Volume 4, Issue 3, March 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Efficient and Scalable CAPTCHA Implementation with No Segmentation Overhead

S.Palani Murugan*, A.Javed Sultan, N.Murali, V.M.Suresh
Department of Information Technology
E.G.S.Pillay Engineering College, Nagapattinam
India

A.Ramesh Kumar
Department of Information Technology
Sree Sowdambika College of Engineering, Arupukottai
India

Abstract— A few years ago, the number of internet users has been on the lower side. But the rapid increase in technologies almost takes internet to all the people of the world who do not need to be aware of computer or any other devices. Nowadays, internet has been accessed through many devices viz., mobiles, laptops etc. The increase in internet users and improvement in technologies has raised a lot of questions on security over use. The people who are using internet may not be aware of what happens behind the closed doors, and more importantly the system/service providers may not be aware of whether human is accessing the service. So it needs a mechanism that tells the service it is human and not machine (robot) otherwise it needs to stop machines from accessing the service. This series of questions evolved to the emergence of human verification system. A mechanism that possibly was the answer for all queries comes in name of CAPTCHA. It was efficient enough but its readability and segmentation are very complex, the former describes user point of view and the latter the system perspective. We propose a simple to use and easy to manage human verification system that uses letters, digits and symbols to comprise the challenge.

Keywords— Human Verification System, CAPTCHA, challenge, reCAPTCHA, OCR.

I. INTRODUCTION

Human verification system is a very sensitive part in the area of web security where it has to go for a lot of scans and vulnerability checks before an algorithm can be adjudged useful. Already the world of web security has evolved with a so called CAPTCHA [3] in specific applications like E-Mail, where an enormous growth of user accounts by now and it needs a mechanism that can take the attackers away from accessing the web sites to make it unavailable or attempting to reduce the performance of a service that runs in. In the environment of web, perhaps, availability of a resource is an important thing as a minute of stall or problem will cause a number of users into a big trouble.

Are you a human? That's the question predominantly plays a vital role in human verification systems. The techniques that are already in use are good enough to protect against robots to create and access e-mail accounts anonymously. Still there is a matter of concern when it comes to some overhead faced by both service providers and users of the web application. The amount of memory that the algorithm or technique consumes to act effectively and how easy it is to the end users to use it easily are the primary check, the former in terms of service provider and latter the end user.

II. RELATED WORK

A. Telling Humans and Computers Apart Automatically

A CAPTCHA is a program that protects websites against bots by generating and grading tests that humans can pass but current computer program cannot. For example, humans can read distorted text as the one shown below, but current computer program can't. The term CAPTCHA (for Completely Automated Public Turing Test To Tell Computers and Humans Apart) was coined in 2000 by Luis von Ahn, Manuel Blum, Nicholas Hopper and John Langford of Carnegie Mellon University

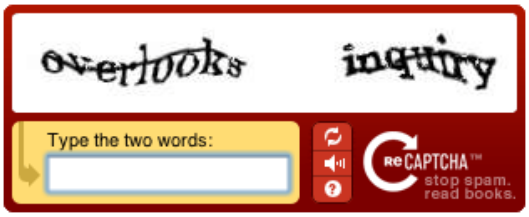

Fig. 1. CAPTCHA

The CAPTCHA requires that the user type the letters of distorted image, sometimes with the addition of an obscured sequence of letters or digits that appears on the screen. Because the test is administered by a computer, in contrast to the standard Turing test that is administered by human, a CAPTCHA is sometimes described as a reverse Turing test. This

term is ambiguous because it could also mean a Turing test in which the participants are both attempting to prove they are the computer.

CAPTCHAs are by definition fully automated, requiring little human maintenance or intervention in administering the test. This has obvious benefits in cost and reliability. By definition, the algorithm used to create the distorted texts must be made public, though it may be covered by a patent. This is done to demonstrate that breaking it requires the solution to a difficult problem in the field of artificial intelligence (AI) rather than just the discovery of the algorithm that is secret, which could be obtained through reverse engineering or other means.

