Web Content Filtering using Cloud Infrastructure

[1]Harishankar Dubey, [2]Nehal Tembhurkar, [3]Kalpesh Pimpalekar, [4]Sarang Lokhande, [5]Rounak Harde
PGDCT, OCE,
RTM Nagpur University, Nagpur, India

Abstract: Nowadays, as the technology is advancing rapidly, the need to learn about it with a significant pace can also be seen. To do so, students take help of the Internet. However, while using internet there are so many issues that obstruct learning and even distract the students. Therefore, developing a system that can help every learner to get what is required, becomes necessary. In this paper we introduce a method to filter irrelevant web content and allowing the access to the content which is relevant with respect to the studies. In this system we use a cloud infrastructure and used websites of different types for our database. We have also introduced an "Ombudsman Feature" to allow Guardian control for monitoring the proper and ethical performance of the application. All these aspects are discussed in the paper in detail.

Keywords—Openstack, Content Filtering Algorithm, Multi-node Cloud Infrastructure, Virtual Machine, Instance, Packets.

I. INTRODUCTION

Today use of the internet is an integral part of learning for learners of every age group, whether they belong to middle school or executives in an organization. We use the internet for almost everything we want to learn, directly or indirectly. With such a significant role on the internet in the field of studies, there are some drawbacks. One drawback is the enormity of the internet; to find what we need is similar to find a needle in a haystack. We have to go through so many websites and web content to get what we exactly want. Another disadvantage of the internet is the number of websites that can be helpful for the students, and there are such website and web content which is distracting and harmful for them. Even if there is a website that can be helpful for a community, it can be offensive for another one. Going down the list, we can also find that there are numerous websites on the internet, and the number is continuously increasing, so we cannot use the technique of blocking websites for filtering the content. There are many other disadvantages of internet. To tackle these issues, we can use Web Content Filtering systems. We mainly focus on the student and learner community that use the internet through networks of schools, libraries, enterprises. In this paper, we propose the idea to use cloud infrastructure for Web Content Filtering, in which we use our content filtering algorithm. This paper is organized as follows: section-1 introduction, section-2 explanation of existing related works and section-3 explanation of proposed methodology.

II. RELATED WORK

There are many of the proposed systems to solve the problem of content filtering, [1] in 2013, Dr. Abbas M. Al-Bakry produce a new Intelligent Web Content Filtering System that used FCM algorithm and some web features to classify the irrelevant websites and to block undesirable websites to reach to end-user. Also, [2] in 2005, N. Churcharoenkrung, Y. S. Kim and B. H. Kang proposed a Dynamic Web Content Filtering on User's Knowledge where the focus of filtering is done on adaptable to the content or domain knowledge changes, uses an incremental knowledge acquisition method called the MCRDR (Multiple Classification Ripple-Down Rules) which is entirely used in the development of commercial knowledge acquisition system. [3] Some of the content filtering methods are based on Artificial Intelligence in 2016 Tanvi Samel, Darshan Vira, Anushka Patil, Durgesh Pandey, Smita Patil gives the theory of Content Filtering Using Artificial Intelligence where they focus on the technique of Bayesian Classifier to filter out the spam content effectively. Naive Bayes classifier uses the Bayes' theorem of conditioned probability to recognize an email to be spam or not. [4] In March 2017, Dr. Sudhir B. Jagtap and Sangita S. Modi came with the idea of Web Content Filtering Using Different Web Mining Techniques in Education System where the use of Information Communication Technology (ICT) for teaching and learning process is highly recommended. There are purposely built filters that analyze each page for inappropriate content before allowing it to the end-user, where parental control is used to put limitations over the use of web access for their child. [5] In 2016 Linbo Tao, Jianjing Shen, Bo Liu, Xiaofeng Guo, proposed the technique of A Privacy Filtering and processing Model in Cloud Computing, they design a classification filtering method on initial data with different processing methods according to their privacy levels, which lay the foundation for the subsequent processing of privacy data.

III. PROPOSED METHODOLOGY

In the proposed methodology we use a Multi-node Cloud Infrastructure. We set up an application server on one of the nodes and one node is required for database server. These two are necessary servers. Other services such as networking, security and management would be installed on a single node.

A. Basic Principle

The content and websites that are stored in the database would only be allowed to access and the application would take reference from the database only. We have used this method because there are numerous websites available on the internet so blocking the web-content and websites would be ineffective.

B. Functioning of the Content Filtering Algorithm.

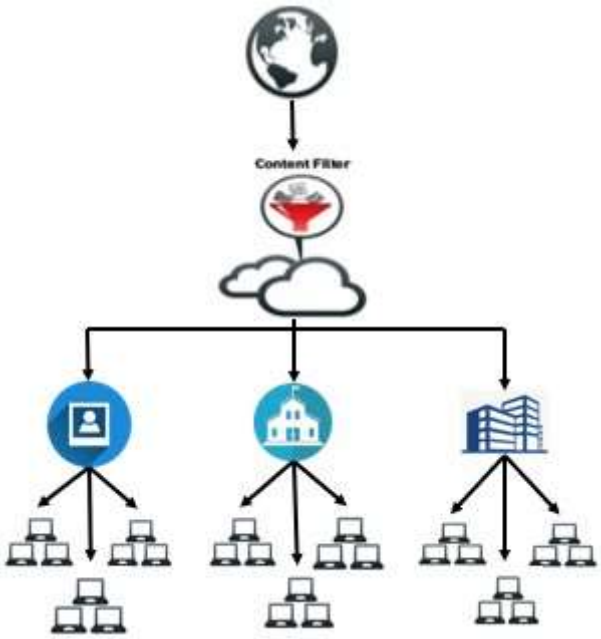


Fig 3.1. Functioning of the Content Filtering Algorithm

There would be an internet gateway through which every packet would pass. This gateway is directly connected to the application server on which the Content Filtering Algorithm is set up. The application server monitors all the traffics that passes through the gateway. The contents and the websites that are already available in the database, could be directly accessed.

