Scientific Journal of Impact Factor (SJIF): 4.14
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 3, March -2016

A Review Paper on Analysis of Decisive and Non-Intrusive Technique to Combat Form Spam

Parita Chandan[1], Chintan Thacker[2], Manish Saxena[3]

[1]Department of Computer Engineering, Gujarat Technological University
[2] Department of Computer Engineering, Gujarat Technological University
[3]MCA Department,Feroze Gandhi Institute of Engineering and Technology

Abstract *— Spammers and the spam e-mails are causing huge losses to businesses & individuals on a regular basis in terms of time & money. Spam filtration is an automated technique to identify SPAM or HAM. Various types of spam filters are designed with different approaches, each having their own pros and cons. But after studying various research works it was found that among all, the two most widely used methods are HONEYPOT and CAPTCHA. These two methods are also available in lots of variants. Through this research work, I have tried to analyse and identify best from the above two methods to combat form spam. At first,I studied HONEYPOT and CAPTCHA. Then, I implemented these two methods on various forms. Later, compared and contrasted these two methods on certain parameters, most importantly on time scale basis and tried to find the method which will be able to identify more spams in less time, which is more users friendly, provide easy accessibility, is more secured, and most importantly easily manageable.*

Keywords- *Honeypots, Spambots, CAPTCHA, SPAM and HAM.*

I. INTRODUCTION

Web mining can be defined as the discovery and analysis of useful information from the World Wide Web. Web mining is broadly divided into three categories:

- WEB CONTENT MINING.
- WEB STRUCTURE MINING.
- WEB USAGE MINING.

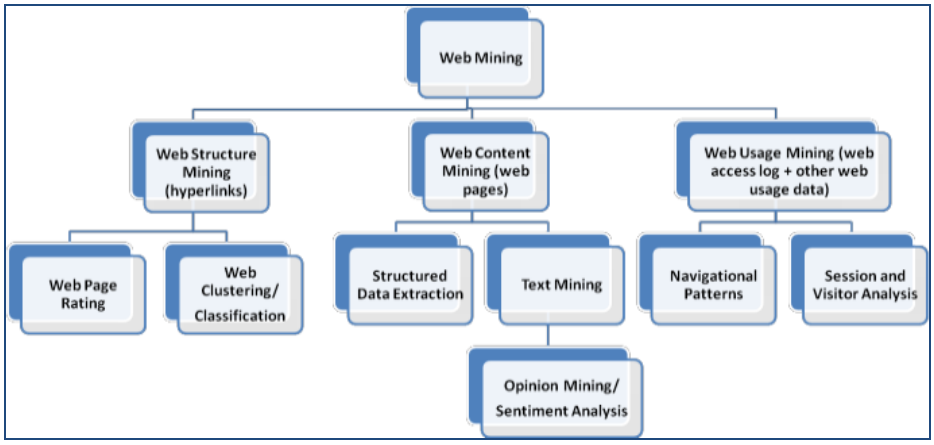

Figure 1. Web Mining Categories

Web Structure Mining focuses on analysis of the link structure of the web and one of its purposes is to identify more preferable documents. Web structure mining helps in discovering similarities between web sites or discovering important sites for a particular topic or discipline or in discovering web communities [17]. The word *spam* means junk mails. The unsolicited emails that are received by any person in his / her mailbox are called spam. These junk mails are usually sent in bulk for advertising and marketing some products. Lots of space in your mailbox is occupied by these junk mails. Sometimes it eats up your valuable space so that the genuine mails are bounced back to the sender if the whole lot of space is occupied by these junk mails. Hence, there comes a necessity to filter out those junk mails from your mailbox [15].

II. BACKGROUND

Figure 2. Spammy Ads

The objective of such spam mails is to retrieve sensitive information of recipients such as, their bank account numbers, passwords, credit card credentials, private information etc. The spam e-mails are causing losses to businesses usually in billions of dollars. Usually spam messages contain a multimedia message for advertising illegal or worthless products, or may be promoting some event, or just propagating some computer malware which is designed to hijack the spam recipient's computer. These spam messages are quite cheap to send such information and even if one in a thousand spam recipient will respond to such messages, the spam sender will be in huge profits [1].

The spam filtering is an automated technique to identity SPAM and Non-Spam, also known as HAM. Generally on the basis of message contents Spam filter is taking its decision about SPAM / HAM, on the basis of sender's and receiver's characteristics [1].

Figure 3. Spam Filtering Technique

The Anti-Spam strategies can be categorized as follows:

Detection based anti-spam strategies attempts to identify spam and remove it or reduce its prominence; whereas Demotion based anti-spam strategies attempts to lower the ranking of spam in ordered lists; and Prevention based strategies attempts to make contribution of spam more difficult by changing interfaces or limiting user actions [1].

