INTERNETINIŲ SISTEMŲ APSAUGOS PRIEMONIŲ NUO BOTŲ PILDOMŲ DUOMENŲ ANALIZĖ

Mindaugas Ruzgys, Simona Ramanauskaitė
Šiaulių universitetas

Įvadas

Šiuo metu internete yra daug informacijos. Nemažą jos dalį sukuria vartotojai, rašydami atsiliepimus, komentarus, diskutuodami forumuose ar dalydamiesi medijų failais. Kiekviena informacinė sistema (IS) turi informacijos įvedimo formų. Jei naudojamos formos neturi apsaugų nuo automatizuotų sistemų (*bot*), vos IS suindeksavus paieškoms, labai tikėtina, kad ims plūsti nepageidaujamas turinys, generuojamas botų.

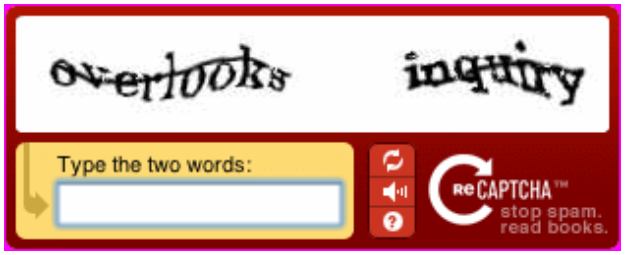
Norint suvaldyti vartotojų įvedamą turinį, reikia skirti papildomų žmogiškųjų ir laiko išteklių, o tai yra nenaudinga verslui. Todėl apsaugos metodų apžvalga ir tyrimas padės pasirinkti tinkamiausią apsaugos strategiją konkrečiai IS.

Tyrimo tikslas – ištirti esamus apsaugos būdus bei strategijas nuo duomenų įvedimo informacinėse sistemose, siekiant apsisaugoti nuo automatizuotų sistemų.

Uždaviniai: išnagrinėti tradicinius, mažiau populiarius ir naujus apsaugos nuo automatizuotų sistemų įrankius; išsiaiškinti minimalius sistemos reikalavimus šių įrankių sėkmingam darbui.

Teksto atpažinimu pagrįstos apsaugos sistemos

Didžioji dalis dabar naudojamų apsaugų pagrįstos žmogaus gebėjimu atpažinti iškraipytą tekstą (Bursztein ir kt., 2010). 1 pav. matome vieno populiariausio šiuo metu *recaptcha* įrankio tipinę užduotį vartotojui (Ahn von L., 2008).



1 pav. reCAPTCHA įrankio, naudojamo nustatyti, ar vartotojas yra žmogus, vaizdas

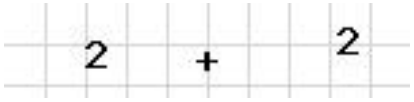
Šio apsaugos tipo populiarumą lemia keletas veiksnių. Pirma – didelis efektyvumas. Teisingai parinkus tekstinę apsaugą, specializuotiems robotams įvestos informacijos apsauga tampa beveik neįveikiama. Tačiau didindamas apsaugos sudėtingumą (teksto iškraipymą), žmogus irgi turi skirti papildomų pastangų, kad atpažintų tekstą. Todėl reikia stengtis išlaikyti optimaliausią variantą, kad patikimai apsaugotų nuo botų ir būtų gana lengva atpažinti žmogui. Kita priežastis, kodėl tekstinės apsaugos sistemos tokios populiarios, – paprastas techninis įgyvendinimas. Pavyzdžiui, viena populiariausių sistemų *recaptcha* teikiama kaip servisas. Dažnai tokių apsaugos sistemų funkcionavimui nereikalinga papildoma programinė įranga ir pakanka numatytųjų įrankių, pavyzdžiui, PHP GD bibliotekos (Bursztein ir kt., 2010). Tekstinės apsaugos priemonės gali būti kelių variacijų. Keletas iškraipyto teksto pateikto atpažinimui pavyzdžių:

- Pateikiami kelių spalvų simboliai, tačiau prašoma įvesti tik tam tikros spalvos simbolius

