Developing a crowdsourcing strategy for SIMSSA

Charalampos Saitis

July 23, 2014

Contents

1 Introduction

The Single Interface for Music Score Searching and Analysis (SIMSSA) project is a recently funded initiative to investigate the development of a fully searchable digital musical document collection, including document digitization, high-volume optical music recognition (OMR), large-scale symbolic search systems, and digital document display. The goal of SIMSSA is to teach computers to recognize the musical symbols in digitized notation and assemble the data on a single website, making online search and analysis of musical scores possible for the first time. Scholars and the general public will be able to consult millions of musical works in the blink of an eye, a task that would previously have required many lifetimes.

Central to the SIMSSA project is the use of Web 2.0 technologies to facilitate the development of online crowdsourcing tools for the collection of large amounts of human-checked transcriptions. Distributed proofreading and correction can be potentially performed by anyone with a web browser and an Internet connection, anywhere in the world, thereby allowing for a constantly-improving search and retrieval system (adaptive OMR (Fujinaga 1996)), as well as for the collection of human-provided ground-truth data to further reduce the amount of error in the final product while maintaining a level of quality and efficiency beyond simple automatic recognition (Hankinson et al. 2012).

In the following sections we outline the prime challenges facing the design of the crowd-sourced, adaptive OMR system envisioned by the SIMSSA initiative. This position paper is part of an ongoing research programme that aims to provide SIMSSA with a theoretical framework for developing effective crowdsourcing tools.

http://simssa.ca

2 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.

3 Hello

Forecasting is the art of saying what is going to happen and then to explain why it didn't.

(Anonymous)

3.1 Challenges

Unlike small-scale public engagement activities such as tagging or rating a song, crowdsourcing requires a greater level of commitment from the contributor in terms of effort, time and intellectual input, as well as continuous assessment by the system owners (Holley 2010). Given these considerations, we have identified the following key challenges for integrating crowdsourcing in the SIMSSA project:

- **How to engage contributors?** According to Doan, Ramakrishnan, and Halevy (Doan et al. 2011), one can either *require users* (e.g, students may be required to help correct transcriptions as part of the class curriculum), *make users "pay" for service* (i.e., *implicit* collaboration; e.g., reCAPTCHA (von Ahn et al. 2008)), *piggyback on the user traces* (e.g., search query logs may be used for spelling corrections (Ahmad and Kondrak 2005) or keyword generation (Fuxman et al. 2008)), or simply *ask for volunteers.*

- **How to reward and retain contributors?** Provided that the contributions of a user are reliable (see next challenge), a strategy to reward and retain them should be encouraged. For example, users may be offered exclusive beta versions of certain (possibly relevant) software or free access to digital libraries. Ideas may also be drawn from the emerging area of *gamification* (Deterding et al. 2011): users can earn badges that they can then redeem as points in online games; high-volume data collection and verification can be turned into a game where participants are rewarded points for their work (e.g., the ESP game (von Ahn and Dabbish 2004); MajorMinor (Mandel and Ellis 2008)).

- **How to evaluate users and their contributions?** Crowdsourcing, as with any open call to an undefined group of people, is prone to incompetent and sometimes malicious use. A manual approach to this challenge includes the further distribution of contributions by ordinary users to an independent team of trusted users, whereby "bad" work is flagged on message boards (Doan et al. 2011). For example, the same task can be assigned to 3 contributors. If one result differs substantially from the other two, then that contributor gets flagged. The less flags a contributor has, the more trustwordy he is. Such reliability testing can also be done on the fly: a contributor can be given a correction task for which the system already knows the solution, then flagged based on their response (Doan et al. 2011).

Doan et al. (Doan et al. 2011) have highlighted 2 more challenges: what type of tasks can users do and how to combine their contributions to solve the target problem. These questions are equally important and we aim to explore solutions based on both hypothetical and real-world scenarios.

To a certain extent, these challenges are technical. However, large-scale musical databases are becoming increasingly central to musicological scholarship, with computational tools triggering new research questions or helping to re-approach previous topics. We hope that these experiences and tools will result in essential infrastructure components for creating next-generation globally accessible digital music libraries for the future musicologist.

Bibliography

Ahmad, F., and G. Kondrak. 2005. Learning a spelling error model from search query logs. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language*, Vancouver, Canada, 955–962.

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.

Deterding, S., M. Sicart, L. Nacke, K. O'Hara, and D. Dixon. 2011. Gamification: Using game-design elements in non-gaming contexts. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, Vancouver, Canada, 2425–2428.

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

Fujinaga, I. 1996. *Adaptive Optical Music Recognition*. Ph. D. thesis, Faculty of Music, McGill University, Montreal, Quebec, Canada.

Fuxman, A., P. Tsaparas, K. Achan, and R. Agrawal. 2008. Using the wisdom of the crowds for keyword generation. In *Proceedings of the 17th international conference on World Wide Web*, Beijing, China, 61–70.

Hankinson, A., J. A. Burgoyne, G. Vigliensoni, and I. Fujinaga. 2012. Creating a large-scale searchable digital collection from printed music materials. In *Proceedings of the 21st International Conference Companion on World Wide Web*, Lyon, France, 903–908.

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*.

Holley, R. 2010. Crowdsourcing: How and why should libraries do it? *D-Lib Magazine* 16 (3/4).

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

Mandel, M. I., and D. P. W. Ellis. 2008. A web-based game for collecting music metadata. *Journal of New Music Research* 37 (2): 151–165.

von Ahn, L., and L. Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, Vienna, Austria, 319–326.

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.