The Design and Implementation of Online Identification of CAPTCHA Based on the Knowledge Base

Yu'e Song[1,2], Chengguo Wang[1], Ling Zhu[3], Xiaofeng Chen[1], and Qiyu Zhang[1](✉)

[1] Yantai Academy, China Agricultural University,
No. 2006, Coastal Middle Road, Gaoxin District,
Yantai 264670, Shandong Province, China
aeaeae623@163.com, rcraingo@163.com,
wangcg@126.com, cxfeng1979@126.com

[2] School of Electrical and Information Engineering,
Beijing Polytechnic College, Beijing 100042, China

[3] Shandong Institute of Business and Technology,
College of Statistics, Yantai 264005, China
oklab@qq.com

Abstract. The Completely Automated Public Turing Test to Tell Computers and Humans Apart (CAPTCHA) identification is designed to distinguish between computers and humans and it prevents the web application programs from malicious attacks, so it has been applied widely. However, great challenges must be faced with the development of CAPTCHA identification. In order to improve the safety of the professional system, the CAPTCHA online identification based on the knowledge base, which has high security and bases on semantic questions and the professionalization of professional system, is put forward combining with the recessive CAPTCHA. The specific implementation course of the new online identification method is worked out according to the example of animal identification. The application of the verification code is suitable for people who have the corresponding professional knowledge. Because the computer has great difficulty to answer semantic information questions, which are also professional issues, so the new online identification method based on the verification of knowledge has very high security.

Keywords: CAPTCHA · Online identification · Knowledge base · Animal

1 Introduction

With the rapid development of internetwork, security problem of the web application becomes an extremely important issue for us. The HTTP attack based on the form automatically submission is a common way of network attack. According to the HTTP protocol, the attacker can write program to simulate the method of form submission, and submit the abnormal data to site service automatically and rapidly. This constitutes the basic HTTP attacks. An attacker can repeat logging to break a user's password and

D. Li and Z. Li (Eds.): CCTA 2015, Part II, IFIP AICT 479, pp. 92–99, 2016.
DOI: 10.1007/978-3-319-48354-2_9

this will lead to a leakage of users' privacy information. In order to prevent the attacker using program automatic login, Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) technology has been widely used [1].

The CAPTCHA is a kind of program algorithm to distinguish between computers and humans, so the procedure must be able to generate and evaluate computer test which human can easily pass but not for computers [2, 3]. Because the computer cannot solve CAPTCHA question, the user who answer the question can be considered human [4].

In order to protect the network, CAPTCHA has been applied widely, such as preventing spam ads in the blog post, protecting website registration and the E-mail address, online polls, preventing dictionary attacks, the search engine robots, worms and spam, etc.

Since CAPTCHA has been proposed, different research institutions and scholars have developed a variety of CAPTCHA. CAPTCHA has different ways of classification [5]. According to the type of information, CAPTCHA can be divided into text CAPTCHA, image CAPTCHA, graphics CAPTCHA, audio CAPTCHA and video CAPTCHA. According to the way of recognition, CAPTCHA can be divided into dominant CAPTCHA and implicit CAPTCHA. According to the interaction, CAPTCHA can be divided into static CAPTCHA and dynamic CAPTCHA. Along with the development of the CAPTCHA, CAPTCHA recognition technology is also developing and some methods have been put forward, such as the matching shape context [6], template matching [8] and neural network identification methods [7]. This makes the security of the CAPTCHA has a huge challenge. Dynamic CAPTCHA and recessive CAPTCHA have a good security and is the research direction in the future.

The hidden CAPTCHA [5] refers to answering the question of the CAPTCHA expressing according to the semantic of CAPTCHA, for example, CAPTCHA system first randomly generates an expression (5 + 3)*9/4 and requires the user to answer the expression values; CAPTCHA system picks up a few images from the graphics library and users need to rotate the graphics to the right direction. Though artificial intelligence has a rapid development, the computer has much difficulty to answer semantic information questions, so the hidden CAPTCHA is safe.

