WORD GROUPING CAPTCHA-A NOVEL APPROACH FOR SECURING WEB SERVICES

DAYANAND

Department of Information Technology, Birla Institute of Technology, Mesra, Ranchi, India
dayanand1003.11@bitmesra.ac.in

Abstract: "CAPTCHA" stands for Completely Automated Public Turing Test[1] to Tell Computers and Humans Apart[2]. Due to exponential growth of internet, security of web application has become a vital issue and many web applications facing a threat of web bots also known as internet Robot is an automated script which executes over the web forms and occupy web spaces and thus increases network traffic. The problem with current text based captcha (most popular captcha) systems is that most of them have proven to be either not robust enough (they have been broken) or they are too complicated or annoying to read even for humans. Word grouping is a type of captcha in which user has to divide the given words in two subgroups. This Paper proposes a solution for improving web security from Web Bots (Robots) by implementing WORD GROUPING CAPTCHA. This paper also discusses captcha evaluation parameter and comparing text based captcha, picture based captcha, word grouping captcha based on evaluation parameters.

General Terms Security, Human, Tests

Keywords: Text Based Captcha, Picture Based Captcha, Word Grouping

I. INTRODUCTION

"CAPTCHA" stands for Completely Automated Public Turing Test to Tell Computers and Humans Apart. If someone wants to sign up for a free email service, before he can submit web form; he first has to pass a test. The test is not hard. For human, the test should be simple and straightforward. But for a computer, the test should be almost impossible to solve. CAPTCHAs are now almost standard security mechanisms for defending against undesirable and malicious bot programs on the Internet. CAPTCHAs generate tests that most humans can pass but not a computer program. CAPTCHA challenges are based on hard, artificial intelligence. The term "CAPTCHA" was coined in 2000 by Luis Von Ahn, Manuel Blum, Nicholas J. Hopper (all of Carnegie Mellon University, and John Langford (then of IBM). The most commonly used CAPTCHAs are text-based, in which the challenge appears as an image of distorted text. Bots can sign up for thousands of accounts a minute with free email service providers, send out thousands of spam messages in an instant, or post numerous comments in blogs pointing both readers and search engines to irrelevant sites, so CAPTCHA is required to differentiate between a bot and a human. CAPTCHAs are used because of the fact that it is difficult for the computers to extract the text from such a distorted image, whereas it is relatively easy for a human to understand the text hidden behind the distortions. Therefore, the correct response to a CAPTCHA challenge is assumed to come from a human and the user is permitted into the website.

Captcha are sometimes called "reverse Turing tests": because they are intended to allow a computer to determine if a remote client is human or not.

Spammers are constantly trying to build algorithms that read the distorted text correctly. So strong CAPTCHAs have to be designed and built so that the spammers cannot harm web security. This Paper proposes a solution for improving web security from Web Bots (Robots) by implementing WORD GROUPING CAPTCHA. This paper will analyse several properties of text based captcha in terms of their effect on security with respect to resistant to automated attacks. This paper will discuss captcha evaluation parameter in terms of consistency, entropy, fun, ease of generation and implementation.

II. BACKGROUND

The need for CAPTCHAs rose to keep out the website/search engine abuse by bots. In 1997, **AltaVista** sought ways to block and discourage the automatic submissions of URLs into their search engines. Andrei Broder, Chief Scientist of AltaVista, and his colleagues developed a filter. Their method was to generate a printed text randomly that only humans could read and not machine readers. Their approach was so effective that in a year, "spam-add-ons'" were reduced by 95% and a patent was issued in 2001.

In 2000, **Yahoo**'s popular **Messenger** chat service was hit by bots which pointed advertising links to annoying human users of chat rooms. Yahoo, along with Carnegie Mellon University, developed a CAPTCHA called EZ-GIMPY, which chose a dictionary word randomly and distorted it with a wide variety of image occlusions and asked the user to input the distorted word.

In November 1999, s*lashdot.com* released a poll to vote for the best CS College in the US. Students from the Carnegie Mellon University and the Massachusetts Institute of Technology created bots that repeatedly voted for their respective colleges. This incident created the urge to use CAPTCHAs for such online polls to ensure that only human users are able to take part in the polls.

