TEMPS DE CALCUL HUMAIN DISPONIBLE

LE 12 OCTOBRE 2010 ***ROUD***

Tous les jours, 100 millions de CAPTCHA sont déchiffrés par les internautes. Une tâche humaine que le service reCAPTCHA met à profit pour combler les déficiences des ordinateurs. Le but : améliorer la numérisation de livres.

Où comment même ***le troll*** *peut contribuer au progrès de l'humanité…*

De très nombreux blogs utilisent des **CAPTCHA** pour empêcher les robots spammeurs de mener leur sombre ouvrage. Ce joli nom aux consonances un peu soviétiques est l'acronyme de "*Completely Automated Public Turing test to tell Computers and Humans Apart*", soit en bon français "test de Turing Public Complètement Automatisé permettant de reconnaître les Ordinateurs des Humains".

Déchiffrer un CAPTCHA
Turing, l'un des grands génies multidisciplinaires du XXe siècle, avait proposé **le test suivant** : imaginez que vous chattiez avec deux interlocuteur inconnus, l'un étant un homme, l'autre une machine, comment feriez-vous pour les distinguer ? (dans un genre un peu différent il y a **le fameux psy d'Emacs**). A priori, plus la machine est intelligente, plus il va vous falloir du temps pour la distinguer de l'homme.

La voie la plus rapide est de tester les compétences analytiques et synthétiques de votre interlocuteur dans des domaines pour lesquels l'ordinateur est toujours bien inférieur à l'homme. Un exemple est la reconnaissance de formes, utilisée donc dans nos CAPTCHA : avant de pouvoir laisser un commentaire, l'interlocuteur doit reconnaître des caractères déformés, et les taper dans une fenêtre.

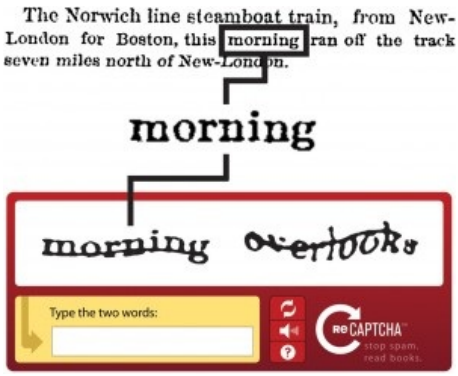
Si cette tâche est typiquement assez difficile pour un ordinateur, elle reste très facile pour l'homme, qui en une fraction de seconde mobilise les ressources de son puissant cerveau pour déchiffrer le CAPTCHA (et avoir le droit de laisser son commentaire pertinent ou son ***flame***).

Du gâchis ? Plus maintenant !
Mais quel gâchis quand on y pense ! Tous les jours, 100 millions de CAPTCHA sont déchiffrés sur les blogs, sites, etc. C'est autant de tâches d'analyse complexe insolubles par des ordinateurs (par définition) et réalisées par des internautes. D'où l'idée derrière **reCAPTCHA** : "recycler" ce temps de calcul humain disponible gratuitement pour suppléer les ordinateurs sur les problèmes qu'ils ne savent pas à résoudre. Les détails de l'algorithme et la philosophie de la méthode sont exposés dans **un article** paru cette semaine dans la prestigieuse revue *Science*.

Le but du jeu ici est d'aider à la numérisation de livres, afin de préserver la connaissance humaine et de la rendre accessible au plus grand nombre. Les pages des livres sont scannées, et chaque mot est transformé en image bitmap par un logiciel approprié. Un autre logiciel de reconnaissance optique de caractères essaie ensuite de reconstituer le mot à partir de cette image.

Le problème est qu'environ 20% des mots ne peuvent être reconnus automatiquement par les logiciels. C'est là qu'entre en jeu reCAPTCHA : il utilise les capacités intellectuelles des internautes laissant des commentaires sur les blogs pour lire les mots que les ordinateurs ne peuvent pas lire.

En effet, reCAPTCHA vous demande de reconnaître deux mots pour pouvoir laisser un commentaire. Pour l'un des mots, reCAPTCHA connaît la réponse : il s'agit d'un des mots déjà déchiffrés par d'autres humains, mais non reconnaissable par un ordinateur. C'est sur ce mot-ci, appelé "mot de contrôle" qu'on testera si vous êtes vraiment un internaute ou un robot spammeur.



L'autre mot est le mot inconnu que l'ordinateur ne sait pas déchiffrer, le mot "test". C'est sur ce mot qu'on vous mettra réellement à contribution : reCAPTCHA va comparer votre proposition aux propositions d'autres internautes sur ce mot. Si les trois premières personnes tombant sur ce mot proposent la même lecture et qu'il n'est pas reconnaissable automatiquement, ce mot est considéré comme déchiffré et sera utilisé comme mot de contrôle futur. Si les réponses humaines sont plus variables, ce mot reste comme un mot test et on additionne le nombre de possibilités proposées par les internautes pour ce mot. Dès qu'une proposition a plus de 2.5 voix **[1]** , ce mot est considéré comme lu.

Évidemment, il y a des cas où les mots ne sont pas du tout lisibles. Dans ce cas, vous pouvez demander un autre CAPTCHA à reCAPTCHA – il y a un petit bouton "*reload*" à côté du CAPTCHA. Si plus de six utilisateurs demandent à changer de CAPTCHA pour un même mot, le mot est considéré définitivement comme illisible et est sorti des bases de données de reCAPTCHA **[2]** .

