Banks	FRSC Official Website — Creating Safe Road in Nigeria	Digital Health Nigeria

6. FRAMEWORK IMPLEMENTATION AND TESTING

The framework shown in Figure 2 was used as the basis for analysis. The web anomalies considered for the framework testing are broken links and non-responsiveness.

6.1 Broken Links Analysis

To test for broken links, the links have to be launched into a web browser, but this is tedious task for checking the sites listed in section 5.2.1 and table 1. To enable optimization and save time, Excel macro was developed for the analysis of broken links. To save and optimize the resources, excel table was used to capture all the website links required for the analysis. The macro was written in Excel 2013 and linked to a table that holds all links to the website. A button is added to the excel sheet to trigger the start of the verification. The idea in the macro is to put a call to the website and analyses the result. All HTTP calls have status codes; it is the status code that is evaluated. Status 200 means that the client's request was successfully received, understood, and accepted. Since this research is focused on the GET request, status 200 “OK” which implies that the response was successful. This was used in checking. Any status other than 200 is highlighted RED as shown in figure 6. Snippet of the code is shown below:

```
If MsgBox("Is the Active Sheet a Sheet with Hyperlinks You Would Like to Check?", vbOKCancel) = vbCancel Then
    Exit Sub
End If
On Error Resume Next
For Each alink In Cells.Hyperlinks
    strURL = alink.Address
    If Left(strURL, 4) <> "http" Then
        strURL = ThisWorkbook.BuiltinDocumentProperties("Hyperlink Base") & strURL
    End If
    Application.StatusBar = "Testing Link: " & strURL
    Set objhttp = CreateObject("MSXML2.XMLHTTP")
    objhttp.Open "HEAD", strURL, False
    objhttp.Send
    If objhttp.statustext <> "OK" Then
        alink.Parent.Interior.Color = 255
    End If
Next alink
Application.StatusBar = False
On Error GoTo 0
MsgBox ("Checking Complete!" & vbCrLf & vbCrLf & "Cells With Broken or Damaged Links are higlighted in red.")
```

Figure 3: Links of all the webpages to be tested

Figure 3 shows the list created in the excel macro sheet that holds the links of websites specified in section 5.2.1 that were tested for broken links

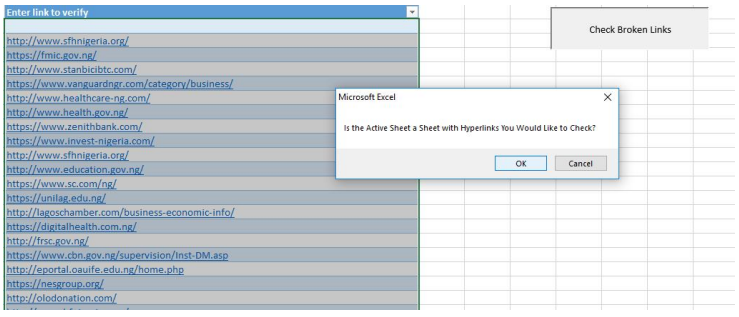

Figure 4: Broken Links Step by Step Testing

Figure 4 shows the start of the test for broken links. The ActiveX button is the trigger for the execution of the macro. For verification, a message box is displayed. The

"CANCEL" button exits the transaction while "OK" commences the execution of the macro

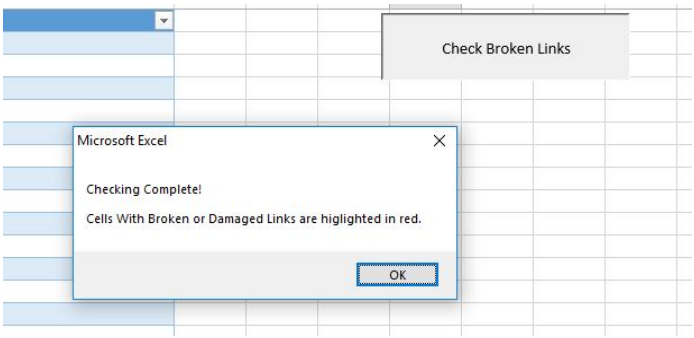

Figure 5: Execution of Testing for Broken Links

During the execution, the status bar shows the link to the site it is currently checking. For a successful check, the macro is been launched with internet access. Once the check is complete, a message box is displayed to notify the user of completion.

