ISSN: 2321-7782 (Online)
Volume 2, Issue 9, September 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

KT - Grand: An Algorithm for Web Content Filtering

M.Thangaraj[1]
Associate Professor
Department of Computer Science
School of Information Technology
Madurai Kamaraj University
Madurai – Tamilnadu – India

V.K.T.Karthikeyan[2]
Research Scholar
Department of Computer Science
School of Information Technology
Madurai Kamaraj University
Madurai – Tamilnad – India

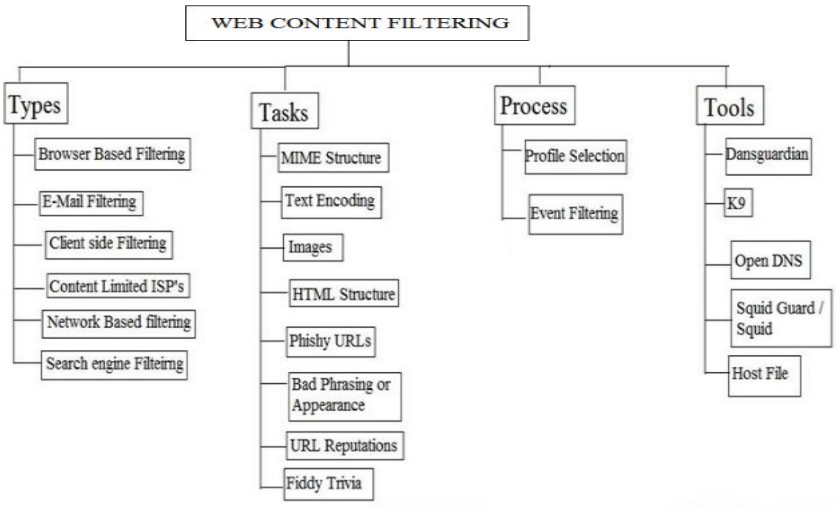
Abstract: ***Normally, Internet accessing are blessed with the mixed ones. It may create serious problem when accessing worst content. According to this incident we need some firewalls to protect ourselves from worst accessing of internet. Web Content Filtering is a type of firewall to block certain web pages from being accessed. There are two different services present in the content filtering: one is web filtering, the screening of web pages and another is e-mail filtering, the screening of e-mail for spam (or) other objectionable content. This paper provides an algorithm "KT - Grand" for filtering the content in webpage. This algorithm can make the analysis in the content part of the web page and make filtering decision either to allow or ban the web page from the access.***

Keywords: Filtering, Screening, keywords searching, String matching, Algorithm analysis.

I. INTRODUCTION

Content Filtering is new subject in the technological area. This issue appears as consequences for the variety of media and advertisement in the internet web sites that lead to unethical and misuse of WWW users. Massive volume of internet content is widely accessible nowadays. One can easily view improper content at will without access control. A modern and effective web content filtering solution scans more than the name of its domain. It is able to break down and analyze web traffic making it capable to accurately pinpoint portions of a web page which should not be allowed into the internal network.

Content Filtering is a type of firewall to block certain sites from being accessed. It usually works by specifying the character and string that, if matched, indicated undesirable content that is to be screened out. Content filtering process and the products that offer this service can be divided into Web filtering as the screening of Web pages, and then e-mail filter as the screening of e-mail for spam or other objectionable content.



There are six types of filtering are present to implement the filters in many different ways, by a software program on a personal computer, via network infrastructure such as proxy servers that provide Internet access, they are Browser based filtering, content limited ISP's, Client-side filtering, E-mail filtering, Search engine filtering and Network based filtering. There are eight types of tasks present in web content filtering they are MIME structure, Text encoding, Images, HTML structure, Phishy URLs, Bad phrasing or Appearance, URL reputations, Fiddy trivia.

There are two types of Process are involved there are Profile Selection - when a collaborative user logs all the profiles which correspond to her or his current context are selected, Event Filtering - it is performed to get only the events which match the selected profiles specifications. There are five types of tools are available in filtering they are Open DNS, K9, Squid guard / Squid, Dansguardian, Host file.

II. RELATED WORKS

This work [30], presented an algorithm for filtering the web content, namely "An Early Decision Algorithm". This algorithm accelerates the filtering process in the web content either to block or pass the web pages. It is implemented using the testing samples of DansGuardian. This work addresses the problem of long delay from text classification algorithms to perform the runtime content analysis of web content. But, it did not combine with more keywords and maintains the URL list.

