IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

Mobile Cache Memory Optimization using Noise Reduction

P. Amudha Bhomini[1*], Jayasudha J.S[2]

[1]Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-627012, India
Department of Computer Science and Engineering, Sree Chitra Thirunal College of Engineering, Trivandrum, India

[*]*Corresponding Author: amudhabhomini@yahoo.co.in*

Available online at: www.ijcseonline.org

Accepted: 21/Nov/2018, Published: 30/Nov/2018

***Abstract*-** Web pages not only contains useful information, but also many features to improve readability and presentation which end up distracting the relevant content as well as occupying more precious memory space. It becomes even more problem while stored in limited mobile cache and prefetch area. While caching and prefetching or when a page is used repeatedly these unnecessary content, called noise such as banner, advertisements, copyright, background images and license information etc. occupy more space, bandwidth while it doesn't add any value to the user of actual content. Eliminating such noises helps in overall performance improvement of mobile caching, and perfecting. If such noises are not removed, they will become nuisance in web content mining as well. There are many contents which can be identified as noise and there are many techniques to remove them. This paper identifies and removes irrelevant noises in web pages such as background images, search panel, copyright, license information, advertisement. Removing image heavy contents reduces cache memory utilisation, improves performance of content rendering considerably. Care is taken only to remove noises identified and leave the useful contents intact. A brief over view of noise removal and its benefits are discussed in this paper.

Keywords: Noise reduction, web content extraction, caching, pre-fetching.

I. INTRODUCTION

Exploding growth of web content available in World Wide Web contains both core information as well as non-essential information which are there to improve user experience. However, with smart phones being available with everyone, with the challenge of making relevant contents available quickly to the mobile users with limited resources, necessitates that unwanted contents are removed and only relevant contents are presented. These non-core content called noise also occupies precious mobile cache and prefetch memory. Hence it is essential to identify and remove noises while keeping the useful content. Irrelevant details in webpage causes serious issues in web mining too [1]. Hence it is important to identify and remove all such noises.

Noise refers to less important, irrelevant or unnecessary information, which is not part of the main content of a web page [2]. In web environment noise can be categorised in to two groups [3, 4]. They are local noise and global noise. Local noise is also known as intra page noise. It deals with noisy elements within a web page. It includes advertising segments, banners, background images, copy right information, licence etc. Global noise is also known as inter page noise. It includes objects like mirror sites, redundant web pages and old version web pages.

There are various studies on identifying noises and removing them such as A user centric approach towards learning noise in web data [5], Automatic web page segmentation and noise removal for structured extraction using path sequences [6], Noise removing from web pages using neural network [7], Hyper clique-based (HC) cleaner [8], A comprehensive data cleaning system [9], Outlier detection method [10,11], Distance based [12,13,14], Density based [15] and Clustering based methods for enhancing data analysis with noise removal [16], Clustering high dimensional data sets (CLUTO) [17], Local Outlier Factor(LOF) [18], Document Objet Model (DOM) analysis and Natural Language Processing (NLP) [19] etc.

This paper focuses only on removal of local noises present within a page for mobile cache memory optimisation purposes. Most of the webpage's main informational block is only one third to half of the page, and rest are surrounded by the noises such as advertisements, banners, navigational links, copyright tags, decorative images and privacy statements [20]. Useful content should be extracted without any loss of information and rest of the content which is noise can be removed. In this paper we identify noises one by one and remove them. It improves cache memory and prefetch memory and also improves performance while presenting content in a limited resource environments like Personal Digital Assistants (PDA), handheld devices and mobile screens.

The rest of the paper is organized as follows: Section II deals with identifying and removing noise, Section III describes performance evaluation, Section IV presents the results and discuss the performance evaluation and Section V concludes the paper of the proposed solution.

II. IDENTIFYING AND REMOVING NOISE

Noise removal process has got few steps to identify and remove noise content information from web pages. Minimum four different categories of noise pattern may be present within web pages of any web site. Noise pattern such as banners including search panels, advertisements, decorative noises like back ground images, copy right and privacy notices can be structured by using various patterns of tags. The HTML tags such as <DIV>, <TABLE> etc. are used to detect noise patterns in a web page.

The processes with noise removal are pictorially represented as follows

The following noises are identified and removed.