2 pav. Simbolių spalva pagrįsta tekstinė apsauga

Šio būdo minusas – vartotojas turi suprasti kalbą, kuria parašytas reikalavimas;

- reikalaujama atlikti matematinius veiksmus



3 pav. Matematiniais veiksmais pagrįsta apsauga

Tai gana efektyvu, tačiau kai kurie botai supranta, kad reikia atlikti veiksmus, o ne įrašyti tekstą;

- naudojamas realaus pasaulio paveikslėlis kartu su sugeneruotu tekstu

4 pav. reCAPTCHA variantas naudojant nuotraukas

Šitas būdas veiksmingas, kartu vartotojai prisideda prie „Google“ žemėlapių tikslinimo. Deja, bet kartais paveikslėlis gali būti sunkiai suprantamas ir žmogui;

„ASCII art“ pagrįstos apsaugos sistemos

5 pav. „ASCII art“ tekstinė apsauga

Tai gana veiksmingas būdas, nes užduočiai pateikti nenaudojamas paveikslėlis, o pateikta simbolių seka neturi jokios loginės reikšmės. Tačiau botai, nutaikyti prieš šitokią apsaugą, gali gana nesunkiai ją apeiti imdami momentinį atvaizdą (*screenshot*) ir atpažindami tekstą. Šiuo atveju atpažinti tekstą net lengviau nei įprastos užduoties metu, nes neiškraipomas tekstas ir fonas.

Ne teksto atpažinimu paremtos apsaugos priemonės

Teksto atpažinimu paremtos apsaugos sistemos yra gana veiksmingos, tačiau reikia įvertinti žmogaus pastangas baigti testą. Jei *captcha* bus paini suprasti, ko prašoma, arba bus neįskaitoma, atgraso vartotoją pildyti formą, tarkime, rašyti komentarą, todėl egzistuoja nemažai ne tekstu pagrįstų sprendimų, siekiant apsisaugoti nuo brukalų botu (Shirbhate, 2012).

Vienas paprasčiausių ir dažnai gana veiksmingas apsaugos būdas – paslėptų formos laukelių (*Honeypot*) naudojimas. Veikimo principas: į įvedimo formą įdedami vienas ar keli laukeliai, kurie su CSS (***C**ascading **St**yle **S**heats*) paslepiami, kad vartotojui nebūtų matomi. Vartotojui patvirtinus formą, reikia patikrinti, ar į paslėptus laukelius yra įvesta simbolių. Aptikus, kad laukeliai nėra tušti, darytina prielaidą, kad tai botas, nes pastarieji formoje aptikę laukelį paprastai netikrina, ar jis yra matomas ir jį užpildo. Šitos metodikos privalumas tas, kad vartotojas neapkraunamas papildomais testais ir apie jį apskritai nežino. Trūkumai: pasitaiko, kad vartotojas būna išjungęs CSS palaikymą (naršo su mob. tel. ar pan.). Tokiu atveju pats vartotojas gali užpildyti šiuos laukeliu ir bus palaikytas botu. Tiesa, gana mažai tikėtina, kad taip nutiks.

Kitas apsisaugojimo būdas – realizuoti laiko, per kurį patvirtinama forma, patikrinimo mechanizmą. Paprastai vartotojas negali užpildyti formos ir jos patvirtinti greičiau nei per kelias sekundes, tuo tarpu robotai tai padaro labai greitai – per kelias milisekundes. Todėl galima uždėti limitą, kad forma turi būti minimaliai pildoma, tarkime, 5 sekundes, o jei patvirtinama greičiau, laikoma, kad tai botas. Šio būdo neigiami aspektai gali būti keli ir priklausyti nuo techninio realizavimo technologijos. Pavyzdžiui, naudojant PHP programavimo kalbą, tai galima įgyvendinti pasinaudojus globaliu sesijos masyvu. Tačiau tikrinimą galima atlikti ir su *JavaScript*, kuris vykdomas kliento pusėje, todėl botams tai didelės įtakos nedaro, nes jie retai turi *JavaScript* palaikymą. Dar vienas didelis trūkumas – vartotojas gali pasinaudoti automatizuotais formų pildymo įrankiais. Beveik visos šiuolaikiškos naršyklės turi įrankius, kurie standartinius formų laukelius (vardas, pavardė ir pan.) leidžia užpildyti vienu mygtuko nuspaudimu. Tokiu atveju yra tikimybė, kad vartotojas gali būti palaikytas botu ir nebus leista patvirtinti užpildytos formos.