This paper [31], provides an interface work which is very helpful for parent, teachers and kids to search the web in secure way and make pleasurable interaction. This algorithm makes kids to focus the websites by redirecting the interest into educational side by not blocking or filtering only by redirecting their WebPages. There is no need for blocking, keep kids interest on educational side, safety form security threats, mainly for play psychologically with their child.

Paper [32], presented a comparative study of Naive Bayes classifiers, linear support vector machines, Mathews Correlation Coefficient (MCC)'s prediction. Then, it conducts the experiments on larger data set as TREC05, TREC06, TREC07. They compare the approaches with other commercial and open source anti spam filters such as Bogo filter, Spam Assassin, among others. The Future Work describes the co evolutionary problem of spam filtering, because while the filter evolves to prediction capacity.

III. ALGORITHM

This is an algorithm which makes a filtering decision on the web pages content. Because, massive volume of Internet content is widely access now-a -days. One can easily view improper content at will without access control. For example, a student may watch social websites during laboratory hours in a schools or colleges. Web filtering products can enforce the access control. The current products have adopted content analysis besides the URL-based approach. This algorithm analyzes the Web content to a certain category first, and makes the filtering decision, either to allow or ban the content. The filtering work of this algorithm present a simple, but effective in the form of observation that the filtering decision can be made scanning through the entire content of the web page, as soon as the content can be analyzed into a certain category. The following figure shows the "KT – Grand" algorithm.


```
Allowable Content ← False
Banned Content ← False
n ← 0;
Do
{
Read the entire content
//Skip the "STOP WORDS"
A ← the total number of words in the content.
B ← the total number of keywords with synonyms present in the content.
if (n==100% content)
{
//Scanning the entire content of the web page
for (calculating the categorical value of the web page)
{
CV ← Categorical Value
BV ← Boundary Value
CV ← B / A;
}
// end of for
if (for Banned Category, CV > BV)
{
Banned content: = True;
Exit;
}
if (for Allowable Category, CV < BV)
{
Allowable Content: = True;
Exit;
}
}
// end of if (n==100% content)
While (not end of the content).
```

Fig: KT-Grand algorithm

The content of the web page is get from the corresponding URL of the web pages. This filtering process neglects some words are called "Stops Words", they are single letter words and preposition words. The category of the web content is analyzed by the keywords which are present in the content. According to this investigation, the searching sense of the keywords in the web content is synonymic search. This process considers the synonym words which are corresponding to the keywords. The keywords are also searched in the HTML tags because there are some unwanted or illegal content may inserted in the web pages' HTML tags by the creators which lead to illegal offense.

Then calculate the Number of words in the web content (B) and the number of times that the keywords appeared in the web content (A). The categorical value (CV) of the web page is derived by the simple formula CV=A/B. There is a default boundary value which analyze the web pages category, when the categorical value is above than the boundary value then the web page is belongs to the Banned category otherwise its belongs to the Allowable category. Here, there is no need to maintain the URL list in a database because, Day-by-day the contents of the web pages may changes in a single URL.

A fast decision is particularly important since most Web content is normally allowable and should pass the filter as soon as possible. Finally, the algorithm can allow or block the web pages while accessing according to their value of the webpage category.

IV. IMPLEMENTATION

For a sample, totally 5 pages are randomly taken and extracted the page content through the corresponding URLs in real time (online or offline). Some keywords are selectively approaches to the filter for the process. (Note: According to this sample investigation the selected keywords are assume to be a Bad keywords to ban the pages). The following Table represents the overall values for the given pages and the following figure shows Graphical Representation of the corresponding results.

S .no	No: Of: Key-words given in filter	No: Of: Words present in the Web page (B)	No: Of: Times that the keywords appeared in page (A)	Category Value of the Web page (CV=A/B)	Boundary Value (BV) is assumed by admin	Result (Banned / Allowed)
Page 1	5	435	19	0.0436	0.1	Allowed
Page 2	9	830	35	0.0421	0.1	Allowed
Page 3	12	690	78	0.1130	0.1	Banned
Page 4	14	687	98	0.1426	0.1	Banned
Page 5	7	348	24	0.0689	0.1	Allowed

