

August 2, 2007

Contact: Hugh Powell (831) 459-4353, hpowell@ucsc.edu; Tim Stephens (831) 459-2495, stephens@ucsc.edu

New program color-codes text in Wikipedia entries to indicate trustworthiness

The online reference site Wikipedia enjoys immense popularity despite nagging doubts about the reliability of entries written by its all-volunteer team. A new program developed at the University of California, Santa Cruz, aims to help with the problem by color-coding an entry's individual phrases based on contributors' past performance.

The program analyzes Wikipedia's entire editing history--nearly two million pages and some 40 million edits for the English-language site alone--to estimate the trustworthiness of each page. It then shades the text in deepening hues of orange to signal dubious content. A 1,000-page demonstration version is already available on a web page operated by the program's creator, Luca de Alfaro, associate professor of computer engineering at UCSC.



Luca de Alfaro

Other sites already employ user ratings as a measure of reliability, but they typically depend on users' feedback about each other. This method makes the ratings vulnerable to grudges and subjectivity. The new program takes a radically different approach, using the longevity of the content itself to learn what information is useful and which contributors are the most reliable.

"The idea is very simple," de Alfaro said. "If your contribution lasts, you gain reputation. If your contribution is reverted [to the previous version], your reputation falls." De Alfaro will speak about his new program this Saturday, August 4, at the Wikimania conference in Taipei, Taiwan.

The program works from a user's history of edits to calculate his or her reputation score. The trustworthiness of newly inserted text is computed as a function of the reputation of its author. As subsequent contributors vet the text, their own reputations contribute to the text's trustworthiness score. So an entry created by an unknown author can quickly gain (or lose) trust after a few known users have reviewed the pages.

A benefit of calculating author reputation in this way is that de Alfaro can test how well his reliability scores work. He does so by comparing users' reliability scores with how long their subsequent edits last on the site. So far, the program flags as suspect more than 80 percent of edits that turn out to be poor. It's not overly accusatory, either: 60 to 70 percent of the edits it flags do end up being quickly corrected by the Wikipedia community.

The exhaustive analysis of Wikipedia's seven-year edit history takes de Alfaro's desktop PC about a week to complete. At present he is working from copies of the site that Wikipedia periodically distributes. Once the initial backlog of edits is calculated, however, de Alfaro said that updating reliability scores in real time should be fairly simple.

While the program prominently displays text trustworthiness, de Alfaro favors keeping hidden the reputation ratings of individual users. Displaying reputations could lead to competitiveness that would detract from Wikipedia's collaborative culture, he said, and could demoralize knowledgeable contributors whose scores remain low simply because they post infrequently and on few topics.

"We didn't want to modify the experience of a user going in to Wikipedia," de Alfaro said. "It is very relaxing right now and we didn't want to modify what has worked so well and is so welcoming to the new user."

De Alfaro's color-coded Wikipedia pages can be found on his [demonstration site \(http://trust.cse.ucsc.edu/\)](http://trust.cse.ucsc.edu/).

Note to reporters: You may contact de Alfaro at (831) 459-4982 or luca@soe.ucsc.edu. (He will be in Taiwan August 1-7.)

####