Dar vienas paprastas ir veiksmingas sprendimas yra į formą įdėti *checkbox* tipo laukelį, kurį vartotojas privalo pažymėti prieš patvirtindamas formą. Šalia pažymėto būtino laukelio dažnai pateikiamas klausimas, liepiantis patvirtinti, kad esate žmogus. Esminis šio būdo aspektas: šis laukelis privalo būti generuojamas dinamiškai, jau įsikrovus puslapiui, su *JavaScript / JQuery*. Tokiu atveju kenkėjiška programa, naudodama *curl* ar panašius įrankius, negali matyti šio laukelio, nes nesuveikia *JavaScript* kodas. Vartotojui patvirtinus formą, telieka patikrinti, ar yra apibrėžta laukelio reikšmė, jei neapibrėžta, darytina prielaida, kad formą bando patvirtinti robotas. Šios apsaugos trūkumas yra tas, kad laukelis generuojamas kliento pusėje. Tai reiškia, kad pas klientą privalo būti įjungta ir tinkamai veikti mūsų pasirinkta *Client Side* technologija, pavyzdžiui *JavaScript*. Taip pat gali kilti problemų naršant su senesnio modelio mobiliaisiais įrenginiais ir *Internet Explorer* naršyklės versijomis, nes jie ne visada sugeba teisingai interpretuoti *JavaScript* kodą.

Pastaruoju metu gana populiari apsaugos nuo botų koncepcija – paveikslėlių naudojimas. Skirtingai nei teksto atpažinimo atveju, paveikslėlyje teksto dažniausiai nėra. Paveikslėliai apsaugai panaudojami keliais būdais. Vienas iš jų vieno ar kelių paveikslėlių pasukamas atsitiktiniu kampu. Tokiu atveju vartotojas su pele spaudinėdamas ant paveikslėlių turi parinkti teisingą jų poziciją. Paveikslėliai turi būti parenkami taip, kad kompiuteriui būtų sunku nustatyti teisingą jo padėtį. Dėl šios priežasties rekomenduojama vengti gamtos vaizdų su horizontu ar žmonių veidų. Pastarųjų poziciją galima nustatyti pasitelkus veidų atpažinimo įrankius (*facial recognition*), todėl botams tampa lengviau „nulaužti“ tokias apsaugas.

6 pav. „Google“ kompanijos pateikiamas pasukto paveikslėlio pavyzdys.

Antras būdas, naudojant paveikslėlius, – paveikslėlio iškraipymas. Tokiu atveju vartotojas turi nustatyti slankiklį į reikiamą padėtį, kad paveikslėlis nebūtų iškraipytas:

7 pav. Iškraipyto paveikslėlio pavyzdys.

Tai gana patikimas ir vartotojo neapkraunantis apsaugos būdas, tačiau turi ir keletą trūkumų. Pirma, pas vartotoją turi būti įjungtas *JavaScript* palaikymas. Kitas

trūkumas tas, kad rinkoje nėra labai daug tokiu pagrindu veikiančių apsaugos sistemų, tuo labiau nemokamų. Pavyzdyje pateikta sistema yra monetizuota, dažnai pateikiama reklaminių tekstų ir siūloma kaip servisas, todėl ją naudodami turime pasitikėti paslaugos teikimo stabilumu.