Table: Overall values for the filtering WebPages

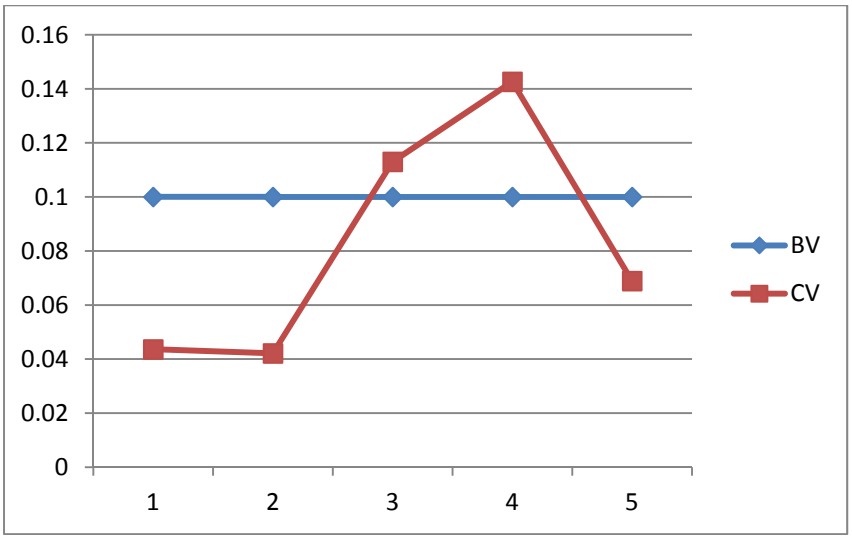


Fig: Graphical Representation of the above results.

The above figure is a Line Graph which represents the measures of content filtering algorithm. According to the graph, CV stands for Categorical Value of the web page which is calculated by the KT-Grand algorithm and BV stands for Boundary Value assumed by the Admin. There are two parameters are used, they are Serial number of the webpage on x-axis and Accurate Value of the web page calculated by the algorithm on y-axis. From the above graphical output the pages 1, 2&5 are allowed to access because values of their pages are lower than the boundary value and pages 3, 4 are banned from access because values of their pages are greater than the boundary value.

V. Conclusion

This paper presents a detail description of implementing a new "KT-Grand" algorithm with the neat survey. This algorithm decides pages to either block or pass the web content as soon as the decision can be made is presented. This algorithm is simple but effective. This algorithm makes filtering work in both an online and offline time content analysis. The philosophy behind this algorithm is to make the filtering decision from the textual part of the Web content. The same rationale behind this method can be applied to other content filtering applications as well, such as anti-spam. The filtering performance of this algorithm is more efficiency than the earlier algorithms.

References

1. T. Almeida, A. Yamakami, and J. Almeida, "Filtering spam's using the minimum description length principle," in Proceedings of the 25th ACM Symposium On Applied Computing, Sierre, Switzerland, March 2010, pp. 1–5
2. T. Almeida, A. Yamakami, and J. Almeida, "Probabilistic anti-spam filtering with dimensionality reduction," in Proceedings of the 25th ACM Symposium On Applied Computing, Sierre, Switzerland, March 2010, pp.1–5.

