Response Time Analysis of Text-Based CAPTCHA by Association Rules

Darko Brodić[1(✉)], Alessia Amelio[2], and Ivo R. Draganov[3]

[1] University of Belgrade, Technical Faculty in Bor, V.J. 12, 19210 Bor, Serbia
dbrodic@tf.bor.ac.rs
[2] DIMES, University of Calabria, Via Pietro Bucci Cube 44, 87036 Rende (CS), Italy
aamelio@dimes.unical.it
[3] Technical University of Sofia, Blvd. Kliment Ohrdski 8, 1000 Sofia, Bulgaria
idraganov@tu-sofia.bg

Abstract. The paper introduces and discusses the usability problem of text-based type of CAPTCHA. In particular, two types of text-based CAPTCHA, with text and with numbers, are in the focus. The usability is considered in terms of response time to find a solution for the two aforementioned types of CAPTCHA. To analyze the response time, an experiment is conducted on 230 Internet users, characterized by multiple features, like age, number of years of Internet use, education level, response time in solving text-based CAPTCHA and response time in solving text-number-based CAPTCHA. Then, association rules are extracted from the values of these features, by employing the Apriori algorithm. It determines a new and promising statistical analysis in this context, revealing the dependence of response time to CAPTCHA to the co-occurrence of the feature values and the strength of these dependencies by rule support, confidence and lift analysis.

Keywords: Text-based CAPTCHA · Text-number-based CAPTCHA · Association rules · Statistical analysis · Apriori algorithm

1 Introduction

CAPTCHA is an acronym for "Completely Automated Public Turing Text to tell Computers and Humans Apart". It is based on "Turing tests" which have to recognize the difference between humans and machine by the humans [13]. In fact, CAPTCHA is a program that uses reversed "Turing tests". Accordingly, it has to recognize the difference between humans and machine (computer). This presents the most intriguing part, because the computer has to have an ability to discriminate humans from the machine [11]. Essentially, it is a test evaluated by computers, which "only" humans can pass [14].

In some way, the CAPTCHA has analogies with cryptographic problem, in which humans have a key to decrypt the problem unlike the bots. If we take into account "Turing test", CAPTCHA has to fulfill the following conditions [11]: (i) it should be easy to generate many instances of the problem linked with

© Springer International Publishing Switzerland 2016
C. Dichev and G. Agre (Eds.): AIMSA 2016, LNAI 9883, pp. 78–88, 2016.
DOI: 10.1007/978-3-319-44748-3_8

their solutions, (ii) humans have to solve it easily, (iii) the most advanced bots, i.e. programs will fail to solve the problem, and (iv) the problem is succinctly explained to be easily understood by the humans. Also, it should not be forgotten that one of the most important CAPTCHA criteria is to be publicly available (open code) [11].

The CAPTCHA can be used for the following reasons [6]: (i) prevention and reduction of spams on forums, (ii) prevention of opening a large number of orders by users, (iii) protection of user accounts from bots attacks, and (iv) validation of online surveys by determining whether the humans or bot access them.

The CAPTCHA quality can be considered in the following directions [3]: (i) usability, (ii) security, and (iii) practicality. The usability considers difficulties of solving the CAPTCHA. It means that it quantifies the time to find the solution to the CAPTCHA. Security evaluates the difficulty of finding solutions to CAPTCHA by the computer (computer program, i.e. bot). The practicality refers to the way of creating and realizing the CAPTCHA by programmers tools.

In this study, the text-based type of CAPTCHA will be considered. Still, the only element that will be taken into account is its usability. Hence, we shall explore the complexity of solving CAPTCHA from the user's viewpoint. A similar statistical analysis was made with a population of 24 participants, which represents a very small population from the factor analysis point of view [8]. Our experiment will include a population of 230 different Internet users, which is a more representative population sample. Two types of text-based CAPTCHA will be under consideration: (i) with text only, and (ii) with numbers only. The tested Internet users vary by the following attributes: (i) age, (ii) years of Internet use, and (iii) educational level. The experiment will show the differences of using different text-based CAPTCHAs which include text or number according to its usability value. This approach is important for choosing adequate types of CAPTCHA and their implementation in different web environments. Also, it identifies the suitability of the CAPTCHA type to a certain group of Internet users. In this paper, the usability of the CAPTCHA will be evaluated by data mining tools like association rules, which is to the best of our knowledge for the first time used for analyzing the CAPTCHA in the known literature. The association rules method is primarily focused on finding frequent co-occurring associations among a collection of items. Hence, its goal is to find associations of items that occur together more often than you would expect from a random sampling of all possibilities. As a consequence to CAPTCHA time response analysis by association rules, the results will show the real dependence between different variables which are of great or small importance to the value of CAPTCHA users response time.