Pritaikymas neįgaliesiems

Kuriant ir naudojant apsaugos sistemas, reikia atkreipti dėmesį, kad jos turėtų būti pritaikytos žmonėms, turintiems negalią. Šioje vietoje pranašumą turi apsaugos priemonės, kurioms nereikia papildomų žmogaus veiksmų, kad veiksmas būtų užbaigtas (*Optimized for Non-Visual Use*). Prie tokių priemonių galima priskirti laiko, per kurį užpildoma forma, kontrolę arba paslėptų laukelių įdėjimą. Tačiau šios priemonės nėra itin veiksmingos, todėl dažnai naudojamos kaip šalutinės šalia tekstinių ar kitokio pobūdžio testų. Naudojant teksto atpažinimu paremtas apsaugas, turėtų būti galimybė išklausyti garso įrašą. Tokiu būdu silpnai matantys ar visai nematantys žmonės gali sėkmingai užbaigti testą. Garsinė pagalba turėtų būti ir testuose, kuriuose reikia atlikti kažkokius veiksmus su pele, tarkime, nustatyti slankiklį į tam tikrą padėtį.

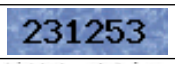
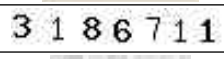
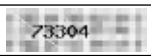
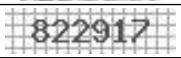
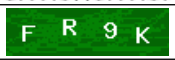
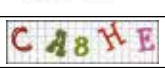
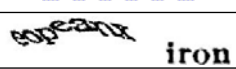
Garsinė pagalba padeda neįgaliesiems, bet kartu gali būti papildoma sistemos saugumo spraga. Robotų kūrėjai gali bandyti analizuoti garso medžiaga, siekdami atpažinti sakomas raides ar kt. Todėl dažnai garsas yra iškraipomas specialiais algoritmais (J. P. Bigham, A. C. Cavender).

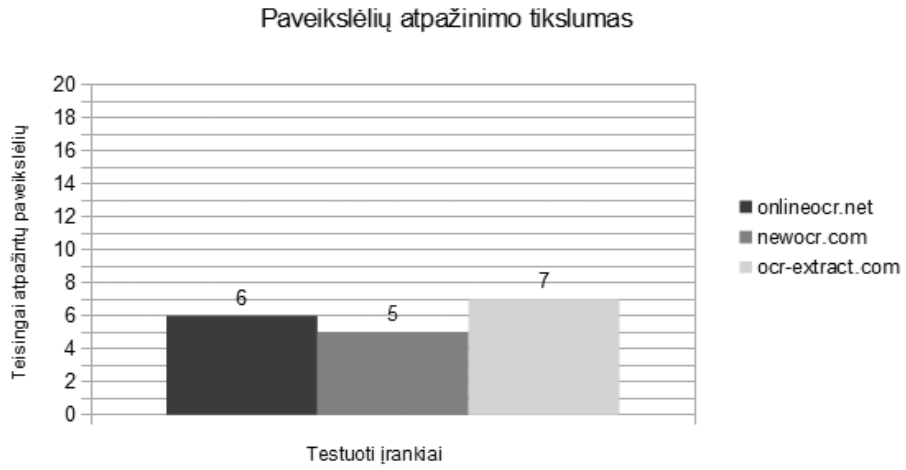
Nors automatizuotas garso atpažinimas ir gali būti rimta saugumo spraga, tačiau iki šiol nėra plačiai žinoma atvejų, kai ja buvo pasinaudota. Dar vienas iškraipyto garso trūkumų: neretai jis tampa sunkiai suprantamas ir žmogui. Pavyzdžiui vienas populiariausias apsaugos įrankis *recaptcha* turi garsinę pagalbą, tačiau nemažai žmonių ja skundžiasi, nes ne anglakalbiam žmogui suprasti smarkiai iškraipytą tekstą gana sunku.

Teksto atpažinimu paremtų apsaugos sistemų patikimumo tyrimas

Siekdami nustatyti, kokią tekstinę apsaugos sistemą pasirinkti, atlikome tyrimą. Tyrimo objektą sudaro