L' efficacité de reCAPTCHA a été évaluée sur 50 articles extraits des archives du *New York Times* entre 1860 et 1970, pour un total de 24 080 mots. 99.1 % des mots ont pu être déchiffrés par reCAPTCHA (216 erreurs), contre 83.5 % des mots pour de simples logiciels de reconnaissance automatique. Un taux de 99 % est en général le taux de réussite "professionnel" : pour l'anecdote, un des professionnels humains "témoin" travaillant sur les mêmes textes a effectué 189 erreurs, presque autant que reCAPTCHA.

Cependant, les erreurs de reCAPTCHA et les erreurs humaines sont de natures différentes : reCAPTCHA est plutôt sensible aux erreurs de "reconnaissance" des mots, par exemple le logiciel va couper des mots au milieu ou au contraire grouper des mots ensemble, tandis qu'un humain va "bêtement" se tromper en faisant une faute typographique ou orthographique…

reCAPTCHA fonctionne depuis 2007. Après un an, il avait été installé sur 40 000 sites web et 1.2 milliards de CAPTCHA avaient été déchiffrés, soit 440 millions de mots déchiffrés correctement. Si on considère que 25 % des mots dans un livre scanné sont mal reconnus, cela correspond à la bagatelle de 17 600 livres transcrits manuellement. En 2008, les créateurs de reCAPTCHA estimaient que l'équivalent de 160 livres par jour sont déchiffrés par reCAPTCHA, qui fournit l'équivalent du travail de 1 500 personnes déchiffrant un mot par seconde et travaillant 40 h par semaine.

En plus d'être utile, reCAPTCHA est plus efficace qu'un CAPTCHA traditionnel : en effet, les algorithmes des robots spammeurs peuvent "apprendre" à lire des CAPTCHA dans la mesure où les déformations habituelles sur les CAPTCHA sont faites numériquement. Or, les distorsions des mots imprimés sont beaucoup plus aléatoires, puisqu'il s'agit de "vraies" distorsions dues à des problèmes d'impression, des problèmes du papier, sans compter le bruit numérique dû au passage au scanner, etc.

Du crowdsourcing à toutes les sauces
Comme l'expliquent les auteurs, reCAPTCHA est la mise en pratique d'une idée fascinante, qu'ils appellent "**human computation**" (pouvoir de calcul humain ?). C'est une petite rupture dans notre vision de l'ordinateur : au lieu d'essayer d'améliorer les machines pour en faire des équivalents humains, tâche peut être simplement impossible à long terme, on utilise la puissance d'internet pour mettre en réseau des hommes afin de résoudre les problèmes complexes insolubles par la puissance de calcul brute des ordinateurs.

On parle ici de numérisation de livres, mais il existe aussi d'autres projets, comme **Fold It**, un jeu en ligne dans lequel les gens essaient de déterminer la structure des protéines, ou encore **Galaxy Zoo dont nous avait parlé Dr Goulu**. On peut imaginer que pour les grands défis numériques de demain, la mission de l'ordinateur ne sera alors que la mise en réseau et l'exécution en parallèle de tâches lourdes mais simples ; mais le vrai pouvoir de pensée, la vraie créativité resteront humaines.

J'aime bien cette idée également pour son corollaire un peu utopique. Qui pourra breveter une protéine dont la structure aura été trouvée grâce au calcul humain parallèle volontaire ? L'utilisation massive et volontaire des capacités des gens ne pourra pas être marchandisée, et le pouvoir de calcul humain deviendrait alors un pouvoir tout court, authentiquement démocratique …

[1] Une proposition d'un logiciel de reconnaissance optique de caractères comptant pour 0.5 voix
[2] Je suis personnellement un peu paresseux et j'ai tendance à abuser de ce bouton, dorénavant j'essaierai quand même de proposer quelque chose…

Références
Le site de **reCAPTCHA**.
Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, Manuel Blum, "**reCAPTCHA: Human-Based Character Recognition via Web Security Measures**", *Science* 12 September 2008: Vol. 321. no. 5895, pp. 1465 – 1468
Une vidéo d'une conférence de Luis von Ahn
Enfin, Dr Goulu avait déjà publié deux billets sur le même sujet : **les Ordinateurs Humains : des Captchas à PeekaSearch** et **ReCaptcha : quand l'internet utilise les cerveaux humains**

>> Ce billet a été initialement publié sur le blog de Tom Roud.

JICE - SITES SCIENTIFIQUES
le 29 novembre 2010 - 14:53 • SIGNALER UN ABUS - PERMALINK

J'ai aussi récemment publié un rapide article sur l'utilisation marketing de ces outils du web :
http://jice.lavocat.name/blog/2010/11/la-pub-tuera-t-elle-les-spammeurs/

C'est plutôt astucieux aussi, mais on ne voit toujours pas pointer de captcha vraiment porteur de projet humanitaires pour le moment.

VOUS AIMEZ 0 VOUS N'AIMEZ PAS 0 LUI RÉPONDRE

2 pings

Zoouniverse, classification partagée d'objets celestes le 26 novembre 2010 - 11:25

[...] à zoouniverse le but est de classifier des objets astronomiques, c'est un article de owni sciences sur la « disponibilité du temps de calcul humain qui m'en a donné [...]

Le retour en grâce d'Alan Turing » Article » OwniSciences, Société, découvertes et culture scientifique le 4 février 2011 - 14:59

[...] Temps de calcul humain disponibleLa reconnaissance vocale est morte : pet à son âmeScience Blogs : "Un modèle complètement nouveau" pour le GuardianWikileaks et biologie, utilisation similaire des données?L'influence de Mandelbrot dans la finance [...]