1. Removal of Background images

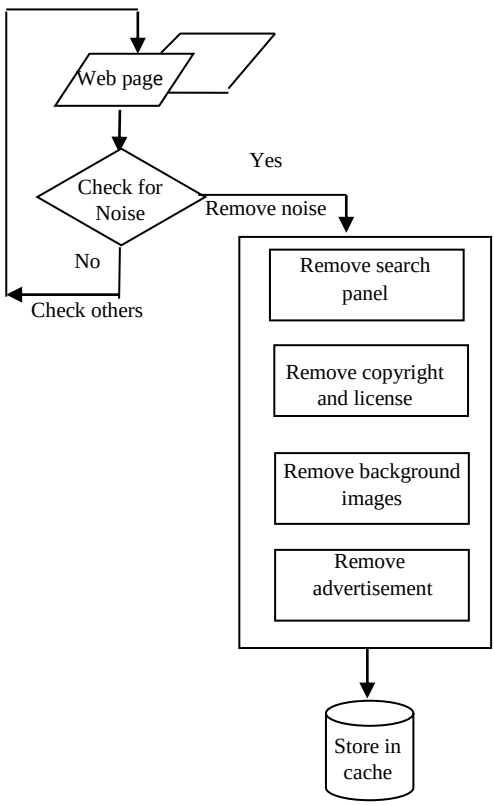

Figure 1. *Steps involved in noise removal*

Background images are the images mainly used for decorative purposes in a web page. The absence of these images does not make any change in the web content. Thus it can be considered as a noise and can be eliminated from the web page. The background images in a web page are identified using table, div, td, tr and style in HTML. They can be cleared by removing the nodes with

- Background attributes of td, table, tr, style, body tag
- Background image attribute of style tag
- Background image feature of style attribute of HTML elements.

2. Removal of advertisements

Advertising is a form of unwanted distraction as it delivers promotional marketing messages to customers. Usually advertisements are images which can be identified using the attributes such as image dimension, name of tag, locations and object tags embedded in anchor tag. The steps involved in removal of advertisement are given in figure 2.

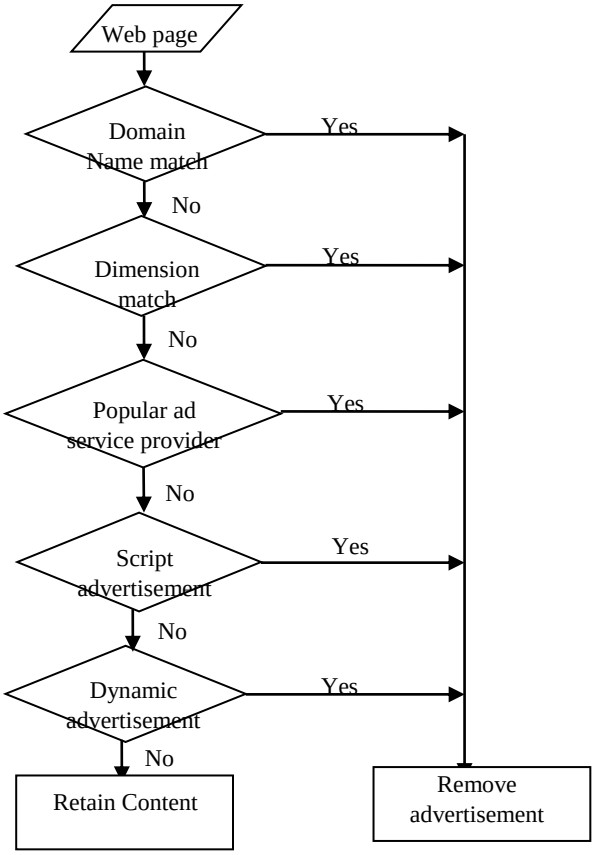

Figure 2. *Process involved in removing advertisement*

The following are the rules used to identify advertisement in a web page [21]

2.1 Based on Domain names

If image URL is different from web page URL then it may be an image advertisement. Relevant or irrelevant images are differentiated from their URL name.

2.2 Based on Image size

Banner advertisements are identified by its size. Image for ads are normally like 468 X 60 pixels, banners e.g. 150 X 500 pixels, 120 X 600 pixels, 160 X 600 pixels

2.3 Predefined list of known Advertisement providers

Maintain a list of popular advertisement providers and block content that comes from them in the list simply by matching URLs against the domain names.

2.4 Advertising by Scripting

Using the <script> tags advertisements can be identified. Certain terms present along with the tag indicates that the advertising elements are added to the web page.

2.5 Dynamic Advertisements

The <INS> tag is used to indicate content that is inserted into a page and indicates changes to a document. INS is semantic tag describing something that is inserted to the text after the text was already published. So it is not from the original author hence it is considered as advertisement.