Forms of attack [3]:

Whether a captcha is based on pictures, text, sound, or puzzle–solving, certain similarities can be seen in terms of how captcha are attacked by malicious users. Typical attack models are

Bypass attacks- Any attack that circumvents the need to solve the captcha at all. Such attacks are not always a weakness of the captcha itself; they may instead be aweakness of the service using the captcha**.**

Challenge replay attacks-If the captcha system can produce only a limited number of unique challenges, then the automated agent may record all or most of the possible challenges. A human associateprovides a library of correct answers for the challenges. The automated agent can thenreplay the correct answer whenever it is faced with a particular challenge for which it knowsthe correct solution. Some image–based captcha are vulnerable to this weakness, particularly those based upon a finite library of images.

Mechanical Turk attack-Here, the problem of solving the captcha is automatically 'outsourced' to a paid human agent. They immediately solve the challenge and quickly return the answer to the automatedagent in real time. The automated agent then presents the human–provided answer, and is able to programmatically exploit the online resource.

Trivial guessing attack- If there is an unlimited range of challenges, but a very limited range of possible answer (*e.g.*, 'which of these 10 choices is correct?'), a high success rate may be achieved by attacking program by merely guessing randomly from the available answers. Particularly,any graphical captcha that requires the user to select a correct position within an imagebut which has a wide error tolerance for user inaccuracy may be vulnerable to a trivialguessing attack.

Brute force attacks- If there is a somewhat limited range of possible answers, *e.g.*, a numerical 4–digit captcha would have 10,000 possible answers, then it is possible for a distributed group of automated agents to attack the captcha by exhaustively trying answers at random or according to a selected sequence. This differs from the 'trivial guessing attack', in that it relies upon having access to a large number of attacking agents.

CAPTCHAs and the Turing Test

CAPTCHA stands for "Completely Automated Public Turing Test[4] to Tell Computers and Humans Apart". It should be difficult for someone to write a computer program that can pass test generated by CAPTCHA even if they know exactly how Captcha works. CAPTCHAs are like Turing test. In original Turing test, a human judge was allowed to ask a series of questions to two players, one of which was computer and other a human being. Both players pretended to be the human and the judge has to distinguish between them. CAPTCHA are similar to the Turing testing that they distinguish computers from humans, but they differ in that the judge is now computer. A CAPTCHA is an automated Turing test.

It's also important that the CAPTCHA application is able to present different CAPTCHAs to different users. If a visual CAPTCHA presented a static image that was the same for every user, it wouldn't take long before a spammer spotted the form, deciphered the letters, and programmed an application to type in the correct answer automatically.

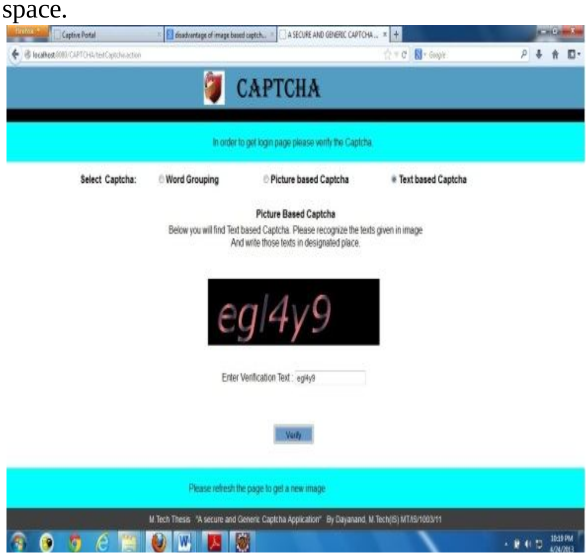
But not all CAPTCHAs rely on visual patterns. In fact, it's important to have an alternative to a visual CAPTCHA. Otherwise, the Web site administrator runs the risk of disenfranchising any Web user who has a visual impairment. One alternative to a visual test is an audible one. An audio CAPTCHA usually presents the user with a series of spoken letters or numbers. It's not unusual for the program to distort the speaker's voice, and it's also common for the program to include background noise in the recording. This helps thwart voice recognition programs.

Another option is to create a CAPTCHA that asks the reader to interpret a short passage of text. A contextual CAPTCHA quizzes the reader and tests comprehension skills. While computer programs can pick out key words in text passages, they aren't very good at understanding what those words actually mean.

III. TYPES OF CAPTCHAS

Text based captcha

The most commonly used CAPTCHAs are text-based where distorted text is displayed. To solve the CAPTCHA, users must recognize the distorted characters and correctly enter them in a designated space.



