Developing a crowdsourcing strategy for SIMSSA

Charalampos Saitis

July 23, 2014

1 Background

Humans are critical components for performing quality control in a recognition process, correcting the inevitable errors that automated systems make and ensuring these errors do not compound themselves in subsequent workflow steps (Hankinson et al. 2012). Correction, however, can be very time, labour and cost intensive. In the case of optical character recognition (OCR)—the textual counterpart of OMR—some unique solutions have been developed to help offset the costs of this task. The Australian Newspapers Digitisation Project (ANDP) (Holley 2009) has created a distributed correction system, where more than 9,000 volunteers have now corrected more than 12.5 million lines of text, with more corrections added all the time. The reCAPTCHA project (von Ahn et al. 2008) has produced over 5 billion human-corrected OCR words by presenting the correction task as a spam-fighting challenge to prove that the corrector is a human and not an automated system.

Both these examples are manifestations of crowdsourcing, a Web-enabled, distributed problem-solving model that has emerged during the past decade (Brabham 2008). A crowdsourcing system is a system where a large number of people, known as contributors, are enlisted to help solve a problem defined by the system owners, which would normally require intensive (and often tedious), costly labour. Apart from helping reaching out to potentially thousands of users around the globe, the Internet offers a high degree of automation and unique possibilities for user management (e.g., through social software such as wiki, discussion group, blogging and tagging) (Doan et al. 2011). Although the term *crowdsourcing* was coined in 2006 (Howe 2006), Linux (released in 1991) and Wikipedia (created in 2001) are two prime examples of such systems.

1.1 Something

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum

ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

1.1.1 Challenges

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

References

Brabham, D. C. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence: The International Journal of Research into New Media Technologies* 14 (1): 75–90.

Doan, A., R. Ramakrishnan, and A. Y. Halevy. 2011. Crowdsourcing systems on the World-Wide Web. *Communications of the ACM* 54 (4): 86–96.

Hankinson, A., J. A. Burgoyne, G. Vigliensoni, A. Porter, J. Thompson, W. Liu, R. Chiu, and I. Fujinaga. 2012. Digital document image retrieval using optical music recognition. In *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, 577–582.

Holley, R. 2009. Many hands make light work: Public collaborative OCR text correction in Australian historic newspapers. *National Library of Australia Staff Papers*.

Howe, J. June 2006. The rise of crowdsourcing. *Wired*.

von Ahn, L., B. Maurer, C. McMillen, D. Abraham, and M. Blum. 2008. reCAPTCHA: Human-based character recognition via web security measures. *Science* 321: 1465–1468.