3. Removal of Search panel

Search panel is identified using the tags for the text box and search button. Search may be associated with web site linked to some other search engine. In either case it is considered as noise and it is removed.

4. Removal of Copyright and License information

Most of the web pages have copyright reserved and license information. These are considered as noise and removed.

III. PERFORMANCE EVALUATION

Reduction of noisy information in a web page reduces the size of source code and size of web page. It also decreases load time of web page as well. If noise is removed before the web page is stored in mobile cache area it reduces cache memory usage and allows more pages without noise can be stored in cache memory. Hence it also reduces network bandwidth. Performance is evaluated in terms of reduction in size of source code, reduction in size of web page and reduction in load time of web page.

i) Reduction in size of source code

The compression ratio of source code of web page is the ratio of size of source code before elimination to the size of source code after elimination. It is calculated using equation 1.

$$CR(SC) = SCBNE / SCANE \quad \text{---------- (1)}$$

Where

CR(SC) - compression ratio of source code
SCBNE - size of source code before noise elimination and
SCANE - size of source code after noise elimination

ii) Reduction in size of web page

The compression ratio of web page is the ratio of size of web page before noise elimination to the size of web page after noise elimination. It is computed using equation 2.

$$CR(WP) = WPBNE / WPANE \quad \text{---------- (2)}$$

Where

CR(WP) - compression ratio of a web page
WPBNE - size of web page before noise elimination
WPANE - size of web page after noise elimination

iii) Reduction in load time of web page

The percentage of reduction in load time of a web page is the ratio of load time of web page prior to noise elimination to the load time of web page after noise elimination. It is calculated using equation 3.

$$LT(WP) = LTBNE / LTANE * 100 \quad \text{---------- (3)}$$

Where

LT(WP) - percentage of decrease in load time of a web page
LTBNE - load time of a web page before noise elimination
LTANE - load time of a web page after noise elimination

IV. RESULTS AND DISCUSSIONS

Noise removal techniques were applied on MSDN website. Steps applied and final result are shown in following figures. The reduction in size is given in the final table.

Removal of Search panel.

Figure 3 shows a web page with search panel. The screen shot of web page with search panel removed is shown in figure 4.