In this paper, the CAPTCHA technology is studied deeply. Based on the implicit CAPTCHA and combining with the characteristics of professional system, a new kind of CAPTCHA is proposed based on the knowledge base and the security of the system can be effectively improved using the new kind of CAPTCHA.

2 Knowledge Representation

In the knowledge base, knowledge representation methods are logical notation, production representation, frame representation and object-oriented representation, semantic representation and the XML representation and representation of ontology [9], etc. According to the characteristics of the CAPTCHA, we choose production knowledge representation description.

Shortliffe firstly introduced the concept of production in the famous expert system MYCIN. The structure IF (E1 & E2 & … & En) THEN A is called the rule. It means that if the logical expression of E1 & E2 & … & En established, the conclusion A is right. The expression E1 & E2 & … & En is called former part of the rule and is any legal logical expressions. It is the prerequisite for reasoning by using the rule. A is called later part of the rule and is the result of reasoning using the rule. [10]. The rule knowledge representation has many advantages, such as simple and clear reasoning, the reasoning machine design and implementation is simple and has a good characteristics in some specific application environment, etc.

3 The Design of CAPTCHA Based on the Knowledge Base

For some professional systems, CAPTCHA can be structured based on knowledge base. Because users have the corresponding knowledge and can reason the related results according to the precondition. Let us use a simple animal identification as an example to illustrate how to construct CAPTCHA.

We give the following rules about animal identification:

IF the animal has hair THEN the animals are mammals.

IF the animal has milk THEN the animals are mammals.

IF the animal has feathers THEN the animal is a bird.

IF the animal can fly AND lay eggs THEN the animal is a bird.

IF the animal eats meat THEN the animal is a carnivorous animal.

IF the animal has a canine tooth AND claw AND eyes staring at front THEN the animal is a carnivorous animal.

IF the animal is mammals and has claw THEN the animal is a hoof animal.

3.1 The Design of the Database and Table for Knowledge Base

According to the rules of reasoning above, we designed the rules table, inferences table and synonym table. Rules table save the atomic conditions of precondition, which are the minimum condition of premise condition. The above animal identification rules are in the rules table as shown in Table 1.

Table 1. Animal identification rules

Serial number	Rules
1	Have hair
2	Have milk
3	Have feathers
4	Can fly
5	Can lay eggs
6	Eat the meat
7	Have canine tooth
8	Have claws
9	Eye star at the front
10	Have hoof

The result of reasoning is text messages. There are different representations for the same text messages and the computer can't recognize it very well, therefore automatic word segmentation can be used for the results and CAPTCHA. In this process, the word which not be used can be removed and the keywords will be extracted, then we can match the keyword. For Chinese word segmentation, IK Analyzer 2012 can be used. The IK Analyzer is an open source lightweight Chinese word segmentation toolkit based on Java language. In the 2012 version, we support configuring IKAnalyzer. CFG.XML file to expand proprietary dictionary and stop using dictionary and dictionary format is utf-8 without BOM in Chinese text files [11]. Stop using words are not really meaning of function words in both English and Chinese [10] and can be ignored because they does not affect the understanding of sentence meaning. The stop using dictionaries are built on the basis of the literature [10, 11]. In order to assist CAPTCHA judgment, two options are increased which must be contained keywords and must not contained keywords. Meanwhile, in order to reduce the complexity of the system reasoning, the result is made as easy as possible. Inferences table is shown in Table 2.

Table 2. Inference table data

Premise condition	Results	Whether the word is segmented	Must contained keywords	Must not contained keywords
1	Mammals	no	no	no
2	Mammals	no	no	no
3	Birds	no	no	no
4, 5	Birds	no	no	no
6	Predators	no	no	no
7, 8, 9	Predators	no	no	no
1, 10	Hoofed animals	no	no	no
2, 10	Hoofed animals	no	no	no

Synonym of the word in the results is stored synonym table, including Chinese, English and acronyms.