3. T. Almeida, A. Yamakami, and J. Almeida, "Evaluation of approaches for dimensionality reduction applied with naive bayes anti-spam filters," in Proceedings of the 8th IEEE International Conference on Machine Learning and Applications, Miami, FL, USA, December2009, pp. 1–6.
4. I. Androutsopoulos, G. Paliouras, and E. Michelakis, "Learning to filter unsolicited commercial e-mail," National Centre for Scientific Research "Demokritos", Athens, Greece, Tech. Rep. 2004/2, March.
5. Arasu, A. and Garcia-Molina, H (2003). Extracting Structured Data from Web Pages. SIGMOD-03.
6. A. Bratko, G. Cormack, B. Filipic, T. Lynam, and B. Zupan, "Spam filtering using statistical data compression models," Journal of Machine Learning Research, vol. 7, pp. 2673–2698, 2006.
7. Cao, Jiuxin , Mao, Bo and Luo, Junzhou, 'A segmentation method for web page analysis using shrinking and dividing', International Journal of Parallel, Emergent and Distributed Systems, 25: 2, 93— 104, 2010.
8. J. Carpinter and R. Hunt, "Tightening the net: A review of current and next generation spam filtering tools," Computers and Security, vol. 25, no. 8, pp. 566–578, 2006.
9. G. Cormack, "Email spam filtering: A systematic review," Foundations and Trends in Information Retrieval, vol. 1, no. 4 , pp. 335–455, 2008.
10. G. Cormack and T. Lynam, "Online supervised spam filter evaluation," ACM Transactions on Information Systems, vol. 25, no. 3, pp. 1–11, 2007.
11. Chen, Z., O. Wu, M. Zhu, and W. Hu (2006) A novel web page filtering system by combining texts and images. In WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, pp. 732–735. IEEE Computer Society.
12. Dontcheva, M., S. Drucker, D. Salesin, and M. F. Cohen, Changes in Webpage Structure over Time, TR2007-04-02, UW, CSE, 2007.
13. Du, R.; Safavi-Naini, R.; Susilo, W.; Web filtering using text classification, The 11th IEEE International Conference on Networks, 2003. ICON2003.pages:325 – 330.
14. T. Guzella and W. Caminhas,"A review of machine learning approaches to spam filtering, Expert Systems with Applications" 2009, in press.
15. V.K.T.Karthikeyan, - "Web Content Filtering Techniques: A Survey", - IJCSET 14-05-03-038 Pages 203-208, March (2014)
16. V.K.T.Karthikeyan, - "New Web Content Filtering: An Implementation", -IJSRCSAMS ijsrcsamsv3i4p47, July (2014.)
17. Kim, J. K., and S. H. Lee. An empirical study of the change of Web pages. APWeb '05, 632-642, 2005.
18. Kwon, S. H., S. H. Lee, and S. J. Kim. Effective criteria for Webpage changes. In Proceedings of APWeb '06, 837-842, 2006.
19. D. Losada and L. Azzopardi, "Assessing multivariate be rnoulli models for information retrieval," ACM Transactions on Information Systems, vol. 26, no. 3, pp. 1–46, June 2008.
20. V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with naive bayes - which naive bayes?" in Proceedings of the 3rdInternational Conference on Email and Anti-Spam, Mountain View, CA, USA, July 2006, pp. 1–5.2004.
21. Neha Gupta and Dr.Saba Hilal, - "algorithm to filter and redirect the web content for kids",- IJET 13-05-01-024.
22. Philip B. Stark, "The Effectiveness of Internet Content Filters".
23. Robin van Meteren and Maarten van Someren,"Using Content-Based Filtering for Recommendation".
24. K. Schneider, "On word frequency information and negative evidence in naive bayes text classification," in Proceedings of the 4th Inter-national Conference on Advances in Natural Language Processing, Alicante, Spain, October 2004, pp. 474–485.
25. F. Sebastiani, "Machine learning in automated text categorization",ACM Computing Survey, vol. 34, No. 1 March (2002) 1-47.
26. A. Seewald, "An evaluation of naive bayes variants in content-based learning for spam filtering," Intelligent Data Analysis, vol. 11, no. 5, pp. 497–524, 2007.
27. Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, Maybank S., "Recognition of Pornographic Web Pages by Classifying Texts and Images", Pattern Analysis and Machine Intelligence, IEEE Transactions on, On page(s): 1019 - 1034, Volume: 29 Issue: 6, June 2007.
28. White Paper, Meeting the challenges of Web Content Filtering - March 2007.
29. Y.Yang and X.Liu, "A re-examination of text categorization methods", Proc. of SIGIR'99, 22nd ACM International Conference on Research and development in Information Retrieval (1999) 42-49.
30. Ying-Dar Lin, Po-Ching Lin, Yuan-Cheng Lai. "An Early Decision Algorithm to Accelerate Web Content Filtering", IEICE TRANS. INF. & SYST., VOL.E91–D
31. Neha Gupta, Dr.Saba Hilal, "An Algorithm to Filter & Redirect the Web Content for Kids", IJET13-05-01-024.
32. Tiago A.Almeida and Akebo Yamakami, "Content Based Spam Filtering", Neural Networks (IJCNN), The 2010 International Joint Conference on 18-23 July 2010.

AUTHOR(S) PROFILE

M.Thangaraj, received his post-graduate degree in Computer Science from Alagappa University, Karaikudi, M.Tech degree in Computer Science from Pondicherry University and Ph.D degree in Computer Science from Madurai Kamaraj University, Madurai, Tamilnadu, India in 2006. He is now the ASSOCIATE PROFESSOR of Computer Science Department at Madurai Kamaraj University. He is an active researcher in Web mining, Semantic Web and Information Retrieval and has published more than 50 papers in Journals and Conference Proceedings. He is a senior member of IACSIT. He has served as program chair and program committee member of many international conferences held in India for about one decade.

V.K.T.KARTHIKEYAN, received his Post-Graduate degree in Computer Science in 2013 from Madurai Kamaraj University, Madurai, Tamilnadu, India and M.Phil degree in Computer Science in 2014 from Bharathidasan University, Tiruchirappalli, Tamilnadu, India. He is currently a Ph.D SCHOLAR in Department of Computer Science, School of Information Technology, Madurai Kamaraj University, Madurai, Tamilnadu, India. His research interests include Context-based Search, Web Mining, Information Retrieval, Text Mining, and Content Filtering.