The paper is organized in the following manner. Section 2 describes the text-based CAPTCHA. Section 3 describes the experiment. It includes the data mining elements like association rules, which will be used as the main evaluation tool. Section 4 presents the results and gives the discussion. Section 5 draws the conclusions.

2 Text-Based CAPTCHA

First CAPTCHA was designed by Broder's team in 1997 for Altavista, to prevent automatic adding URL to a database of a web browser [9].

Text-based CAPTCHA is the most widespread type of CAPTCHA. It asks the user to decrypt the text which is usually distorted in some way [10]. The text-based scheme typically relies on sophisticated distortion of the text images rendering it unrecognizable to the state of the art. It is popular among programmers, because it can be easily created. Hence, it is characterized by high-raking practicality. Its security depends on the solution quality of different element combinations in text and its background. However, it is attacked by the most advanced types of bots, which include the OCR system in. Figure 1 shows samples of typical text-based type of CAPTCHA.

Fig. 1. The samples of the text-based CAPTCHA

Figure 2 illustrates the typical samples of text-based type of CAPTCHA, which include numbers, sometimes called number-based CAPTCHA.

Fig. 2. The samples of the number-based CAPTCHA

Although the aforementioned text and number-based CAPTCHAs seem almost identical, the number-based CAPTCHA can include also some arithmetic operation. Hence, the user response to it can be different. However, from the user standpoint, it is extremely valuable to research the eventual response time differences in order to evaluate their suitability for certain group of Internet users. This is the main point that is going to be explored in the experimental part of our study.

3 Experiment

The experiment includes the testing of 230 Internet users, which have to solve given CAPTCHAs. Accordingly, each user is asked to solve a text and a number-based CAPTCHA. The value of 5 features including the following: (i) age, (ii) number of years of Internet use, (iii) education level, (iv) response time to solve text-based CAPTCHA, and (v) response time to solve number-based CAPTCHA is registered for each user, and then stored into a dataset. Hence, the dataset contains a total of 230 instances of 5 feature values for each Internet user.

The aim of the analysis is to find the association among the different feature values. In particular, we investigate if there is some implication between multiple feature values concerning the time spent by the user to efficiently solve text and number-based CAPTCHAs, and the strength of this implication. In fact, it is useful to understand in which users' conditions a given response time is obtained as well as the correlation between multiple feature values and the response time of the user. To pursue this goal, we find sets of feature values co-occurring frequently together, which are associated to a given response time of the user.

3.1 Association Rules

The most natural method to detect the frequent sets of feature values is the association rules one [1]. Next, we briefly recall the main concepts underlying the association rules and introduce the constraints for solving our task.

Let $T = T_1 \cup T_2 \cup, ..., \cup T_n$ be the set of items corresponding to all possible values of n features. In particular, let $T_i = \{t_i^1, t_i^2, ..., t_i^x\}$ be the set of x items representing all possible values of feature i. In our case, n is equal to 5 and T contains all the possible values of age, number of years of Internet use, education level, text-based CAPTCHA response time, and number-based CAPTCHA response time. A transaction I is a subset of n items extracted from T, such that $\forall t_i^j, t_k^h \in I$, $i \neq k$, and j and h define two possible values respectively of feature i and k. It indicates that items in I correspond each to a value of a distinct feature. In our case, given i.e. the possible values of "education level" feature, only one of them can be contained inside transaction I. Consequently, each transaction represents an instance of the aforementioned dataset.

