PHYS.ORG

Computer users are digitizing books quickly and accurately with Carnegie Mellon method

August 14 2008

Millions of computer users collectively transcribe the equivalent of 160 books each day with better than 99 percent accuracy, despite the fact that few spend more than a few seconds on the task and that most do not realize they are doing valuable work, Carnegie Mellon University researchers reported today in *Science Express.*

They can work so prodigiously because Carnegie Mellon computer scientists led by Luis von Ahn have taken a widely used Web site security measure, called a CAPTCHA, and given it a second purpose — digitizing books produced prior to the computer age. When Web visitors solve one of the distorted-letter puzzles so they can register for email or post a comment on a blog, they simultaneously help turn the printed word into machine-readable text.

More than a year after implementing their version, called reCAPTCHA, recaptcha.net/ on thousands of Web sites worldwide, the researchers conclude that their word deciphering process achieves the industry standard for human transcription services — better than 99 percent accuracy. Their report, published online today, will appear in an upcoming issue of the journal *Science.*

Furthermore, the amount of work that can be accomplished is herculean. More than 100 million CAPTCHAs are solved every day and, though each puzzle takes only a few seconds to solve, the aggregate amount of time translates into hundreds of thousands of hours of human effort that can potentially be tapped. During the reCAPTCHA system's first year of operation, more than 1.2 billion reCAPTCHAs have been solved and more than 440 million words have been deciphered. That's the equivalent of manually transcribing more than 17,600 books.

"More Web sites are adopting reCAPTCHAs each day, so the rate of transcription keeps growing," said von Ahn, an assistant professor in the School of Computer Science's Computer Science Department. "More than 4 million words are being transcribed every day. It would take more than 1,500 people working 40 hours a week at a rate of 60 words a minute to match our weekly output."

Von Ahn said reCAPTCHAs are being used to digitize books for the Internet Archive and to digitize newspapers for The New York Times. Digitization allows older works to be indexed, searched, reformatted and stored in the same way as today's online texts.

Old texts are typically digitized by photographically scanning pages and then transforming the text using optical character recognition (OCR) software. But when ink has faded and paper has yellowed, OCR sometimes can't recognize some words — as many as one out of every five, according to the Carnegie Mellon team's tests. Without reCAPTCHA, these words must be deciphered manually at great expense.

Conventional CAPTCHAs, which were developed at Carnegie Mellon, involve letters and numbers whose shapes have been distorted or backgrounds altered so that computers can't recognize them, but humans can. To create reCAPTCHAs, the researchers use images of words from old texts that OCR systems have had trouble reading.

Helping to make old books and newspapers more accessible to a computerized world is something that the researchers find rewarding, but is only part of a larger goal. "We are demonstrating that we can take human effort — human processing power — that would otherwise be wasted and redirect it to accomplish tasks that computers cannot yet solve," von Ahn said.

For instance, he and his students have developed online games, available at www.gwap.com , that analyze photos and audio recordings — tasks beyond the capability of computers. Similarly, University of Washington biologists recently built Fold It, fold.it/ , a game in which people compete to determine the ideal structure of a given protein.

In addition to von Ahn, authors of the new report include computer science undergraduate Benjamin Maurer, graduate students Colin McMillen and David Abraham, and Manuel Blum, professor of computer science.

Source: Carnegie Mellon University

Citation: Computer users are digitizing books quickly and accurately with Carnegie Mellon method (2008, August 14) retrieved 27 April 2024 from https://phys.org/news/2008-08-users-digitizing-quickly-accurately-carnegie.html