Volume 3, Issue 9, September 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

CAPTCHA Generation Using Markov Text

Kanika Singhal
Computer Science and Engineering, ITS College
Greater Noida, India

Abstract—Today internet has become need of the hour. Registration is required on every internet site for availing It's Services. Often we come across a distorted image of pseudorandom letters and numbers at the end of registration form. That image is called CAPTCHA (A Completely public Turing test to tell computers and humans apart) .It is a test that humans can pass but computers cannot..In this paper we are presenting a novel approach of captcha generation using markov text . The Captcha image is generated using the concept of probability and then it's analysis is done using bits of assurance.

Keywords— Captcha, Markov text, Bits of assurance

1. INTRODUCTION

Automated programs are always looking for a chance to enter the internet sites and send spam email. This puts an extra burden on the server and the resources on the sites is also wasted. Here comes the security issue. The proliferation of the publicly available services on the Web is a boon for the community at large. But unfortunately it has invited new and novel abuses. Programs (bots and spiders) are being created to steal services and to conduct fraudulent transactions. Free online accounts are being registered automatically many times and are being used to distribute stolen or copyrighted material. Recommendation systems are vulnerable to artificial inflation or deflation of rankings. For example, EBay, a famous auction website allows users to rate a product. Abusers can easily create bots that could increase or decrease the rating of a specific product, possibly changing people's perception towards the product. To avoid this ,CAPTCHA(*A **C**ompletely public **T**uring test to tell **C**omputers and humans apart)* is provided.

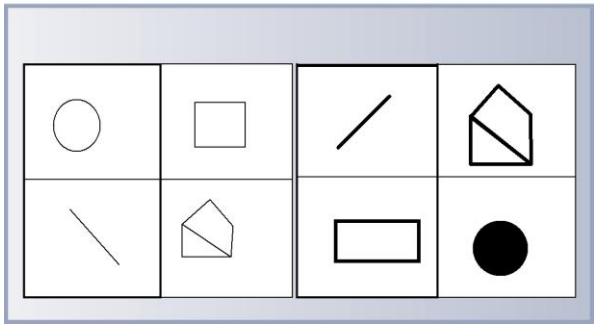
FIGURE1: TYPES OF CAPTCHA

There are several types of Captcha used today. Many OCR and non OCR methods have been proposed. CAPTCHA is now almost a standard security mechanism for defending against undesirable or malicious Internet bot programs, such as those that spread junk email and grab thousands of free email accounts. It has found widespread application on many web sites including Google, Yahoo, and Microsoft's. Captcha was initially devised by Andrei Broder and his colleagues in

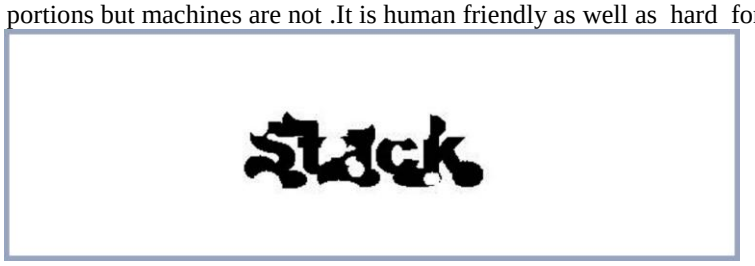
1997 and in the same year AltaVista used this method as a HIP. This method used distorted English word that a user was asked to type. The distorted word was easier for user to understand but difficult for bots to recognize using OCR techniques. Text based CAPTCHAs are inthe form of an image containing a difficult to recognize text string to be identified and typed by the user in a text box provided near the captcha image on the web page. The following summarizes the most popular types.

- **Gimpy**
  Gimpy works by selecting words from dictionary and presents them in distorted manner. To gain entry into the service, user must enter the words. It is easy for humans to guess, but the partially overlapped text can be hard to guess. People who know English has advantage.

- **Pix**
  Pix uses large database of animated objects of everyday life. It has automation and gradability but the labels can be ambiguous. It does not have universality as some objects does not exist in some countries .It is resistant to non effort attacks.



- **Baffle text**
  Baffle text is the extension of gimpy test. It uses gestalt psychology which states that humans are very good in filling the missing portions but machines are not .It is human friendly as well as hard for machines to guess.



- **Pessimal**
  Pessimal print work by combining a word, font and image degradations. People who know English has much advantage. It is easy for humans to guess as well as hard for machines.It is resistant to non effort attacks.