http://nigeriaworld.com/news/business/
http://lindaikeji.blogspot.com/
https://zoompf.com/2012/0/lose-the-wait-optimizing-gif-images
http://jpegclub.org/jpegtrane/
http://pmt.sourceforge.net/pngcrush/
http://www.cerias.purde.edu/
https://en.m.wikipedia.org/wiki/Dynamic_HTML
https://en.m.wikipedia.org/wiki/JavaScript
https://en.m.wikipedia.org/wiki/Cascading_Style_Sheets
https://en.m.wikipedia.org/wiki/Netscape
https://en.m.wikipedia.org/wiki/Microsoft
https://en.m.wikipedia.org/wiki/Browser_wars
https://en.m.wikipedia.org/wiki/Netscape_Communications
https://en.m.wikipedia.org/wiki/W3C
https://en.m.wikipedia.org/wiki/Mosaic_browser
https://en.m.wikipedia.org/wiki/Eric_Bina

Figure 6: Broken Links Analysis Test Results.

The links are analyzed based on their HTTP status code. The highlighted cells are links that didn't return a status code 200 which means not OK

It was observed that 10% (6 websites out of a population of 67) of the selected websites returned a red highlight which signifies a bad request.

6.2 Non Responsive Analysis

The invention of smart phones has evolved the viewing of websites from traditional view to the use of mobile devices. To determine that a website is responsive or not, a web browser using mobile view and a mobile phone browser was used.

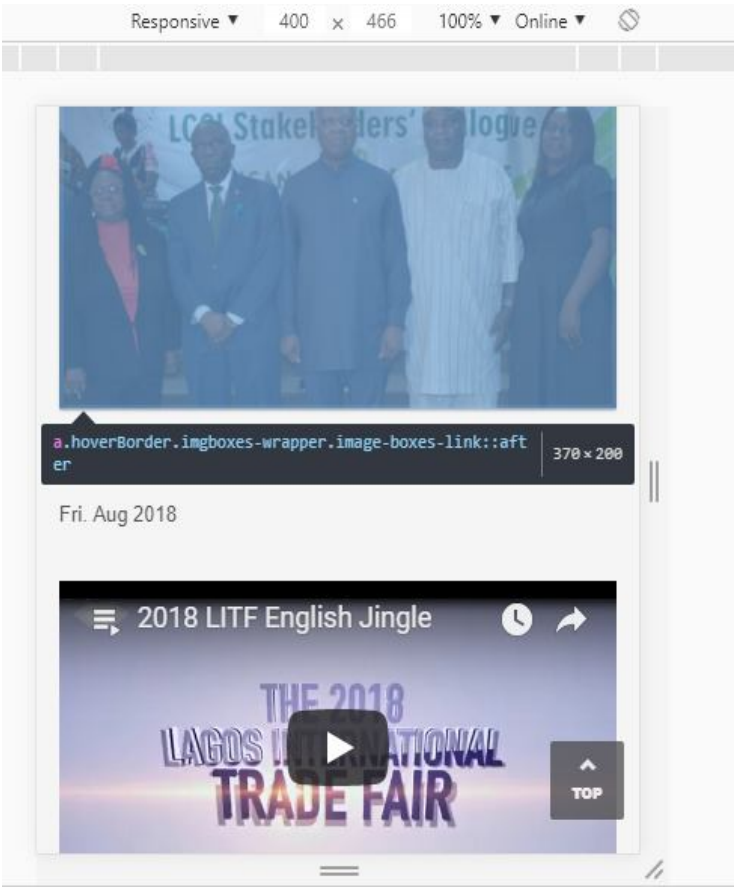

Figure 7: Test for Responsiveness

This check for a responsive site is done through the use of a mobile device or a web browser with a mobile view. The web browser mobile view is selected as the choice of checking as shown in figure 7. The result of whether the website is responsive or not is recorded in a spreadsheet and percentage calculated. The manual process of checking was favored because even though HTML5, CSS3 and some responsive components are used in a website doesn't imply that the website is responsive. Some components of a site might be responsive while some aren't. For example, a site might text might be responsive but the pictures and/or videos will overflow. It was discovered that 17% of the selected website views were not responsive.