CAPTCHA

In order to get login page please verify the Captcha

Select Captcha: ○ Word Grouping ○ Picture based Captcha ● Text based Captcha

Picture Based Captcha

Below you will find Text based Captcha. Please recognize the texts given in image
And write those texts in designated place.

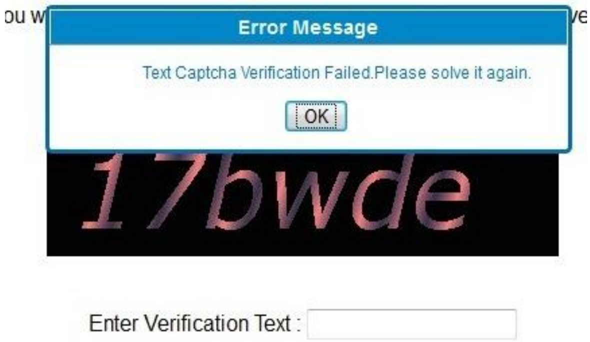
Enter Verification Text : egl4y9

Please refresh the page to get a new image

M.Tech Thesis "A secure and Generic Captcha Application" By Dayanand, M.Tech(IS) MT/IS/1003/11

Captive Portal
disadvantage of image based captch...
A SECURE AND GENERIC CAPTCHA...
localhost:8080/CAPTCHA/textCaptcha.action
Google
eg/4y9
Verify
10:19 PM
6/24/2013

Figure-1



Error Message

Text Captcha Verification Failed.Please solve it again.

Enter Verification Text :

Figure-2

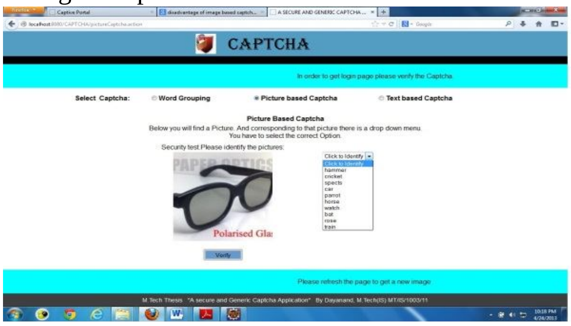
Weakness of Text based Captcha:

The number of classes of characters and digits are very small. When the noise and distortion is added to the text based captcha they often create a problem in recognizing them. Although some alphabets and digits have very different shapes, but when they are distorted, it is become difficult to recognize them. For example 7 may look like 1, ‘cl’ can be confused with d, ‘nn’with m.

In January 2008 article published in informationweek.com claiming Yahoo’s [4] captcha security had been broken. In February 2008 in www.theregister.co.uk claiming that Google’s [5] captcha had been broken by spammers. In May, Microsoft’s [6] captcha security had been broken.

Picture based Captcha[7]

In picture recognition the motivation here is that humans are much better at recognizing picture than computers are, and that perhaps we can use that advantage to make a good CAPTCHA. IdentiPic[8] is photo based CAPTCHA system where user has to identify picture. Pictures are shown and corresponding to each pic there is drop down list having few options.



Captive Portal
disadvantage of image based captch...
A SECURE AND GENERIC CAPTCHA...
localhost:8080/CAPTCHA/pictureCaptcha.action
CAPTCHA

In order to get login page please verify the Captcha

Select Captcha: ○ Word Grouping ● Picture based Captcha ○ Text based Captcha

Picture Based Captcha

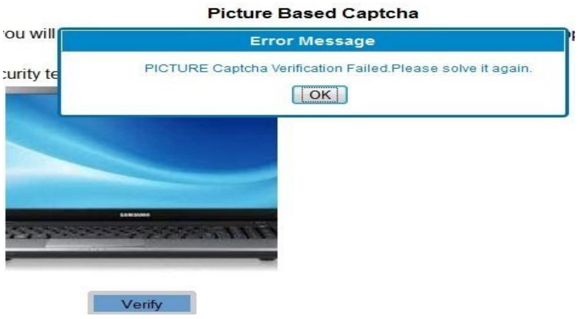
Below you will find a Picture. And corresponding to that picture there is a drop down menu. You have to select the correct Option.

Security test.Please identify the pictures:

Please refresh the page to get a new image

M.Tech Thesis "A secure and Generic Captcha Application" By Dayanand, M.Tech(IS) MT/IS/1003/11

Figure-3



Picture Based Captcha

Error Message

PICTURE Captcha Verification Failed.Please solve it again.