The websites that are not available in the database would be first analyzed by the application and if there is anything malicious or irrelevant available in it then the application blocks it. If the website doesn't have any such thing then the application allows the access to it and the website gets stored in the database for next time.

C. Role of Cloud Infrastructure

In schools, colleges, enterprises and organizations there might be several constraints applied to prevent the access to irrelevant and malicious contents, attempts to access those contents can be seen. Due to this the resources could be misused and it may provide a passage to harmful contents such as viruses, trojans and bugs to infiltrate the network. To tackle such issues, we use the cloud infrastructure. We establish a multi-node cloud infrastructure using Openstack.

As more and more records are going digital, we can use that data to ensure the security of the system. As a student gets enrolled to a school, college or library a virtual instance would be allotted to him/her, that could be accessed with the help of Enrollment ID and Password.

As the severs and physical desktops are interconnected through LAN (Local Area Network), there would be no need of internet to login to the portal and access the allotted virtual machine instance. Once the instance is initiated, user can access the internet through it. Hence no one can access the internet without using the content filtering application. Even though, if someone tries to access the barred contents and websites, there is an "Ombudsman Feature" introduced that allows guardian control. Through this feature, the supervisor would be informed immediately and the instance would freeze until admin reset it.

D. Significance of Multi-Node Cloud Infrastructure

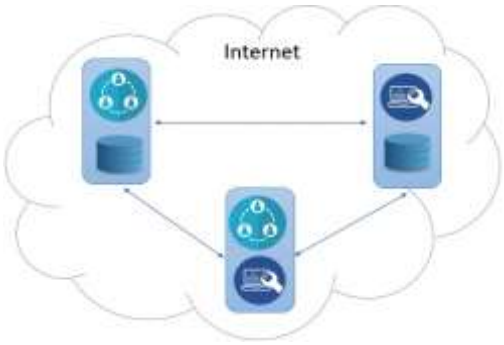


Fig 3.2. Significance of Multi-Node Cloud Infrastructure

We use a Multi-Node Cloud Infrastructure for our system. This is because, with the help of multi-node architecture the availability could be ensured. In this architecture, the burden of the process would be shared as well as we can use other nodes as backup node in case any of the servers goes down.

IV. CONCLUSION

We have studied different papers and came with the conclusion that different strategies are used to filter content that is present on internet for the end user. But still there is some lack which we have studied and came with new idea to solve the problem of filtering the content over the internet. We have discussed the outline of our idea in present paper and the brief working of our project will be stated in next paper.

V. REFRENCES

[1] Dr. Sudhir B. Jagtap and Sangita S. Modi, *Web Content Filtering Using Different Web Mining Techniques, ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering* Volume 7, Issue 3, March 2017.

[2] Ansam A. AbdulHussien, *Comparison of Machine Learning Algorithms to Classify Web Pages, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 11*, 2017.

[3] Tanvi Samel, Darshan Vira, Anushka Patil, Durgesh Pandey, Smita Patil, *Content Filtering Using Artificial Intelligence, ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* Volume 5, Issue 9, September 2016.

[4] Linbo Tao, Jianjing Shen, Bo Liu and Xiaofeng Guo, *A Privacy Filtering and Processing Model in Cloud Computing, 2016 9th International Symposium on Computational Intelligence and Design.*

[5] M. Thangaraj and V.K.T. Karthikeyan, *KT - Grand: An Algorithm for Web Content Filtering, ISSN: 2321-7782 International Journal of Advance Research in Computer Science and Management Studies,* Volume 2, Issue 9, September 2014.

[6] Dr. Abbas M. Al-Bakry and Talib T. AL-Fatlawii, *Intelligent Web Content Filtering System using MAS, 2013. 978-1-4799-2975-7/13.*

[7] W. A. AWAD, *Machine Learning Algorithms In Web Page Classification, International Journal of Computer Science & Information Technology (IJCSIT)* Vol 4, No 5, October 2012.

[8] XIAOGUANG QI and BRIAN D. DAVISON, *Web Page Classification: Features and Algorithms, ACM Computing Surveys, Vol. 41, No. 2, Article 12,* Publication date: February 2009.

[9] N. Churcharoenkrung, Y. S. Kim and B. H. Kang, *Dynamic Web Content Filtering on User's Knowledge, 2005*. 0-7695-2315-3/05.

[10] Richard DAZELEY and Byeong-Ho KANG, *Rated MCRDR: Finding Non-Linear Relationships Between Classifications in MCRDR, 3rd International Conference on Hybrid Intelligent Systems*, December 2003.