1 lentelė. ***Teksto atpažinimo iš apsaugoje naudojamų paveikslėlių bandymas***

Paveikslėlis	onlineocr.net	newocr.com	ocr-extract.com	Vartotojo vertinamas sudėtingumo lygis
	Atpažintas	Atpažintas	Atpažintas	1
	Neatpažintas	Neatpažintas	Neatpažintas	3
	Neatpažintas	Neatpažintas	Neatpažintas	2
	Neatpažintas	Neatpažintas	Atpažintas	2
	Atpažintas	Neatpažintas	Atpažintas	2
	Atpažintas	Atpažintas	Atpažintas	1
	Neatpažintas	Atpažintas	Atpažintas	3
	Neatpažintas	Atpažintas	Neatpažintas	2
	Atpažintas	Neatpažintas	Atpažintas	2
	Neatpažintas	Neatpažintas	Neatpažintas	2
	Neatpažintas	Neatpažintas	Neatpažintas	2
	Neatpažintas	Neatpažintas	Neatpažintas	1
	Neatpažintas	Neatpažintas	Neatpažintas	2
	Neatpažintas	Neatpažintas	Neatpažintas	3
	Neatpažintas	Neatpažintas	Neatpažintas	3
	Neatpažintas	Neatpažintas	Neatpažintas	4
	Neatpažintas	Neatpažintas	Neatpažintas	6
	Atpažintas	Atpažintas	Atpažintas	1
	Neatpažintas	Neatpažintas	Neatpažintas	4
	Neatpažintas	Neatpažintas	Neatpažintas	7



8 pav. Paveikslėlių atpažinimo tikslumas

20 atsitiktinai pasirinktų tekstinių apsaugų paveikslėlių. Tiriant buvo bandyta keletas nemokamų ar dalinai apribotų internetinių teksto atpažinimo įrankių (Bursztein ir kt., 2012). Šie įrankiai nėra specializuoti apsaugų apėjimui, bet veikia tokiu pačiu principu: pirmiausia pritaikant įvairius filtrus bandoma kuo labiau pašalinti fono įtaką, paskui seka simbolių segmentacija ir galiausiai teksto atpažinimas pasitelkus įvairius euristinius algoritmus ar neuroninius tinklus (K. Chellapilla, P. Y. Simard). Tiesa, neuroniniai tinklai gana veiksmingi atpažįstant skanuotą ar fotografuotą tekstą, tačiau apsaugai naudojami smarkiai iškraipyti simboliai, todėl atpažinimo veiksmingumas smarkiai krenta.

1 lentelėje nurodomas pasirinktas tekstinės apsaugos paveikslėlis ir nurodoma, ar atpažinimo sistemai pavyko jį atpažinti. Šiuo atveju nebuvo bandoma įvertinti atpažinimo sistemų galimybių, o statistiškai įvertinti paveikslėlių atpažįstamumą ir išsiaiškinti to priežastis. Lentelėje pateikiama subjektyvi nuomonė apie teksto atpažinimo sunkumą žmogui. Vertinama nuo 1 (lengvai atpažįstamas) iki 10 (visiškai neatpažįstamas). Kadangi tai yra subjektyvi autoriaus nuomonė, į išvadas ir tolesnę analizę šitie duomenys nebus įtraukiami.

Iš trijų išbandytų internetinių įrankių, skirtų tekstui iš paveikslėlių išskirti, visos programos pasirodė apylygiai. Skirtumas tarp geriausiai pasirodžiusio įrankio (*www.ocr-extract.com* 7 atpažinti paveikslėliai) ir blogiausiai pasirodžiusio įrankio (*www.newocr.com* 5 atpažinti paveikslėliai) – tik du paveikslėliai.

Pažymėtina, kad kai kuriuos paveikslėlius atpažino visi bandyme išbandyti įrankiai, o didžiosios dalies nepavyko atpažinti nė vienam įrankiui.

Paanalizavę geriausiai atpažįstamų paveikslėlių turinį, matome, kad juose vyrauja nesudėtingas fonas, o simboliai išdėstyti beveik tiesia linija, nepasukus atsitiktiniu kampu. Dėl šių priežasčių atpažinimo programos lengvai atlieka paveikslėlio segmentaciją ir atpažinimą. Tuo tarpu paveikslėliuose, kuriuose vyrauja žymus fonas ir iškraipyti simboliai, nebuvo atpažinta nė vienos bandytos programos.

Tai parodo, kad renkantis teksto atpažinimu paremtą apsaugą reikia atkreipti dėmesį į jos sudėtingumą botams.