Figure-4

Weakness of Picture based Captcha:

It creates a problem to users having low vision or learning disability [9]. Most of the time object recognition becomes cumbersome due to the ambiguity present in image objects. Instead of Turing test it has become almost an IQ test.

DECAPTCHA [10]

In this section we present our captcha breaker, Decaptcha, which is able to break many popular captchas including eBay[11], Wikipedia and Digg [12]. Then we discuss the rationale behind its five stage pipeline.Decaptcha uses the aForge framework [13] and the Accord framework that provide easy access to image manipulation filters, and standard machine learning algorithms such as SVM [14].Decaptcha pipelinestages are:

1.**Pre-processing**: In this first stage, the captcha’s background is removed using several algorithms and the captcha is binarized (represented in black and white) and stored in a matrix of binary values. Transforming the captcha into a binary matrix makes the rest of the pipeline easier to implement, as the remaining algorithm works on a well-defined abstract object. The downside of using a binary representation is that we lose the pixel intensity. However in practice this was never an issue.

2. **Segmentation:** In this stage Decaptcha attempts to segment the captchas using various segmentation techniques, the most common being CFS [15] (Color Filling Segmentation) which uses a paint bucket flood filling algorithm [16]. This is the default segmentation technique because it allows us to segment the captcha letters even if they are tilted, as long as they are not contiguous.

3. **Post-Segmentation:** At this stage the segments are processed individually to make the recognition easier. During this phase the segments’ sizes are always normalized.

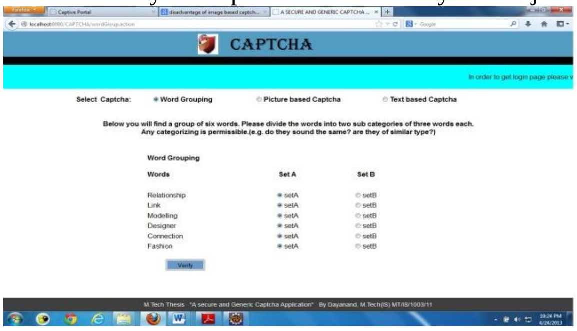
4. **Recognition**: In training mode, this stage is used to teach the classifier what each letter looks like after the captcha has been segmented. In testing mode, the classifier is used in predictive mode to recognize each character.

5. **Post-processing**: During this stage the classifier’s output is improved when possible. For example, spell checking is performed on the classifier’s output for Slashdot because we know that this captcha scheme uses dictionary words. Using spellchecking allows us to increase our precision on Slashdot from 24% to 35% .

IV. PROPOSED CAPTCHA

Word Grouping

Word grouping captcha the user is presented with six words, and is asked to divide the group into two subsets,using any categorizing the user wishes.The words will be easier so that any user can do that.It needn't any expertise in any subject.



Captive Portal
disadvantage of image based captch...
A SECURE AND GENERIC CAPTCHA ...
localhost:8080/CAPTCHA/wordGroup.action
Google
CAPTCHA
In order to get login page please v

Select Captcha: ● Word Grouping ○ Picture based Captcha ○ Text based Captcha

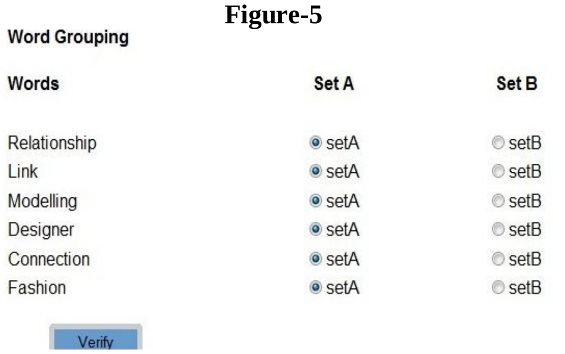
Below you will find a group of six words. Please divide the words into two sub categories of three words each. Any categorizing is permissible.(e.g. do they sound the same? are they of similar type?)