2. Proposed Approach

In this paper we have presented Captcha generation using Markov text. Dictionary strings are easy to read but also easy to guess. Random strings are hard to guess and also hard to read. Markov text lies in between the two approaches. Markov algorithm determines how likely it is that one word is followed by another. These words are not English words and therefore do not yield to dictionary attack. The algorithm work as follows.

1) The text is arranged into the set of prefixes and suffixes.
2) Then an initial prefix is selected and one of the suffix associated with that prefix is chosen at random with probability determined by the input statistics.
3) Then a new prefix is created by removing the first word from the prefix and appending the suffix .
4) The process is repeated until we can't find any suffix for the current prefix or the word limit is exceeded .

Let us assume the text:-

"Can you come with me. Can you accompany "

The text is arranged into the set of prefixes and suffixes.Let us assume that a prefix length of two words and suffix length of one word. The arrangement would be as follows:-

Prefix	Suffix
Can you	come
you come	with
come with	me
with me	can
me can	you
can you	accompany

Then the prefix which has the highest probability, which is repeated maximum number of times is selected.("can you ").Now there are two suffix associated with this prefix.("come" and "accompany").We can choose any of the associated suffix randomly. Let us choose the suffix "come". Now the first word of the prefix is removed and suffix is appended. So the new string formed is "youcome". The process is repeated until we can't find any suffix for the current prefix or the word limit is exceeded.

3. Analysis of the proposed approach

How much assurance do we gain if a CAPTCHA subject correctly identifies a Markov challenge? If the attacker notices that the challenges are all letters, and therefore guesses a letter uniformly, then each letter in a challenge carries lg 26 ~ 4.70 bits of assurance against that attacker. But what if the attacker observes the letter frequencies?

For the King James Bible, the most frequent letter is e at 12.73%, the second most frequent is t at 9.81%, and the least frequent is x at 0.04%.

Suppose that a CAPTCHA chooses a challenge according to a probability vector $p = (p1, p2, ..., pN)$, and that the attacker also chooses a string according to p. Then the attacker's probability of success is the mean probability $(pi)2$. Like all challenges, Markov strings are vulnerable to lucky long shot attacks. Markov strings provide 4.70 bits of assurance per character against an attacker who uniformly guesses one of 26 letters, 3.86 bits against an attacker who guesses letters with the appropriate frequencies, and 2.97 bits against an attacker who guesses common letters. An engineering approximation and experiment together indicate that the average character gives about 2.07 bits of assurance against an attacker who guesses a string produced by the same Markov model. Most common string that occurs with probability is as follows:-

Most common String	Probability of occurence
andth	341 times

That observation reduces the number of bits of assurance per character from 2.07 to just 1.64. Other attacks yield similar reductions in assurance.

3. Conclusion

Security has been of great importance last few decades. Users fill all the entries of the form for the registration on the internet sites. User fills all the required entries before submitting the Captcha. The string of captcha is generated by many methods such as dictionary method,random string method. . Dictionary strings are easy to read but also easy to guess. Random strings are hard to guess and also hard to read. Markov text lies in between the two approaches. In this approach the string is generated using the probability of suffix. So it will be a great challenge for the bots to determine how the captcha string is generated. This will add more to the security.

References

[1] Prem Shankar Yadava, Chandra Prakash Sahu, Sanjeev Kumar shukla, "Time variant captcha: Generating strong captcha security by reducing time to automate computer programs", journal of emerging trends in computing and information sciences.vol2 no. 12 december 2011.

[2] http://golang.org/doc/codewalk/markov.go?h=os
[3] Kwan Woo Park, "Analysis of Captcha", Computer Science Department University of Southern California Los Angeles, CA 90089-0781 USA.
[4] Clark pope and khushpreet kaur, "Is it human or computer: defending ecommerce with captchas".
[5] Ahmad Salah El Ahmad, Jeff Yan "The Robustness of a new captcha", school of Computing Science Newcastle University, UK.
[6] Yong Rui, Zicheng Liu, "ARTIFACIAL: Automated Reverse Turing test using FACIAL features", Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA.
[7] Ahmad S El Ahmad, Jeff Yan and Mohamad Tayara, "The Robustness of Google CAPTCHAs".
[8] Jeremy Elson, John R. Douceur, Jon Howell, "Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization", Microsoft Research.
[9] Rich Gossweiler, Parkway Mountain View, Maryam Kamvar Google, Shumeet Baluja Google, Inc. 1600 Amphitheatre Parkway Mountain View,Philip Brighten godfrey, "What's Up CAPTCHA? A CAPTCHA Based on Image Orientation".