Given these concepts, an Association Rule (AR) is defined as an implication $W \Rightarrow Z$, where W and Z are disjoint sets of items, representing respectively antecedent and consequent, characterized by *support* and *confidence*. Support measures how many times the items in $W \cup Z$ co-occur in the same transaction inside the dataset. It evaluates the statistical significance of AR. Confidence quantifies how many times the items in $W \cup Z$ co-occur in the transactions containing W. It is an estimate of the conditional probability of Z given W.

The aim is to detect the "meaningful" ARs from the dataset, having support $\geq minsupport$ and confidence $\geq minconfidence$, whose Z contains only items of text-based or number-based response time. It allows to analyze the co-occurrences of age values, number of Internet years use and education level,

determining a given response time, and the strength of these co-occurrences, realizing a statistical analysis.

In order to detect the "meaningful" ARs, the well-known *Apriori* algorithm is employed [2]. It consists of two main steps: (i) detection of all the sets of items with frequency $\geq$ *minsupport* in the dataset; (ii) construction of the ARs from detected sets of items, considering the *minconfidence* threshold. The algorithm is based on the *anti-monotonicity* concept, for which if a set of items is unfrequent, also every its superset will be. For this reason, in each iteration, unfrequent sets of items are pruned from the algorithm. In the first step, the algorithm detects the sets of items of size k with frequency $\geq$ *minsupport*. Then, it enlarges them to size $k+1$, by including the only sets of items of size 1 and frequency $\geq$ *minsupport*. These two steps are repeated starting from $k=1$ to a certain value of k, for which no more sets of items of size $k+1$ with frequency $\geq$ *minsupport* can be generated. In the second step, rules are generated for each frequent set of items F, finding all the subsets $f \subset F$ such that confidence value of $f \rightarrow F-f$ is $\geq minconfidence$.

To evaluate the efficacy of predictability of the ARs, the *lift* measure is adopted [5]. *Lift* is the ratio between the confidence of the rule and the expected confidence, assuming that W and Z are independent [7]. If the lift takes a value of 1, then W and Z are independent. On the contrary, if the lift is bigger than 1, W and Z will co-occur more often than expected. This means that the occurrence of W has a positive effect on the occurrence of Z. Hence, the rule is potentially able to predict Z in different datasets.

4 Results and Discussion

All experimentation has been performed in MATLAB R2012a on a notebook quad-core at 2.2 GHz, 16 GB RAM and UNIX operating system.

In the aforementioned dataset, "age" can assume two different values, identifying users below 35 years, and users above 35 years. "Education level" has also two possible values, which are expressed as "higher education" and "secondary education". Furthermore, the values of the "number of years of Internet use" vary in the interval from 1 to 9 in the dataset. However, in order to improve the quality of the extracted ARs, this last feature has been regularly discretized by adopting an Equal-Width Discretization [12], to obtain only 3 intervals of interest, identified as "high Internet use" (>6 years), "middle Internet use" (from 4 to 6 years), and "low Internet use" (<4 years).

The two features, representing respectively the response time in solving text-based CAPTCHA and the response time in solving number-based CAPTCHA, have numerical values too, but corresponding to real numbers ranging from 0.0 to $+\infty$. A test using different discretization methods has been conducted on the numerical values and showed that clustering is the most suitable method for this task. Accordingly, response times have been discretized into 3 different intervals, by adopting the K-Medians clustering algorithm [4], whose advantage is the detection of natural groups of values based on their characteristics, overcoming

the limitations of K-Means in managing the outliers. K-Medians determines the k value intervals, where $k = 3$ in this case, minimizing a function J, which quantifies the total sum of L_1 norm between each value in a given interval and its centroid. Given an interval, its centroid is the median over the values of the interval. Ten different executions of K-Medians algorithm have been performed on the response time values, and the set of intervals with the lowest value of J has been selected as the final solution. At the end, K-Medians determined 3 different intervals corresponding to "low response time" ($<$ 20.09 s), "middle response time" (from 20.09 s to 39.97 s), and "high response time" ($>$ 39.97 s).

Table 1 shows the possible values for each feature of the dataset. For discretized features, their corresponding intervals are reported.