Word Grouping

Words	Set A	Set B
Relationship	● setA	○ setB
Link	● setA	○ setB
Modelling	● setA	○ setB
Designer	● setA	○ setB
Connection	● setA	○ setB
Fashion	● setA	○ setB

Verify
M.Tech Thesis "A secure and Generic Captcha Application" By Dayanand, M.Tech(IS) MT/IS/1003/11
10:24 PM
4/24/2013

Figure-5



Word Grouping

Words	Set A	Set B
Relationship	● setA	○ setB
Link	● setA	○ setB
Modelling	● setA	○ setB
Designer	● setA	○ setB
Connection	● setA	○ setB
Fashion	● setA	○ setB

Figure-6

Advantage over existing Captcha system

No readability issue with word grouping captcha. Words are easily understandable, no confusion in recognising them. We can add large number of word group sets in our database. No problem with people having colour vision problem. User just needs to divide the words in two subgroups. It only requires text based interface. As it is new in comparison with existing captcha system so attacks are less vulnerable.

CAPTCHA EVALUATION PARAMETERS [17]

Consistency- When presented with the same Captcha, how reproducible is a user's answer? The level of consistency will clearly vary across different Captcha, and the acceptable level will vary by application (some may be more lenient than others).

Entropy -By entropy, we mean do different people answer the same Captcha in the same way?

Ease of generation-How difficult is it to generate a given Captcha? Can it be generated given only randomness, or does it require a precomputed/pregenerated corpus.

Implementation-Finally, how easy is it to implement? Does it require complex and elaborate graphics, or can it be implemented for a text-only system? How accessible is it?

V. EVALUATION OF DIFFERENT CAPTCHA

Text based captcha

Consistency –For Text based captcha Consistency is high.

Entropy Nearly everyone provide the same solutions for all of the Text based CAPTCHAs; there is essentially no variation. In some cases user may confused a z with an x and an 'o' with an 'a' , but all of the other answers chances are same. Whereas there was basically no inter-user variation. There are 44 possible letters in each position (upper and lowercase, with commonly-confused pairs like C/G, I/l, Q/O, h/b removed), and five possible positions, yielding $44^5=2^{27}$ possible puzzle answers.

Ease of generation– Text based CAPTCHAs are, by design, fairly easy to generate they simply require the text to be rendered and some randomness.

Implementation- Implementation is easy in comparison with other image based or word grouping Captcha.

Picture Captcha

Consistency –Consistency is good, if user is able to recognize the picture correctly.

Entropy- Entropy is less than text based captcha, since it depends on the quality of picture, difficulty level of picture and user's ability to recognize pictures. Image recognition is a hard problem

Ease of generation- It uses a larger database of photographs and animated images of everyday object.

Implementation -some implementations use only a small fixed pool of CAPTCHA images. Eventually, when enough image solutions have been collected by an attacker over a period of time, the test can be broken by simply looking up solutions in a table, based on a hash of the challenge image.

Word Grouping

Consistency-Word Grouping seems somewhat memorable without much practice. We suspect that this rate can be boosted with a tiny bit of practice.

Entropy- In principle, each word grouping has 2^6 possible outputs and we expect to see a large amount of variation.

Ease of generation-Word Grouping has some of the limitations on its corpus, it cannot become too large or users will not recognize some of the words.

Implementation- The implementation is straightforward, and has several desirable properties. The task is easy and the user interface is simple and accessible which means that it can work on screen readers or in a text-only environment like a login prompt.

APPLICATIONS

CAPTCHAs are usedin various Web applications to identifyhuman users and to restrict access to them. In **Online Polls** [18] CAPTCHAs can be used in websites that have embedded polls to protect them from being accessed by bots, and hence bring up the reliability of the polls. In protecting **Web Registration** [19] CAPTCHAs can effectively be used to filter out the bots and ensure that only human users are allowed to create accounts. In preventing **comment spam** [20]' **E-Ticketing**, **Email spam,** and **Preventing Dictionary Attacks [21][22][23].**

CONCLUSION

Creating a captcha that is so secure that no human can solve it or so user friendly that it is a trivial task for captcha breaking software is very easy to accomplish. A successful captcha by its definition is able to tell humans and computers a part. The goal is to add security features whenever possible as long as they do not significantly or unnecessarily decrease the accuracy of human solvers. Text based captcha are now breakable. If we will increase distortion, blurring and other factors, then it will be hard for human beings also to read those texts while our goal is to differentiate between humans and computers. Word grouping captcha proposes a solution for this problem. It's hard to break and user may not find any difficulty in dividing those words in two subgroups. Since user has to just click on radio buttons, so it is less time consuming also.

