Quand les filtres anti-spam améliorent les logiciels d'OCR

Publié le vendredi 1er juin 2007
Voir en ligne : https://www.france-science.org/Quand-les-filtres-anti-spam.html

Une équipe de Carnegie Mellon University vient de proposer une amélioration de son système de filtre sur des pages Internet connu sous le nom Completely Automated Public Turing Test to Tell Computers and Humans Apart (CAPTCHA). Le procédé est désormais courant : des lettres déformées et barrées apparaissent à l'écran et l'utilisateur doit trouver le mot qu'elles forment. Le mécanisme avait été à l'origine commandé par Yahoo pour éviter l'enregistrement de comptes e-mail frauduleux par des bots. Il a été depuis largement repris par eBay, Hotmail, Paypal et dans de nombreux sites d'entreprises et de particuliers pour s'assurer que la saisie de données a bien été effectuée par un humain et non pas par une machine.

L'évolution de CAPTCHA a été baptisée reCAPTCHA. Elle tire profit d'une double constatation : plus de 60 millions de CAPTCHA sont résolus chaque jour, chaque test prenant environ 10 secondes. Au total, plus de 150.000 heures d'analyse humaine sont ainsi consommées quotidiennement. D'un autre côté, de nombreux projets d'OCR (Optical Character Recognition) sont ralentis dans la numérisation de livres à cause de mots non reconnus, déformés par le vieillissement des ouvrages ou par une mauvaise qualité d'acquisition.

reCAPTCHA apporte une solution originale : le système propose deux mots à déchiffrer. Le premier est un mot connu qui sert à s'assurer que la personne en face de l'écran est bien un humain. Le deuxième est un mot rejeté par un logiciel d'OCR, déformé comme pour un CAPCHA classique. Quand un certain nombre d'utilisateurs a déchiffré de la même manière le deuxième mot, celui-ci est validé et peut alors être substitué dans le processus de numérisation du livre. Cette méthode est aujourd'hui utilisée au profit de l'Internet Archive, une association à but non lucratif de l'Open Content Alliance, pour l'aider dans son travail de numérisation de contenus libres de droits.

Intel a développé un web-service qui permet à tous les webmestres d'utiliser le mécanisme de filtrage sur leur site Internet. Il est disponible sur le site de reCAPTCHA sous forme de plugins compatibles avec la plupart des CMS (WordPress, MediaWiki, Typo3, …), de bibliothèques pour langages orientés web (PHP, Python, Perl, Ruby) et d'une API Java. Les avantages du web-service sont doubles : au cas où un logiciel arriverait à déchiffrer le premier mot, le système de déformation des lettres peut être mis à jour sans intervention des webmestres. De plus, le service inclut un système de détection et de bannissement d'adresses IP pour ces logiciels frauduleux.

Par la même occasion, l'équipe de Carnegie Mellon University propose sur son site un service gratuit de protection d'adresse mail par reCAPTCHA et une version audio du système de protection pour les personnes malvoyantes. L'objectif clairement affiché est de remplacer un maximum de CAPTCHA par leur successeur.

Source :

- Carnegie Mellon Project Boosts Book Digitization Efforts, 24/05/2007
http://www.cmu.edu/news/archive/2007/May/may24_recaptcha.shtml
- A new twist on anti-spam tech can help digitize books, 25/05/2007
http://arstechnica.com/news.ars/post/20070525-anew-twist-on-anti-spam-tech-can-help-digitize-books.html

Pour en savoir plus, contacts :

- Site officiel de reCAPTCHA par Carnegie Mellon University : http://recaptcha.net/
- Site Internet de l'Internet Archive : http://www.archive.org/index.php

Code brève
ADIT : 43050

Rédacteur :

Vincent Reboul, deputy-stic.mst@ambafrance-us.org