In the MySQL database we design different table structures, which are shown in Tables 3, 4 and 5.

Table 3. Rule table

Field	Data type	Note
Id	int	Automatic numbering, primary key
Rule	varchar(100)	

Table 4. Inferences table

Field	Data type	Note
Id	int	Automatic numbering, primary key
Condition	varchar(100)	
Result	varchar(100)	
Segmentation	char(1)	
Key	varchar(200)	
Antonym	varchar(200)	

Table 5. Synonym table

Field	Data type	Note
Id	int	Automatic numbering, primary key
Key	varchar(100)	
Synonym	varchar(100)	

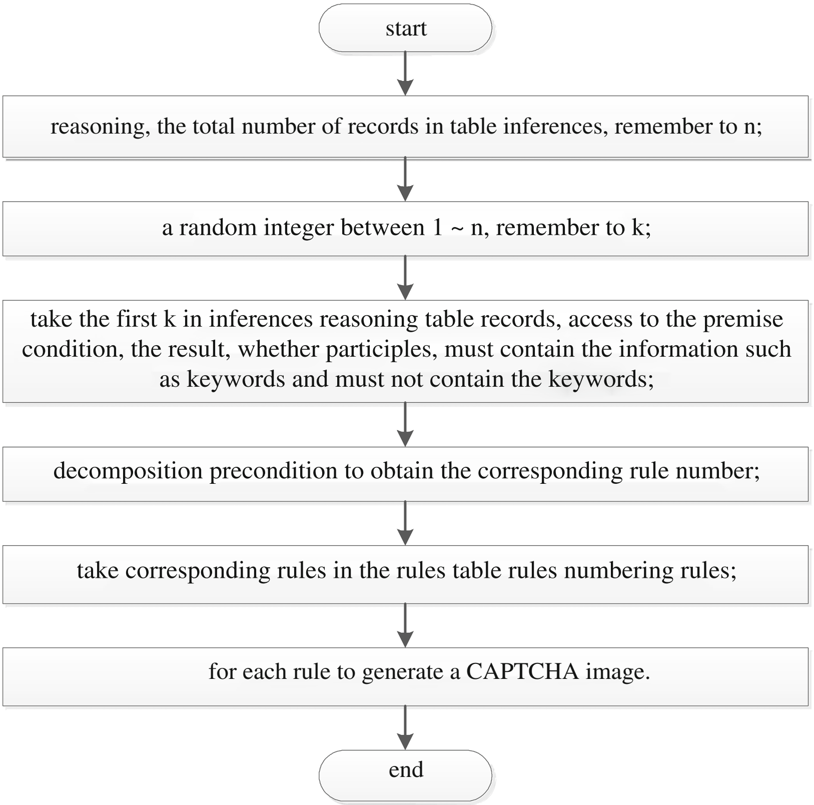
start
reasoning, the total number of records in table inferences, remember to n;
a random integer between 1 ~ n, remember to k;
take the first k in inferences reasoning table records, access to the premise condition, the result, whether participles, must contain the information such as keywords and must not contain the keywords;
decomposition precondition to obtain the corresponding rule number;
take corresponding rules in the rules table rules numbering rules;
for each rule to generate a CAPTCHA image.
end

Fig. 1. CAPTCHA generation algorithm

3.2 CAPTCHA Generation Algorithm

(1) Reason the total number of records in inferences table and remember to n;
(2) Randomly select the integer between 1–n, remember to k;
(3) Take the kth records in the inferences table and access the premise condition, the result, whether participles, the keywords which must be contained and which must not be contained;
(4) Decompose precondition to obtain the corresponding rule number;
(5) Take corresponding rules in the rules table rules numbering rules;
(6) Generate a CAPTCHA image for each rule.

The algorithm flow chart is shown in Fig. 1.