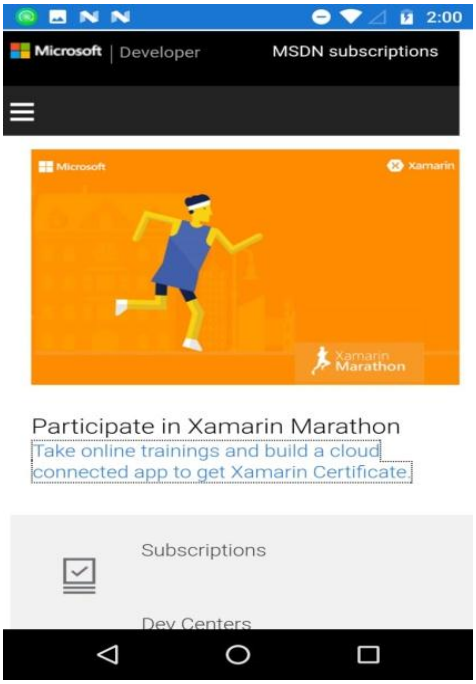

Figure 3. *Web page with search pane*

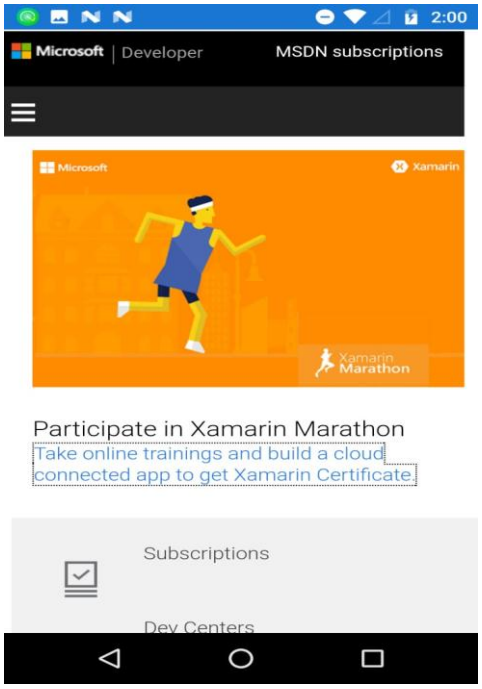

Figure 4. *Web page with search panel removed*

Advertisement Removal

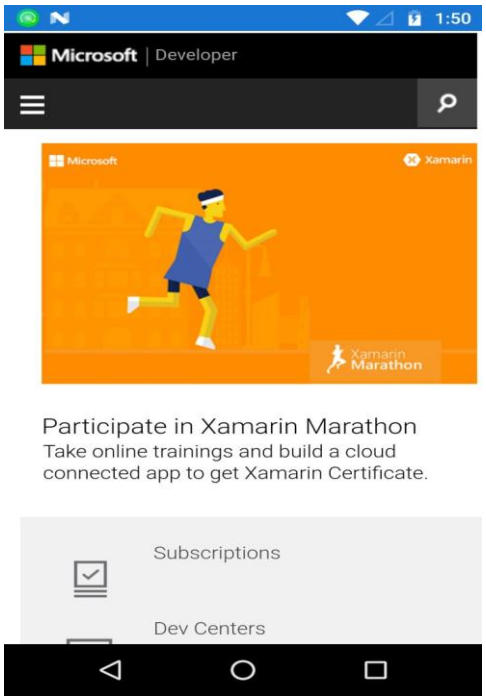

Figure 5. Web page with advertisement

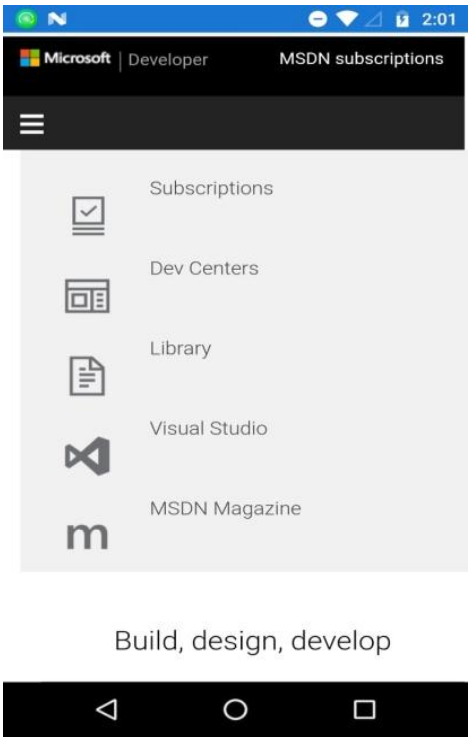

Figure 6. *Web page with advertisement removed*

Figure 5 shows a web page with advertisement. The screen shot of web page with advertisement removed is shown in figure 6

Copy right removal

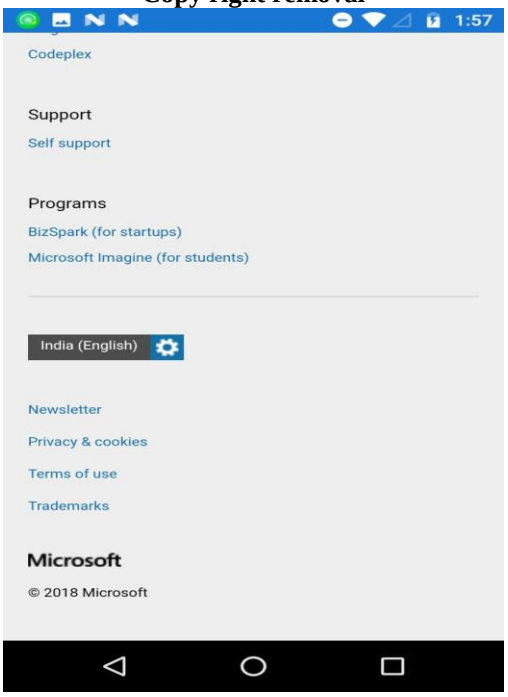

Figure 7. Web page with copyright information

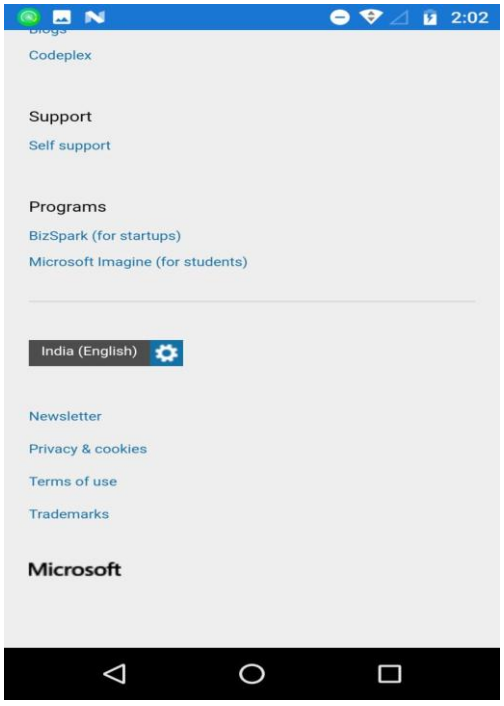

Figure 8 Webpage without copyright information

The compression ratio of source code of few web pages is given in Table 1.