Unlike computers, humans excel at the task of solving CAPTCHAs. While segmentation and recognition are two separate processes necessary for understanding an image for a computer, they are part of the same process for a person. For example, when an individual understands that the first letter of a CAPTCHA is an "a", that individual also understands where the contours of that "a" are, and also where it melds with the contours of the next letter. Additionally, the human brain is capable of dynamic thinking based upon context. It is able to keep multiple explanations alive and then pick the one that is the best explanation for the whole input based upon contextual clues. This also means it will not be fooled by variation in letter.

B. Modern CAPTCHAs

Earlier CAPTCHAs present different variations of characters that are often collapses together, making segmentation almost impossible. The newest iterations have been much more successful at warding off automated tasks. An example of the CAPTCHA challenge in the earlier days is the one that shown in Fig.1. The waviness and horizontal stroke were added to increase the difficulty of breaking the CAPTCHA with a computer program.

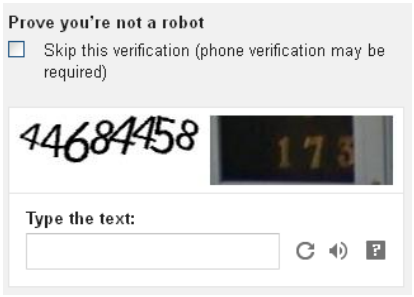

Fig. 2. Modern CAPTCHA (used by Google)

Fig.2 shows the one that has been used Google makes the challenge very complex to break it down but lesser in showing it with waviness and horizontal stroke. The forums and university websites are using a very diverse challenge such as the one shown below.

Fig. 3. Modern CAPTCHA (used by Universities)

The above is another way of authenticating humans but not exactly in the fashion of typical CAPTCHAs yet an effective solution of performing simple arithmetic calculation rather than posing an image that was possibly waved, striped, lined. Although a very diverse and efficient way of authenticating, still it needs users to perform some kind of mathematical operation.

Fig. 4. Modern CAPTCHA (used by blogger)

The above is another effective CAPTCHA mechanism proving humans accessing the website. It's like the previous one, needs to perform arithmetic operation but in a slightly different manner.

Typical applications of CAPTCHA includes preventing comment spam in blogs, protecting website registration, protecting against dictionary attacks, search engine bots, worms and spam. There are many CAPTCHA implementations, some better than others. The following are some strong recommendations before to adapt to viz., Image security, script security, etc.

III. PROBLEM DEFINITION

While CAPTCHA [1] is very useful and widely adapted across the Internet, there are some drawbacks to using it. In fact, even people who are not vision impaired may have difficulty in using CAPTCHA. Sometimes the generated images are just really hard to read. So make sure they are random (e.g. even if the phrase stays the same, then the noise image and/or text placement changes). Also indicate to the user that they can refresh the page so the image is recreated and possibly easier to read. One last note to be aware of, is that CAPTCHA is not totally foolproof [4]. People have written bots that do OCR (Optical Character Recognition) in order to foil these tests. Obviously the more complex the CAPTCHA image become, the harder it is to do text recognition. An alternate for CAPTCHA is already in use in name of phone verification. But it needs another device called phones. People who go through this verification can have access a fewer times than giving access unconditionally.

We propose a less complex and yet effective challenge which is easy to read for the users. Keeping a large number of images needs it to be more robust and vibrant enough storage mechanism. But this proposed approach aims at giving a simple and vibrant technique of providing challenge by means of comprising alphabets, digits as well symbols.

IV. PROPOSED APPROACH

CAPTCHA has been playing a vital role wherever it is used. The problem that lies with CAPTCHA is readability that the users suffer with. Still CAPTCHA is being used predominantly well by number of organizations and service providers after it has gone through lot of changes viz reCAPTCHA, etc. There is some desperate need for some techniques that may not effectively replace CAPTCHA but which reduces the complexity of CAPTCHA.

CAPTCHA uses texts that consist of letters and numbers hidden inside an image to make it complex for machines or computer programs from reading it effectively. It doesn't mean breaking CAPTCHA is impossible but it needs some intelligence to have done it. People use bots that do OCR is still a fair idea. So it is observed that the enormous amount of growth the technology undergone would make CAPTCHA less effective in days to come. So many technologists are trying to prevent CAPTCHAs being breakable by introducing new methodologies and doing changes with already available techniques.