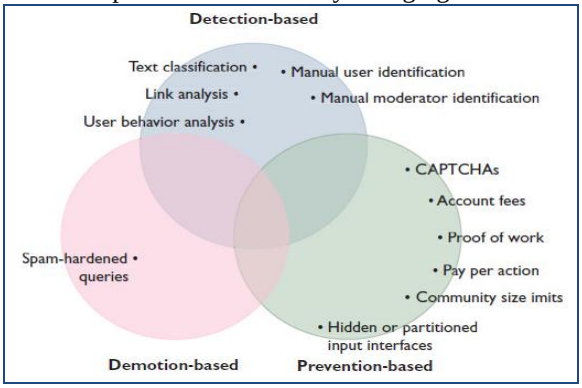

Figure 4. Anti-Spam strategies

III. SCOPE OF RESEARCH

Objective of this research work is to find the most efficient, user friendly and the most secured spam filter technique. This research work is carried out in two phases.

Phase 1: Study of all the available spamizer techniques from the available research work done by previous researchers and finding the two widely use spamizer methods.

Phase 2: Developing an experimental tool to implement the two methods and compare and contrast these two methods on certain parameters, most importantly on time.

IV. PROPOSED METHODOLOGY

I. Check box CAPTCHA

The check box option works by placing a check box on a form which users are asked to select or unselect before submission. Again, this is a simple interaction, and in the context of web forms, check boxes are common place so users should be able to complete it quickly without too much thought. The simplicity of this option means it rates highly on list of CAPTCHA alternatives. Only hesitation would be to make sure the terminology is easy for users to understand. 'I'm not a spambot' is likely to confuse users who don't know what a spambot is [13].

Figure 5. Traditional anti-spam CAPTCHA & Checkbox anti-spams CAPTCHA

II. Puzzle CAPTCHA

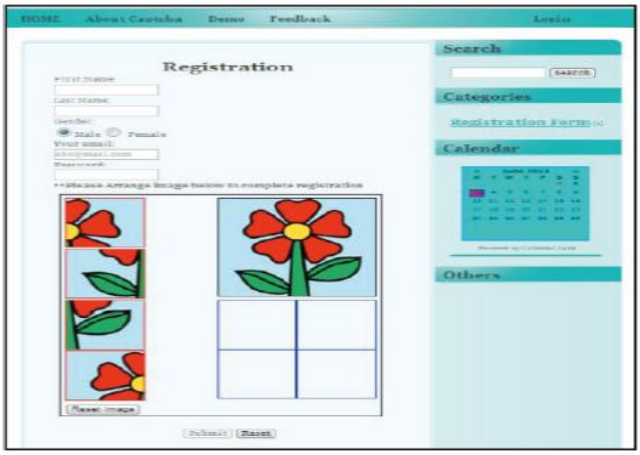

Figure 6. Puzzle anti-spam CAPTCHA

The journeys of process in authentication by using CAPTCHA system based on puzzle begin with interface viewing by a user. A user should drag and drop all the images into the given place and organize it with appropriate sequences. If the all the images had been organized in a right sequence and place, that user will authorize as a human not even a single error will be tolerated. The architecture of CAPTCHA system based on puzzle consists of three main parts, which are collection of images, image processing engine and user interface. The number of collection image must be increased if the potential number of users increases. But this application had been developed with limit for the small access of the system or server [3].

III. Slider CAPTCHA

The slider tool uses the simple interaction of clicking on a button and sliding it from left to right to validate the user as human is likely to be familiar to iPhone or iPad users that need to 'slide to unlock' their device. The tool works because the task is easy for humans to complete while the tool remains invisible to spambots [13].

Figure 7. Slider anti-spam CAPTCHA

It doesn't seem to cover accessibility issues since it requires a mouse or drag gesture to work (no keyboard support). And at large scale, it won't keep spammers out effectively. CAPTCHAs are incredibly cheap rates (moving a slider would be even faster & easier) and it's likely trivial to develop a script that will adjust the slider automatically to submit the form [12].

IV. Socio CAPTCHA

As social media irrespective of the specific platform contains a lot of personal information, this information can be used to produce CAPTCHA [12]. This system can work based on the following assumptions: User is already logged in or social media account is accessible and synchronized with the system accordingly. User/Account holder has provided enough information while creating profile which can be used as Socio-CAPTCHA. Socio CAPTCHA string examples can be used in any CAPTCHA generation type: Friends Name, Book Names, Friends List, Friend's Phone Numbers, Played games, etc. This new proposed scheme will improve accessibility and personalization as personal profile CAPTCHA strings will be used by everyone irrespective of the CAPTCHA type [4].