Table 1: Compression ratio of source code of few web pages

Web page	Compression Ratio
msdn	1.05
nasscom	1.60
sify	1.06
wordpress	1.12
bbc	1.29

The compression ratio of web pages is given in Table 2.

Table 2: Compression ratio of web pages

Web page	Compression Ratio
msdn	2.16
nasscom	1.26
sify	1.21
wordpress	4.46
bbc	1.39

Percentage of decrease in load time of web pages is given in Table 3.

Table 3: Decrease in load time of web pages

Web page	Decrease in web page load time (%)
msdn	25.0
nasscom	37.5
sify	44.4
wordpress	37.5
bbc	18.18

Figure 9 shows the analysis of memory optimisation for different web pages. Hence the experimental results shows that removal of noise in web pages reduces bandwidth usage and memory. Thus it reduces latency and web traffic.

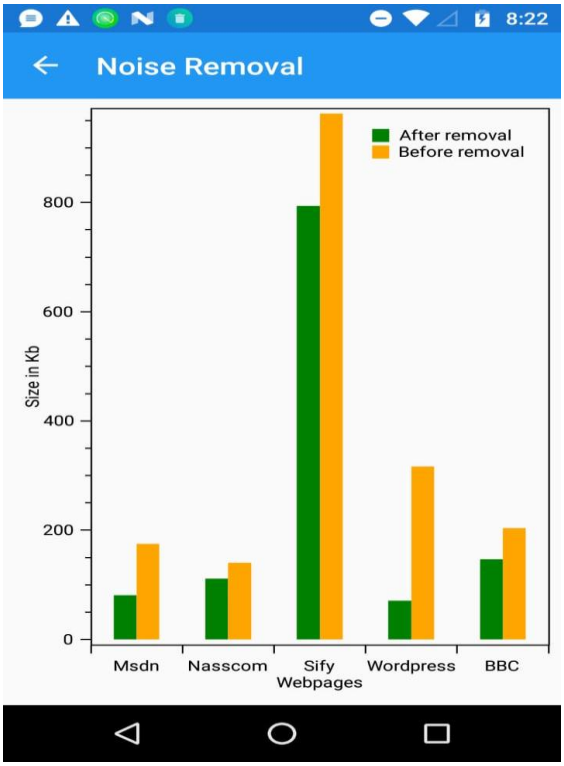

Figure 9. Analysis of memory usage before and after noise removal

Corresponding table for percentage of memory reduction of web pages is given in Table 4.

Table 4. Reduction in memory

Web Page	Before noise removal (kb)	After noise removal (kb)	Reduction in (%)
msdn	175	81	46.3
nasscom	140	111	79.3
sify	963	794	82.5
wordpress	317	71	22.4
bbc	204	147	72.1

V. CONCLUSION

By identifying and removing noise, efficiency has been improved minimum 18% to maximum 44% in terms of load time. Compression ratio achieved varies from 1.21 to 4. 46. Percentage is memory reduction is much higher, ranging from 22% to 82%. It is observed that benefit is more in sites heavy with images and advertisements. However, any reduction in noise helps to improve memory utilisation, load time without compromising on usage.

REFERENCES

[1] Yogita K Patel, Nrendasinh Limbad, "Noise Removal from Web Pages for Web Content Mining", IJAIIE, Vol. 2, Issue 3, pp 2293 – 2299, 2016.

[2] Dauta, S. Paria and D.K. Kole , "Structural Analysis and Regular Expressions Based Noise Elimination from Web Pages for Web Content Mining", International conference on advances in computing, communications and informatics", pp 1445 – 1451, 2014.

[3] L. Yi et al., "Eliminating Noisy Information in Web Pages for Data Mining", Proceedings of 8th ACM SIGKDD international conference on Knowledge discovery and data mining", 2003.

[4] L. Yi et al., "Web Page Cleaning for Web Mining Though Feature Weighting", Proceedings of 18th international joint conference on Artificial Intelligence, 2003.