7. CONCLUSION

The growth of the World Wide Web over the past decade has grown exponentially in all facets i.e. users, sites, design etc. With this growth come vulnerabilities of websites, incompatibility with devices, lack of contents etc. Website anomalies can cause breaches which can lead directly/indirectly to fraud, identity theft, regulatory fines, brand damage, lawsuits, downtime, malware propagation, and loss of customers. It is evident from this research that web anomalies affect users experience and functionality of websites.

8. SUGGESTION FOR FUTURE WORK

This study focused on designing a conceptual framework for website anomalies behavior analysis and the framework was tested with two anomalies namely broken links and non-responsiveness. This study could be extended in future research to test all the phases of the framework, explore adaptive and selective monitoring of the anomalies and then designing a recommender system to proffer solution to the anomalies.

REFERENCES

1. Antal, B., Bogers, B., & Kunder, M. (2016). Estimating Search Engine Index Size Variability: A 9-year Longitudinal Study. *Scientometrics, vol. 107, no 2.*, 839-856.
2. Berners-Lee. T(2012). *Longer Biography*. Retrieved from W3.org: http://www.w3.org/People/Berners-Lee/Longer.html
3. De-la-Pena-Sordo, J., Pastor-López, I., Ugarte-Pedrero, X., Santos, I., & García, B. P. (2014). Anomalous User Comment Detection in Social News Websites. *International Joint Conference SOCO'14-CISIS'14-ICEUTE'14*, 517-526.
4. Evans, D. (2011). The internet of things: How the next evolution of the internet is changing everything. *CISCIO White Paper, vol. 1*, 1-11.
5. Hodge, Victoria J., and Jim Austin. "A survey of outlier detection methodologies." Artificial Intelligence Review 22.2 (2004): 85-126.
6. Howell, J. W. (2018, April 9). *Anomaly Detection in Azure Stream Analytics*. Retrieved from github.com: https://github.com/MicrosoftDocs/azure-docs/blob/master/articles/stream-analytics/stream-analytics-machine-learning-anomaly-detection.md
7. Isham, M. (2013, April 22). *5 Common Causes of Slow Website Performance*. Retrieved from Zoompf's Web Performance Blog: https://zoompf.com/blog/2013/04/top-5-causes/
8. Ilya Grigorik 2012: Web Performance Anomaly Detection with Google Analytics https://www.igvita.com/2012/11/30/web-performance-anomaly-detection-with-google-analytics/ *November 30, 2012 retrieve on 112062021*
9. Kim, W. (2017, November 2). *The 5 Most Common Problems of Business Websites - The #Hashtag*. Retrieved from www.atlanticwebworks.com: https://www.atlanticwebworks.com/blog/5-common-problems-business-websites/
10. Niederst, J. (2006). Web Design In a Nutshell. *United States of America: O'Reilly Media,* 12-14.
11. Ravneet, k and Sarbjeet S (2016). A survey of data mining", Egyptian Informatics Journal (2016)17,199-216.
12. Swint, R (2019). What is a Website Performance Anomaly? Retrieve from https://www.yottaa.com/website-performance-anomaly/
13. Security,W. (2015). *Website Security Statistics Report 2015*. Retrieved from whitehatsec.com: https://info.whitehatsec.com/rs/whitehatsecurity/images/2015-Stats-Report.pdf
14. Ahmad, S and Purdy, S. 2016: Real-Time Anomaly Detection for Streaming Analytics Published 2016 Computer Science ArXiv
15. Zainuddin N.B. Abdollah, M.F.B, Yusof, R.B and Sahib, S.B.(2014) A Study on Abnormal Behaviour in Mobile Application. Open Access Library Journal, 1: e1229. http://dx.doi.org/10.4236/oalib.1101229
16. Zook, C. (2017, November 2). *6 Common Website Problems in 2015 (and How to Fix Them)*. Retrieved from WebpageFx: https://www.webpagefx.com/blog/internet/6-common-website-problems-in-2015-and-how-to-fix-them/
17. Zwicky, E., Cooper, S., & Chapman, D. (2000). Building Internet Firewalls. *United States: O'Reily & Associates*, 804.