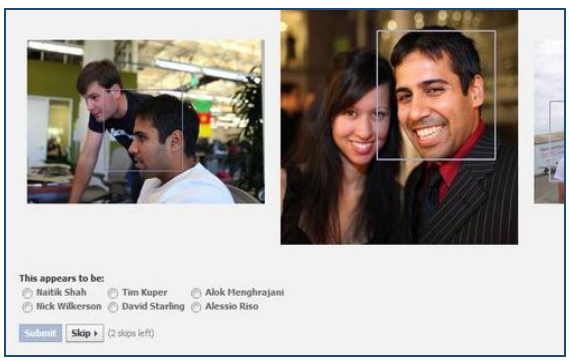

Figure 8. An example of a Socio anti-spam CAPTCHA

V. HONEYPOT METHOD: Using Hidden Form Field

Spam-bots always struggle to read CSS or JavaScript on a webpage. One simple solution is to add a completely junk field in every web form, to hide this field using CSS or Java Scripts. For example:

<input type = "some text" name= "secret" style="display:none;">

Users hate using sites with CAPTCHA. Alternative solutions are available which are not as frustrating as CAPTCHA. The best solutions are those that don't require users to prove they are not spam-bots.

Another CAPTCHA which are less intrusive than traditional CAPTCHAs are honeypots. They can stop some spambots, but not all. They may also create accessibility issues for some users. Honeypot CAPTCHAs work by hiding a text field from users through CSS. It'll only accept entries that leave the field blank. Users can't fill out this field because they can't see it. But spambots will see and fill it in. The form will then reject the spambot's entry. Some spambots have learned to avoid honeypot text fields if they're labelled in a way that tells users to avoid it. This presents accessibility issues for screen reader users who have CSS disabled. If the label on the honeypot field doesn't tell them not fill out the honeypot, they won't know to avoid it. You could give the honeypot field a common label, such as "name", to trick the spambot into filling it in. But it would also trick screen reader users to fill it in too. Honeypot CAPTCHAs are not 100% effective at stopping spambots, nor are they accessible to all users. But they are far better than traditional CAPTCHAs [11].

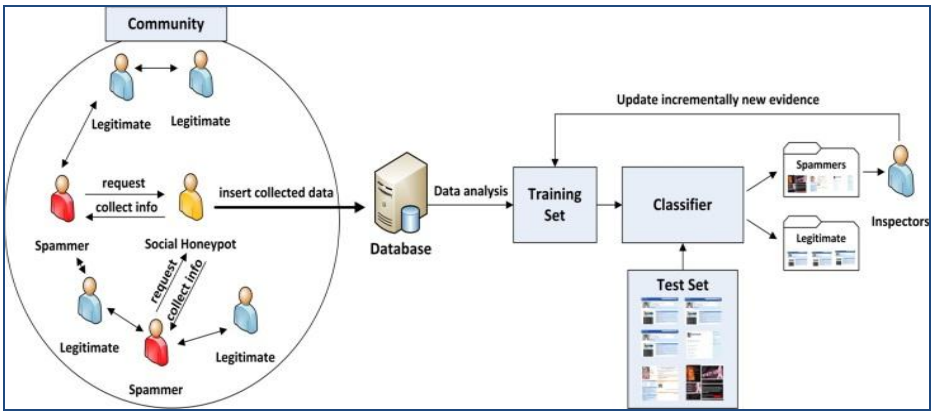

Figure 9. Prevention based honeypot technology

VI. Verified sign-in

Another option for confirming visitors are human is to require them to sign in with an account such as, Facebook, Twitter, etc. This solution is popular for blogs because it includes the side benefit of removing the anonymity that mean-spirited users rely on when they leave rude or offensive comments. Tying comments to a social account adds a level of responsibility that discourages trolls. The obvious problem here, though, is that not all users have the required social account. But there's still one problem remaining: many users aren't comfortable using their social account information to log into an unfamiliar website. They might be concerned that this is an invasion of privacy, or that the website will post updates to their account without their permission [16].

Figure 10. Social log-in functionality via Janrain.

VII. Recording User Time Expenditure

Another technical alternative which is hidden from users is the time-based form. The idea behind this is to detect a spambot based on the time it takes to complete a form. Genuine users take a few moments to complete a form, whereas spambots complete forms instantly. Therefore any forms submitted too quickly would be identified as a bot. We can see this solution working quite well, as long as the time-frame set is practical for users to achieve.In this approach for Spam Detection, recording the 1st Time Stamp when user starts working on a Form, then we are recording the final Time Stamp when user submits a Form [1].

Application Time = Submit Time – Initial Time

There is a Min. Submit Time of every URL containing a Contact Form, stored in a table by the admin, which we can retrieve in parallel while user is filling the form.