[5] Julius Onyancha and Valentina Plekhanova, "A User Centric Approach Towards Learning Noise in Web Data", 12th International conference on Intelligent systems and Knowledge engineering, IEEE, 2017.

[6] Roberto Panerai Velloso and Carina F. Doneles, "Automatic Web Page Segmentation and Noise Removal for Structured Extraction Using Tag Path Sequences", Journal of information and data management, Vol. 4, No. 3, pp 173 – 187, 2013.

[7] Thanda Htwe and Khin Haymar Saw Hla, "Noise Removing from Web Pages Using Neural Network", IEEE, Vol. 1, pp 281 – 285, 2010.

[8] H.Xiong, P.N. Tan and V. Kuma, "Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution", Proceedings in 3rd international conference on data mining, IEEE, pp 387 – 394, 2003.

[9] H. Galhardas, D. Floescu, D. Shasha and E. Simon, "An Extensible Data Cleaning Tool", Proceedings in ACM SIGMOD International conference on Management of data, 2000. V.J. Hodge and J. Austin, "A Survey of Outlier Detection Methodologies", Artificial Intelligence Rev., Vol. 22, pp 85-126, 2004.

[10] P.N. Tan, M. Steinbach and V. Kumar, "Introduction to Data Mining", Addison Wesley, 2005.

[11] Angiulli and C. Pizzuti, "Fast Outlier Detection in High Dimensional Spaces", Proceedings of 6th European conference on Principles of data mining and knowledge discovery, 2002.

[12] E.M. Knorr, R.T. Ng and V. Tucakov "Distance Based Outliers: Algorithms and Applications", Very large databases, Vol. 8, pp 237 – 253, 2000.

[13] S.D. Bay and M. Schwabacher, "Mining Distance Based Outliers in Near Linear Time With Randomization and a Simple Pruning Rule", Proceedings in 9th ACM SIGKDD international conference on Knowledge discovery and data mining", pp 29-38, 2003.

[14] S. Ramaswamy, R. Rastogi and S. Kyuseok, "Efficient Algorithms for Mining Outliers from Large Data Sets", Proceedings ACM SIGMOD International conference on management of data, 2000,

[15] Hui Xiong, Gaurav Panddey, Michael Steinbach and Vipin Kumar, "Enhancing Data Analysis with Noise Removal", Transactions on knowledge and data engineering, IEEE, Vol 18, No 3, pp 304 – 319, 2006.

[16] C. Karypis, "Cluto: Software for Clustering High Dimensional Data Sets", 2005.
[17] M.M. Breunig, H.P. kriegel, T. Ng and J. Sander, "LOF: Identifying Density Based Local Outliers", Proceedings of ACM SIGMOD International conference on Management of data, 2000.
[18] P.M. Joshi and S. Liu, "Web Document Text and Images Extraction using DOM Analysis and Natural Language Processing", Proceedings of 9th ACM symposium on Document engineering, pp 218 – 221, 2009.
[19] Hitesh Kumar Azad, Rahul Raj, Rahul Kumar, Harshit Ranjan, Kumar Abhishek and M.P. Sing, "Removal of Noisy Information in Web Pages", ICTCS, ACM, 2014.
[20] H.R. Parmar and J. Gadge, "Removal of Image Advertisement from Web Page", International journal of computer applications, Vol. 27, Issue 7, pp 1-5, 2011.

BIOGRAPHY

P. Amudha Bhomini received the B.Sc. (Computer Science) degree from Madurai Kamaraj University, Madurai in 1990, M.C.A. degree from University of Madras, Madras in 1993 and M.Phil (Computer Science) degree from Mother Teresa Women's University, Kodaikanal in 1999. She is working as an Associate Professor in Computer Applications Department, Nesamony Memorial Christian College, Marthandam since 1999. Her current research interest focuses on Computer Networks, Wirless Communications and Network Security.

Jayasudha J. S. has been working in Sree Chitra Thirunal College of Engineering in the department of Computer Science & Engineering since November 1996. She has 22 years of teaching experience. Now she is working as Professor and Dean (Academic). She has assigned the duty of full additional charge of Principal from 4thMarch 2016 to 18th April 2017 and from 2nd June 2010 to 5thMarch 2011. During her tenure as Principal in-charge, this institution has received National Young Leadership award (NYLP) and Best NSS unit Award in the state in the year 2016. She has organized many community development programmes, short term courses and conferences. She received her B. E. degree from Madurai Kamaraj University and M. E. degree from National Institute of Technology, Trichy and doctorate degree

