Капшук О.О., Бєлушкін М.Є.
Інститут прикладного системного аналізу НТУУ "КПІ", Київ, Україна

Надійність захисту веб-сайтів від автоматичного розпізнавання символів CAPTCHA

Захист інформації у веб-просторі з кожним роком набуває все більшої актуальності. Зокрема, захист веб-сайтів від роботів, спамерів та інших автоматичних засобів отримання доступу до них відіграє важливу роль в цьому процесі. Однією з найпоширеніших технологій захисту веб-сайтів в середовищі Internet є використання символьних CAPTCHA (Completely Automated Public Turing Tests to Tell Computers and Humans Apart) - повністю автоматизованих публічних тестів Тюринга для розрізнення комп'ютерів і людей в Internet, які використовується для того, щоб визначити, хто використовує систему — людина чи комп'ютер. CAPTCHA генерує тести, які можуть бути виконані людиною, але вони мають бути недоступними для існуючих на сьогоднішній день комп'ютерних програм. CAPTCHA часто використовується для того, щоб запобігати чисельним автоматичним реєстраціям та відправленням повідомлень програмами-роботами [1].

На сьогоднішній день існує багато реалізацій тесту Тюринга для мережі Internet від примітивних до найскладніших. В доповіді проводиться аналіз надійності захисту веб-сайтів з використанням символьних CAPTCHA та проблем побудови стійких до автоматичного розпізнавання CAPTCHA. Для аналізу було розроблено декодер CAPTCHA на базі бібліотеки комп'ютерного зору з відкритим вихідним кодом OpenCV, а також використані відомі OCR системи: GNU Ocrad, Tesseract, GOCR, FineReaderOnline, PWNtcha. Серед них є системи винятково створені для наукових цілей, є комерційна система FineReaderOnline, новітня система з відкритим кодом Tesseract, яка розробляється завдяки фінансуванню корпорацією Google, а також система PWNtcha , яка створена програмістами, спеціально для розпізнавання CAPTCHA. Результати тестування 12 зразків різних типів CAPTCHA вказаними вище декодерами (окрім PWNtcha) наведені на Рис.1. Дані про ефективність PWNtcha отримано з сайту розробника декодера [2]. Для оцінки ефективності автоматичного розпізнавання (%) для кожного типу CAPTCHA використовувалось 10 варіантів зображень. Результати порівняльного аналізу дозволяють зробити наступні висновки:

- використання OCR систем загального призначення не ефективно для автоматичного розпізнавання CAPTCHA;
- декодери, створені спеціально для розпізнавання CAPTCHA, мають високу ефективність і можуть використовуватися для тестування ступеню захисту веб-сайтів.

Розробники CAPTCHA завжди намагаються знайти розумний баланс між двома завданнями: максимально ускладнити розпізнавання програмними методами і при цьому не надто ускладнити прочитання тексту людиною. Згідно з дослідженням, проведеним в Microsoft Research [3], комп'ютерні програми вже на поточному етапі розвитку перевершують людей в розпізнаванні окремих символів. З цієї причини при розробці сучасних CAPTCHA особлива увага приділяється максимальному ускладненню процесу сегментації - автоматичного розділення символів. Робиться це за допомогою додавання різного роду шуму, близького розміщення символів та інших методів. В роботі також проведено дослідження принципів побудови 40 найбільш популярних CAPTCHA, а також їх вплив на стійкість до автоматичного розпізнавання, що дозволяє сформулювати рекомендації до побудови CAPTCHA:

1. Застосування різноманітних шрифтів і кольорів, які кардинально відрізняються один від одного;
2. Застосування кольорових шумів на фоні, причому важливо, щоб хоча б одна складова шуму збігалася з кольором символів;
3. Використання лінійних та нелінійних викривлень – причому параметри цих викривлень завдавати випадково, щоб ускладнити процедуру повернення до нормального вигляду;
4. Ускладнення сегментації шляхом накладання символів один на одного;

5. Спотворення символів шляхом закручування, витягування, тощо.

CAPTCHA	PWNtcha	Ocrad	Tesseract	Gocr	FineReader	Розроблений декодер
	100	0	100	0	0	100
	100	10	100	100	0	100
	100	10	90	90	40	100
	100	0	30	20	0	90
	98	0	10	10	0	100
	100	0	100	100	100	100
	88	0	90	0	0	100
	97	10	40	10	0	100
	100	0	100	100	100	100
	89	0	0	0	0	100
	100	0	0	0	0	100
	49	20	20	20	0	60

Рис. 1. Результати тестування декодерів

Слід зауважити, що застосування цих рекомендацій у повній мірі та збільшення їх показників призводить до покращення ефективності захисту веб-сайтів, але з легкістю може призвести до погіршення юзабіліті CAPTCHA. Тому для загальної оцінки якості CAPTCHA необхідно проводити тестування їх на юзабіліті з використанням якісних показників [4]. Як свідчать результати проведеного тестування з використанням розробленого декодера, імовірність автоматичного розпізнавання символів CAPTCHA залишається дуже високою і для удосконалення системи захисту веб-сайтів слід використовувати додаткові методи, наприклад, метод прихованого поля, фіксування часу, витраченого користувачем, тощо.

Література. **1.** Распознавание некоторых современных CAPTCHA: Блог Хабрахабр, інформаційна безпека (запис у блозі від 25 березня 2011 року) [Електронний ресурс] / Pastafarianist. Режим доступу до блогу: http://habrahabr.ru/post/116222/. **2.** PWNtcha - captcha decoder: сайт розробника ПЗ [Електронний ресурс] Режим доступу до сайту: http://caca.zoy.org/wiki/PWNtcha. **3.** Chellapilla, K., Larson, K., Simard, P., Czerwinski, M. Computers beat humans at single character recognition in reading-based Human Interaction Proofs (HIPs), Microsoft Research 2005.[Електронний ресурс] Режим доступу до сайту: http://research.microsoft.com/en-us/um/people/marycz/czerwinski_cv_june2006_protected.doc. **4.** Jakob Nielsen. Usability 101: Introduction to Usability, 2003. [Електронний ресурс] Режим доступу до сайту: http://www.useit.com/alertbox/20030825.html.