Išvados

1. Išsiaiškinta, kad rinkoje populiariausios apsaugos priemonės, nuo botų – teksto atpažinimu grįstos sistemos. Jų paplitimą lemia didelis veiksmingumas ir nesudėtingas techninis įgyvendinimas. Daugumai IS tai patikimiausias ir veiksmingiausias sprendimas.
2. Renkantis teksto atpažinimu paremtą apsaugos sistemą reikia atkreipti dėmesį, kad paveikslėlio fonas ir pats tekstas būtų pakankamai iškraipomas ir sunkiai atpažįstamas botų. Kitu atveju sistema bus labai pažeidžiama, nes paveikslėlis su tekstu bus lengvai išanalizuojamas.
3. Renkantis apsaugos sistemą, nereikia pamiršti, kad apsaugotomis sistemomis naudojasi ir neįgalieji, kuriems gali būti sunku įveikti vizualines apsaugas, todėl reikėtų pasirūpinti, kad būtų garsinė pagalba atliekant testą.

Literatūra

1. Bursztein E., Martin M., Mitchell J. C., 2010, *Text-based CAPTCHA Strengths and Weaknesses*. Prieiga per internetą: http://www.cin.ufpe.br/~rsc3/temp/text-based-captcha-strengths-and-weaknesses.pdf.
2. Ahn von L., 2008, reCAPTCHA: Human-Based Character Recognition via Web Security Measures. Prieiga per internetą: http://users.df.uba.ar/marcos/reCAPTCHA.pdf.
3. Shirbhate P., Dhamankar V., Kshirsagar A., Deshpande P. & Kapse S., 2012, *Overview of Honeypot Security System for E-Banking*. Prieiga per internetą: http://www.irnetexplore.ac.in/IRNetExplore_ Volumes/UARJ/UARJ_doc/Volume%201%20Issue%201/paper22.pdf.
4. Bigham J. P., Cavender A. C., *Evaluating Existing Audio CAPTCHAs and an Interface.*
5. *Optimized for Non-Visual Use*. Prieiga per internetą: http://www.annacavender.com/downloads/captchachi09.pdf.
6. Chellapilla K., Simard P. Y. *Using Machine Learning to Break Visual Human Interaction Proofs (HIPs)*. Prieiga per internetą: http://student.eepisits.edu/~henry/upload/NIPS2004_0843.pdf.

ANALYSIS OF TOOLS FOR PROTECTIONAGAINST DATA FLOODING IN INTERNET SYSTEMS

Mindaugas Ruzgys, Simona Ramanauskaitė

Summary

The amount of data on the Internet is growing every day and has never been as large as now. This is influenced by the fact that information is being published on the Internet not only by its authors, but also by other users . However, ensuring the correctness of the data is resource and time consuming. Therefore, the need of protection tools against automated data submission has risen.

In this work, existing techniques and tools for automated data submission to Internet systems are analyzed. The list of traditional and less popular solutions is presented. The research of different text-based protection tools revealed what kind of text-based protection against bots is more suitable to protect Internet systems.

Key words: tools for protection against data flooding, CAPTCHA.

INTERNETINIŲ SISTEMŲ APSAUGOS PRIEMONIŲ NUO BOTŲ PILDOMŲ DUOMENŲ ANALIZĖ

Mindaugas Ruzgys, Simona Ramanauskaitė

Santrauka

Šiuo metu informacijos kiekiai internete milžiniški, nes dalį šios informacijos sukuria ne puslapių bei sistemų autoriai, o jų naudotojai. Norint tikrinti ir įvertinti vartotojų įvedamo turinio tinkamumą, reikia žmogiškųjų bei laiko išteklių, todėl atsiranda poreikis automatizuoti apsaugos nuo kenkėjiško turinio mechanizmą.

Šiame darbe analizuojami ir aptariami apsaugos nuo automatizuotų sistemų (*bot*) metodai bei įrankiai. Pateikiami tradiciniai ir mažiau paplitę apsaugos būdai. Tiriant nustatyta, kokias reiktų pasirinkti teksto atpažinimu pagrįstas sistemas, kad būtų galima patikimai apsisaugoti nuo botų keliamų grėsmių.

Prasminiai žodžiai: apsaugos nuo botų priemonės, *captcha*.

Įteikta 2013-05-14