This might not be secure enough to stand alone, though, as some of the sneakier bots are programmed to take longer to fill out forms to specifically avoid this trap. Plus, for returning visitors with cookies enabled, the form may auto-populate, causing the visitor to be wrongfully identified as a bot [16].

V. COMPARISON OF DIFFERENT TYPES OF SPAM FILTERING TECHNIQUES

Sr. No	Spam Filtering Techniques	Advantages	Drawbacks
1	Checkbox CAPTCHA	Easy Interaction	Confuses a user who doesn't know what a spambot is.
2	Slider CAPTCHA	Easy Interaction for legitimate user	Doesn't cover accessibility issues.
3	Puzzle CAPTCHA	Spammers cannot easily fill the puzzle	Only works for limited no. of users on a system.
4	Socio CAPTCHA (SCAP)	Full of fun & surprises for the user.	User should have social media account. User/Account holder has provided enough information while creating profile
5	Honeypot Method	Legitimate user need no proof	Honeypots can only track activity that interacts with it.
6	Verified sign-in	Includes the benefit of removing the anonymity.	Not all users have the required account.
7	Recording User Time Expenditure	Easy to implement	Some sneakier bots are programmed to take longer to fill out forms to specifically avoid this trap.

VI. CONCLUSION AND FUTURE WORK

At last, it can be concluded that Honeypot technique and CAPTCHA are the most widely implemented Spam prevention techniques but through our research work done till now, it was found that each of these two suffers from various drawbacks and overheads which make them less efficient and less useful. In our further work, we will integrate these two techniques and try to remove their drawbacks.

We will analyse these methods based on certain parameters to find the most efficient of two in terms of security and user friendliness.

REFERENCES

[1]. Manish Saxena, P. M. Khan," Spamizer: An Approach to Handle Web Form Spam" 978-9-3805-4416-8©2015IEEE, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[2]. Yi Zhou, Kai Chen, Li Song, Xiaokang Yang, Jianhua He," Feature Analysis of Spammers in Social Networks With Active Honeypots: A Case Study of Chinese MicrobloggingNetworks " 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 978-0-7695-4799-2© 2012 IEEE
[3]. Firkhan Ali Bin Hamid Ali, Farhana Bt. Karim ,"Development of CAPTCHA System Based on Puzzle" 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014), September 2 -4, 2014 - Langkawi, Kedah, Malaysia, 978-1-4799-4555-9©2014 IEEE
[4]. Hassan Ishfaq, WaseemIqbal and Waleed Bin Shahid,,Attaining Accessibility and Personalization with Socio-Captcha (SCAP), 978-1-4799-6369-0©20 15 IEEE, roceedingsof 2015 12th International Bhurban Conference on Applied Sciences & Technology (IBCAST) 307 Islamabad, Pakistan, 13th -17th January, 2015
[5]. Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina,"Fighting Spam on Social Websites: A Survey of Approaches and Future Challenges", 89-7801© 2007 IEEE,Published by the IEEE Computer Society
[6]. Tao Men, Yan Sun, Deming Wang, Deming Wang, Mingrong Wang," A Novel Dynamic CAPTCHA Based On Inverted Colors ", 978-1-4799-2716-6©2013 IEEE
[7]. Misako Goto, Toru Shirato, RyuyaUda," Text-Based CAPTCHA Using Phonemic Restoration Effect and Similar Sounds", 2014 IEEE 38th Annual International Computers, Software and Applications Conference Workshops.
[8]. Hsin-Chang Yang, Chung-Hong Lee," Post-Level Spam Detection for Social Bookmarking Web Sites", 2011 International Conference on Advances in Social Networks Analysis and Mining, 978-0-7695-4375-8© 2011 IEEE
[9]. Shailaja Tingre, Debajyoti Mukhopadhyay," An Approach for Segmentation of Characters in CAPTCHA", ISBN:978-1-4799-8890-7©2015 IEEE
[10]. https://solutionfactor.net/blog/2014/02/01/honeypot-technique-fast-easy-spam-prevention/
[11]. http://uxmovement.com/forms/captchas-vs-spambots-why-the-slider-captcha-wins/
[12]. http://www.smashingmagazine.com/2011/03/in-search-of-the-perfect-captcha/
[13]. http://www.experienceux.co.uk/ux blog/2014/03/19/5-alternatives-to-captcha-that-wont-baffle-or-frustrate-users/
[14]. http://pageaffairs.com/notebook/contact-form-honeypots
[15]. http://www.beansoftware.com/Tutorials-Articles-Guides/Anti-Spam-Introduction.aspx
[16]. https://www.usertesting.com/blog/2014/04/09/think-your-site-needs-captcha-try-these-user-friendly-alternatives/
[17]. https://sites.google.com/site/assignmentssolved/mca/semester6/mc0088/14