FUTURE WORK

Usability issues with Word Grouping Captcha. The issues may be difficulty level of words in Word grouping. Instead of dividing those six words in two subgroups by just clicking on radio button and submit, we can ask user to write those six words in two subgroups. A lot of work is needed for Consistency evaluation. How much entropy is actually present in each Captcha, as a way of determining how vulnerable they are to guessing attacks?

REFRENCES

1. MoniNaor "**Verification of a human in the loop or identification via the Turing test**". Unpublished Manuscript,1997.A preliminary draft available on http://www.wisdom.weizmann.ac.il/~naor/PAPERS/human.pdf Pages. 2-3.
2. Luis Von Ahn , Manuel Blum , and Jo n Langford "**Telling humans and computer apart automatically**". In communications of the ACM Vol. 47, No. 2, February'2004, Pages. 57-58
3. Graeme Baxter Bell "**Strengthening Captcha based web security**" ,First Monday, Volume 17, Number 2 - 6 February 2012 Pages 2-3 http://firstmonday.org/ojs/index.php/fm/article/viewArticle/3630/3145
4. .http://www.informationweek.com/news/internet/webdev/showArticle.jhtml?articleID=205900620
5. .http://www.theregister.co.uk/2008/02/25/gmailcaptchacrack/
6. . http://blogs.zdnet.com/security/?p=1232&tag=nl.e550
7. Rizwan Ur Rahman "**Survey on Captcha systems"** In *Journal of Global Research in Computer Science,* Volume 3, No. 6, June 2012, *Pages. 55-57* identiPic CAPTCHA available at http://identipic.com
8. Moin Mahmud Tanvee, Mir TafseerNayeem, Md. MahmudulHasanRafee "**Move & Select: 2-Layer CAPTCHA based on Cognitive Psychology for securing web services"**,Taken from ijens.org, available at http://www.ijens.org/Vol_11_I_05/117005-8383-IJVIPNS-IJENS.pdf**.**
9. ElieBursztein, Matthieu Martin, John C. Mitchell "**Text-based CAPTCHA Strengths and Weaknesses**" in ACM Computer and Communication security 2011 (CSS'2011) Pages 11,12
10. E. Bursztein and S. Bethard, 2009. "**Decaptcha: Breaking 75% of eBay audio CAPTCHAs**," *Proceedings of WOOT '09: Third USENIC Workshop on Offensive Technologies* (10 August Montreal, Canada), at http://www.usenix.org/event/woot09/tech/full_papers/bursztein.pdf accessed 10 February 2012. Page-3
11. Jennifer Tam, Jiri Simsa, Sean Hyde, Luis Von Ahn**"Breaking Audio CAPTCHAs"** Carnegie Mellon University page - 5,www.**captcha**.net/**Breaking**_Audio_**CAPTCHAs**.pdf
12. AndrewKirillov. aforge framework. http://www.aforgenet.com/framework/
13. C.Cortesand V. Vapnik." **Support vector networks. Machine learning**", 20(3):273–297, 1995
14. J.Yan and A.S. El Ahmad. "**A Low-cost Attack on a Microsoft CAPTCHA**". In Proceedings of the 15th ACM conference on Computer and communications security, pages 543–554. ACM,2008.
15. Wikipedia. Flood fill algorithm. http://en.wikipedia.org/wiki/Flood_fill
16. WaseemDaher "**POSH: A Generalized Captcha with security application**" In AISec ,Proceedings of the 1st ACM workshop on Workshop on AISec ,2008,Pages. 2-10
17. Luis Von Abn, Manual Blum, Nichlas Hoper and John Langford CAPTCHA: "**Using Hard AI Problems ForSecurity**".Computer Science Dept., Carnegie Mellon University, Pittsburgh PA 15213, USA
18. H. S. Baird and K. Popat."**Human interactive proofs and document image analysis**". Proc. of 5th IAPR Int. Workshop on Document Analysis Systems (DAS 2002), vol. 2423 of LNCS, pp. 507–518, 2002
19. The Official CAPTCHA site located on the internet at http://www.captcha.net/
20. Dictionary Attack, available at http://en.wikipedia.org/wiki/Dictionary_attack.
21. S. Chakrabarti and M. Singhal."**Password-based authentication: Preventing dictionary attacks**". Computer, 40(6): pp. 68–74, June 2007.
22. B. Pinkas and T. Sander. "**Securing passwords against dictionary attacks**". Proc. of 9th Conf. on Computer and Communications Security, pp. 161–170, Nov. 2002

★★★