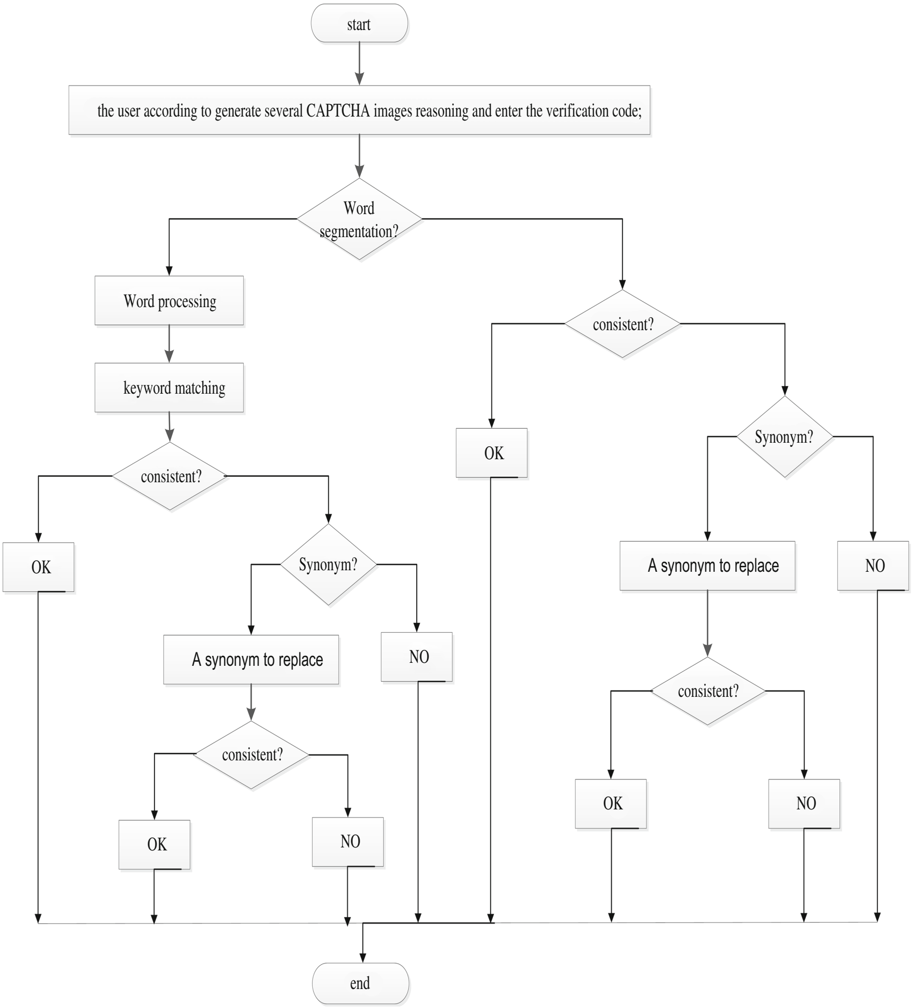

Fig. 2. CAPTCHA validation algorithm

3.3 CAPTCHA Validation Algorithm

(1) The user reasons according to generated CAPTCHA images and enter the CAPTCHA;

(2) Word segmentation? If no, compare the CAPTCHA entered by the user and the results and judge whether they are consistent. If consistent, agree on. If inconsistent, judge whether there is a synonym and whether consistent after replacement. If unanimity, agree on. If inconsistent, not through;

(3) If the words need segmentation, do words segmentation to the CAPTCHA entered by the user and results and match the keyword. If they are consistent, agree on. If inconsistent, judge whether there is a synonym and whether consistent after replacement. If unanimity, agree on. If inconsistent, not through. The algorithm flow chart is shown in Fig. 2.

3.4 CAPTCHA Implementation

The realization of the CAPTCHA is shown in Fig. 3.