Table 1. Possible values and corresponding intervals for each feature of the dataset

Features	Values	Interval
Age	Below 35	-
	Above 35	-
Education level	Higher education	-
	Secondary education	-
Number of years of Internet use	High Internet use	$>$ 6 years
	Middle Internet use	4 - 6 years
	Low Internet use	$<$ 4 years
Response time in solving text-based CAPTCHA	High response time	$>$ 39.97 s
	Middle response time	20.09 s–39.97 s
	Low response time	$<$ 20.09 s
Response time in solving number-based CAPTCHA	High response time	$>$ 39.97 s
	Middle response time	20.09 s–39.97 s
	Low response time	$<$ 20.09 s

An example of transaction obtained from the values, eventually discretized, of the features, corresponding to a row of the dataset, is given in Table 2.

Table 2. Example of transaction

Age	Educ. level	Num. years Internet use	Resp. time text	Resp. time text-number
Above 35	Higher educ	Middle Internet use	Middle resp. time	Middle resp. time

Furthermore, an example of AR defined over the values of the features "age", "number of years of Internet use", and "response time in solving text-based CAPTCHA", is given in Eq. (1).

(age) (Num. years Internet use) (Resp. time text)

$$above\ 35,\ middle\ Internet\ use \rightarrow middle\ response\ time \quad (1)$$

The antecedent W of the AR is $\{above\ 35,\ middle\ Internet\ use\}$, while the consequent Z is $\{middle\ response\ time\}$. It can be observed that $W \cup Z$, which is $\{above\ 35,\ middle\ Internet\ use, middle\ response\ time\}$, is contained inside the transaction in Table 2. This AR expresses that a response time in solving text-based CAPTCHA between 20.09 s and 39.97 s is likely to occur in correspondence to users with age above 35 years and having an experience in Internet use varying from 4 to 6 years.

Tables 3-4 contain the ARs extracted from the dataset by the Apriori algorithm. For each AR, the antecedent W, the consequent Z, the support, confidence and *lift* are reported. Also, the thresholds of *minsupport* and *minconfidence* have been fixed respectively to 5 % and 50 %, because even if the sets of items are not so frequent in the dataset, the corresponding ARs are still interesting to analyze. It is worth to note that the *lift* is > 1 for all the ARs. This means that the antecedent and the consequent are always positively correlated.

Table 3 reports the ARs extracted from the dataset, whose consequent is a value associated to response time in solving text-based CAPTCHA. Looking at the ARs, some general considerations of interesting dependences can be performed. In particular, users with age below 35 years are almost always associated to a low response time in solving the text-based CAPTCHA. Also, users with age above 35 years are mostly associated to a middle-high response time in solving text-based CAPTCHA. The rule with the highest confidence of 0.90 is $\{below\ 35, higher\ educ., middle\ Internet\ use \rightarrow low\ resp.\ time\}$. It indicates that in 90 % of transactions where users are: (i) below 35 years, (ii) have a higher education, and (iii) a middle experience in Internet use, the response time is low. The AR with the second highest confidence value of 0.8723 is $\{below\ 35, higher\ educ. \rightarrow low\ resp.\ time\}$. It denotes that in around 87 % of transactions where users are below 35 years having higher education, the response time to text-based CAPTCHA is low. Furthermore, it can be noted that the confidence of this rule is higher than the confidence of the rule $\{below\ 35, secondary : educ. \rightarrow lowresp.time\}$. It indicates that in most of the cases when users are below 35 years and have a higher education, the response time is low. Also, it is more likely that young users with higher education level, rather than with secondary education level, provide a solution to text-based CAPTCHA in a reduced time. It is also interesting to observe that, when users with middle experience in Internet use are below 35 years, the response time is low with a confidence of 0.6364, while for users above 35 years, the response time is middle with the same confidence. It indicates that the age is discriminatory for obtaining a low response time, while the Internet use is not. Finally, a meaningful rule with lift value of 5.5740 is $\{above 35, low\ Internet\ use \rightarrow high\ resp.time\}$ indicating that users above 35 years with a low Internet use are strongly correlated to a high response time.