A. Motivation

Artificial Intelligence (AI), the most promising area of research in computer science fancies the chance of making machines to think like humans, behave like humans especially makes decision like humans. Machines are becoming more powerful through robust and potent coding techniques. They can make any system or program vulnerable courtesy to the great AI. CAPTCHAs are very secure in one sense (packed by steganography, hidden text) yet a powerful AI blessed machine or algorithm can still break it.

It needs a very smart technique to make the machine not to be able to read/break the challenge (security) rightly intended to make sure only human is accessing the services. As already mentioned CAPTCHAs are making it hard to read the challenge as they are tilt, wavy or strike. reCAPTCHAs also behaves in the same fashion, keeping two separate images, show them together and segmentation is the process that will needs a lot of work to be put in.

B. Methodology

Here we propose two similar methodologies which slightly differ in ideology. This methodology ideally uses the challenge something like CAPTCHAs but no hidden image or which normally seen tilt, wavy or strike required that almost make segmentation tougher. We use a well defined sequence of characters that obviously consist of letters both upper as well as lower case letters, digits that normally range from 0 to 9 and symbols almost anything available through typical keyboard. A sequence of any combination of letters, digits and symbols makes a challenge. The work for the user is slightly an easier one. They have to enter the digits and letters in the same order as they can see from the challenge and more importantly to leave out all symbols displayed.

Here we give how to explore this, consider the challenge looks like 3$a@5^7F*K, it can be processed as 3a57FK, leaving out the symbols $@^* which are seen from the challenge.

C. Considerations

An algorithm should be checked for its consistency and scalability before to be applied in real time. The user may feel good at accessing the challenge and resolving (reading) it may not be a big deal, but the system/host/server where the algorithm runs must be scalable and consistent enough. Internet users are increasing in numbers rapidly. So it needs an efficient as well as effective algorithm that must dwell upon. Considering CAPTCHA it was a huge task which typically has three phases of work that possibly the result of segmentation to be validated against the one entered by the user after successfully reading the challenge displayed.

Here the process is slightly a different one. Because it has no segmentation process yet parsing is to be done. The text that has been entered by the user after seeing the challenge has to be evaluated against the parsed text which normally has

all letters, digits in the same order but leaving out symbols. Parsing [2] doesn't consume too much time because the process is just a token separation effectively. A simple procedure/function that has been developed for token separation will do the simple trick of comparison between original text and the text entered by the user in response to the challenge displayed.

The technique that is being discussed here is theoretically secure. But whether the technique is computationally secure is the biggest question. Still we believe it needs a lot of work to break this challenge.

V. CONCLUSIONS

Even people own blogs has a feedback form for users comment. The emergence of social networks and forums has impressed the internet users for doing a lot of conversation among people. Amount of information that has been shared through internet is immensely increasing. Making sure that information is from humans and not from roBOTs is where service providers need to be careful of. Once enormous amount of data is produced by BOT, it will surely be an uphill task to manage it. Human verification system, the one, which we have proposed here, is a simple idea to prove human and not BOTs. We conclude, this paper is out of research interest and the interest we got after knowing about human verification system. Through this paper, we do not intend to tell the world of readers, the technology discussed here is computationally feasible. Still CAPTCHA is a fair choice for proving "you are human". The technique may have to undergo lot of experiments to prove it computationally secure.

ACKNOWLEDGMENT

We would like to express our gratitude towards our management, principal, heads and colleague friends for their support and special mention for Google search engine which has been throughout a moral support for us.

REFERENCES

[1] Walsh, Eric (October 28, 2013), "CAPTCHA he cracked by artificial intelligence", mybroadband.co.za. Reuters. Retrieved 27 November 2013.

[2] Chellapilla, Kumar; Larson, Kevin; Simard, Patrice; Czerwinski, Mary. "Designing human friendly human interaction proofs (HIPs)". Microsoft Research.

[3] Engber, Daniel (January 17, 2014), "Who made that captcha?", nytimes.com. NYT. Retreived 17 January 2014.

[4] "Breaking a Visual CAPTCHA". Cs.berkeley.edu. 2002-12-10. Retrieved 2013-09-28. M.Young, The technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