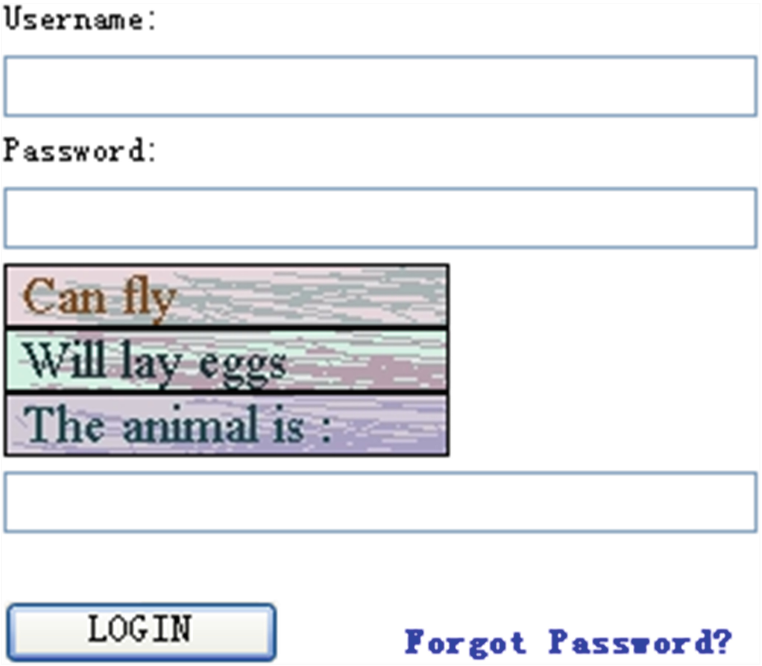

Fig. 3. Authentication code implementation

4 Conclusion

The CAPTCHA has a variety of forms, but the development of CAPTCHA recognition technology causes a hidden danger for the security of the CAPTCHA. In order to improve the security of the CAPTCHA, a new kind of CAPTCHA based on knowledge base is put forward combining the implicit CAPTCHA, which is based on semantic information question and answer and the professional system. This new CAPTCHA can significantly improve the security of the professional system. The CAPTCHA designed in this paper is suitable for professional system but not for general system, such as E-mail.

Acknowledgments. This work was supported by the scientific research fund project of China Agricultural University Yantai academy (YT201311, 201201Ja), the science and technology plan project of Beijing education committee (Grant no. KM201510853006) and key scientific research project of Beijing Polytechnic College (Grant no. bgzykyz201502, bgzykyz201503).

References

1. Ji, Z.: Principles and prevention of HTTP attacks based on identifying code recognization. Comput. Eng. **32**(20), 170–172 (2006)
2. Ying, X.: The research on user modelling for internet personalized services. Ph.D. thesis, National University of Defense Technology (2003)
3. von Ahn, L., Blum, M., Langford, J.: Telling humans and computers apart automatically. Commun. ACM **47**(2), 57–60 (2004)
4. Tao, R., Song, Y.E., Wang, Z.J.: Ambiguity function based on the linear canonical transform. IET Signal Process. **6**(6), 568–576 (2012)
5. Wang, B., Wang, J., Du, K., et al.: Research on attach and strategy of CAPTCHA technology. Appl. Res. Comput. **30**(9), 2776–2779 (2013)
6. Mori, G., Malik, J.: Recognizing objects in adversarial clutter: breaking a visual CAPTCHA. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 124–141 (2003)
7. Zuo, B., Shi, X., Xie, F., et al.: A neural network based approach to recognizing the verification code. Comput. Eng. Sci. **31**(12), 20–22 (2009)
8. Huang, S., Xu, M.: Recognition and improvement of identifying code. J. Nanjing Normal Univ. (Eng. Technol. Ed.) **9**(2), 84–88 (2009)
9. Liu, J.-W., Yan, L.-F.: Comparative study of knowledge representation. Comput. Syst. Appl. **20**(3), 242–246 (2010)
10. Zhang, X., Gao, H., Zhao, Z.: The rule representation for knowledge in database style. Comput. Eng. Appl. **38**(1), 200–202 (2002)
11. Zhang, Q.: Research and design of spam email filter system based on bayesian algorithm spam. M.S. Thesis, Qufu Normal University (2006)