Table 4 shows the ARs extracted from the dataset, whose consequent is a value related to response time to solve a number-based CAPTCHA. It can be

Table 3. Association rules extracted from the dataset, for which the consequent is related to response time in solving text-based CAPTCHA

Antecedent	Consequent	Support	Confidence	Lift
Below 35, Higher educ., Middle internet use	Low resp. time	0.0786	0.9000	1.7466
Below 35, Higher educ.	Low resp. time	0.1790	0.8723	1.6929
Above 35, Low internet use	High resp. time	0.0524	0.7059	5.5740
Below 35	Low resp. time	0.4672	0.6687	1.2978
Below 35, High internet use	Low resp. time	0.1223	0.6667	1.2938
Below 35, Middle internet use	Low resp. time	0.1834	0.6364	1.2350
Above 35, Middle internet use	Middle resp. time	0.0917	0.6364	1.4870
Above 35, Higher educ., Middle internet use	Middle resp. time	0.0611	0.6364	1.4870
Below 35, Secondary educ., High internet use	Low resp. time	0.0742	0.6071	1.1783
Above 35, Higher educ.	Middle resp. time	0.1048	0.6000	1.4020
Below 35, Secondary educ.	Low resp. time	0.2882	0.5841	1.1335
Higher educ., High internet use	Low resp. time	0.0611	0.5833	1.1321
Higher educ.	Low resp. time	0.2183	0.5747	1.1153
High internet use	Low resp. time	0.1354	0.5741	1.1141
Secondary educ., High internet use	Low resp. time	0.0742	0.5667	1.0997
Above 35	Middle resp. time	0.1703	0.5652	1.3208
Below 35, Secondary educ., Middle internet use	Low resp. time	0.1048	0.5217	1.0125
Above 35, Secondary educ.	Middle resp. time	0.0655	0.5172	1.2086

noted that users with an age below 35 years and higher education level are mostly associated to a low response time in solving number-based CAPTCHA, independently from the experience in Internet use. Differently, users with an age above 35 years are strongly associated to a middle response time in solving number-based CAPTCHA, with lift values of around 2. In particular, the AR with the highest confidence of 1.00 is $\{below\,35, higher\,educ. \rightarrow low\,resp.\,time\}$. It indicates that, among the users below 35 years and with higher education level, the 100 % of them is able to quickly solve the number-based CAPTCHA. This means that an age below 35 years together with a higher education level are well-correlated to a low response time. However, also the users only having an age below 35 years or a higher education level are able to characterize the low response time, with smaller confidence values but higher support values

Table 4. Association rules extracted from the dataset, for which the consequent is related to response time in solving number-based CAPTCHA

Antecedent	Consequent	Support	Confidence	Lift
Below 35, Higher educ.	Low resp. time	0.2052	1.0000	1.4967
Below 35, Higher educ., Middle internet use	Low resp. time	0.0873	1.0000	1.4967
Below 35, Higher educ., High internet use	Low resp. time	0.0611	1.0000	1.4967
Below 35	Low resp. time	0.5851	0.8375	1.2535
Below 35, Middle internet use	Low resp. time	0.2402	0.8334	1.2473
Below 35, High internet use	Low resp. time	0.1528	0.8334	1.2473
Below 35, Secondary educ.	Low resp. time	0.3799	0.7699	1.1523
Below 35, Secondary educ., Middle internet use	Low resp. time	0.1528	0.7609	1.1388
Higher educ., High internet use	Low resp. time	0.0786	0.7500	1.1225
Below 35, Secondary educ., High internet use	Low resp. time	0.0917	0.7500	1.1225
High internet use	Low resp. time	0.1747	0.7407	1.1087
Secondary educ., High internet use	Low resp. time	0.0961	0.7333	1.0976
Higher educ.	Low resp. time	0.2664	0.7011	1.0494
Above 35, Higher educ., Middle internet use	Middle resp. time	0.0568	0.5909	2.1479
Above 35, Middle internet use	Middle resp. time	0.0830	0.5758	2.0928
Above 35, Higher educ.	Middle resp. time	0.1004	0.5750	2.0901
Above 35	Middle resp. time	0.1528	0.5072	1.8438

than in the case where both the features are considered. Again, users with: (i) an age below 35 years, (ii) a higher education level, (iii) a middle experience in Internet use, and (iv) a low response time in solving the number-based CAPTCHA, appear in around 9 % of transactions. It is a middle-low value, but with a confidence of 1.00. For this reason, the information retrieved from the AR $\{below\ 35,\ higher\ educ.,\ middle\ Internet\ use \rightarrow low\ resp.\ time\}$ becomes more interesting. It indicates that in the 100 % of the cases where: (i) users are below 35 years, (ii) have a higher education level, and (iii) have a middle experience in Internet use, the number-based CAPTCHA is quickly solved. In contrast, among the users above 35 years, with higher education level and middle Internet use, a middle response time occurs in around 59 % of cases. For this reason, the connection between age above 35, higher education level and middle response time

becomes promising. Finally, it is worth to note that a high Internet use is able to obtain a low response time, independently from education level.

In conclusion, from this analysis we observed that an age below 35 or a higher education level are strongly correlated to a low response time in solving the text-based CAPTCHA. Also, an age below 35 together with a higher education or a middle experience in Internet use are mostly correlated to a low response time in solving number-based CAPTCHA. Furthermore, comparison between text and number-based CAPTCHA shows that: (i) number-based CAPTCHA has higher support and confidence to be solved in low response time (below 35, higher educ.), and (ii) text-based CAPTCHA has higher support and confidence to be solved in the middle and high response time (above 35, low Internet use). Hence, number-based CAPTCHA should be easily solvable.

5 Conclusions

The study introduced a new statistical method for analysis of the response time of Internet users to solve the CAPTCHA. In particular, text and number-based CAPTCHAs were in the focus. Analysis consisted of investigating the dependence between the response time to solve CAPTCHA and typical features associated to the users, like age, education level and number of years of Internet use. It was statistically accomplished by adopting the ARs. Apriori algorithm was employed on the dataset of the users for obtaining the ARs together with their support, confidence and lift. Analysis of the obtained ARs revealed interesting co-occurrences of feature values and their association to a given response time to solve the text and number-based CAPTCHA. Also, it allowed to investigate on the strength of these dependencies by making considerations about the rule support, confidence and lift.

Future work will extend the method to different features, to other types of CAPTCHA and to alternative statistical measures. Also, other closed itemset mining algorithms in Java or C/C++ will be adopted and tested on the dataset.

Acknowledgments. The authors are fully grateful to Ms. Sanja Petrovska for the helpful support in collecting the data.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data - SIGMOD, pp. 207–216 (1993)
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pp. 487–499 (1994)
3. Baecher, P., Fischlin, M., Gordon, L., Langenberg, R., Lutzow, M., Schroder, D.: CAPTCHAs: the good, the bad and the ugly. In: Frieling, F.C., (ed.) Sicherheit. LNI, Vol. 170, pp. 353–365 (2010)

4. Bradley, P.S., Mangasarian, O.L., Street, W.N.: Clustering via concave minimization. Adv. Neural Inf. Process. Syst. **9**, 368–374 (1997). MIT Press
5. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD), pp. 265–276 (1997)
6. CAPTCHA. www.google.com, http://en.wikipedia.org/wiki/CAPTCHA
7. Hahsler, M.: A probabilistic comparison of commonly used interest measures for association rules (2015). http://michael.hahsler.net/research/association_rules/measures.html
8. Lee, Y.L., Hsu, C.H.: Usability study of text-based CAPTCHAs. Displays **32**(1), 81–86 (2011)
9. Lillibridge, M., Abadi, M., Bharat, K., Broder, A.: Method for selectively restricting access to computer systems. United States Patent 6195698, Applied 1998 and Approved 2001
10. Ling-Zi, X., Yi-Chun, Z.: A case study of text-based CAPTCHA attacks. In: Proceedings of International Conference on Cyber Enabled Distributed Computing and Knowledge Discover, pp. 121–124 (2012)
11. Naor, M.: Verification of a human in the loop or Identification via the Turing Test. Report, Weizmann Institute of Science (1996)
12. Sullivan, D.G.: Data mining V: preparing the data. http://cs-people.bu.edu/dgs/courses/cs105/lectures/data_mining_preparation.pdf
13. Turing, A.M.: Computing machinery and intelligence. Mind **59**, 433–460 (1950)
14. Von Ahn, L., Blum, M., Langford, J.: Telling humans and computers apart automatically. Commun. ACM **47**(2), 47–60 (2004)