Crowdsourcing in the Document Processing Practice
(A Short Practitioner/Visionary Paper)

Ehud D. Karnin, Eugene Walach, and Tal Drory

IBM Research Haifa Lab, Haifa University Campus, Haifa 31905, Israel
{karnin,walach,tald}@il.ibm.com

Abstract. The processing of scanned documents calls for automatic recognition of the text by OCR (Optical Character Recognition) computer programs, followed by human validation and correction. Crowdsourcing of these essential manual tasks is a good option, provided one can take care of some key challenges, so that the quality level expected by the customer is met. We show how tools for efficient validation and correction are adapted and enhanced to address issues associated with crowdsourcing, such as data privacy, quality control, crowd monitoring, and job quality assurance. We started to implement these ideas and technologies in our COoperative eNgine for Correction of ExtRacted Text (CONCERT), which is used in book digitization projects.

Keywords: Enterprise Crowdsourcing, documents processing, quality control, productivity tools, quality assurance.

1 Introduction

The process of recognizing text in images is a huge business, which is prevalent in virtually every industry. Forms (e.g., medical and insurance claims) and checks processing, license plate recognition, and book digitization are just few examples. Despite advances in automatic recognition by OCR programs, there are always characters which cannot be identified, or recognized with a low confidence level. The risk of character substitution, which may be painful in many cases, drives the thresholds up. In other words, only high confidence characters are accepted, and all remaining characters need to be verified or corrected by a human operator.

The naïve way for an operator to work is to get the form with the unrecognized or unsure characters highlighted in some way within the image of the document. Operators move over these areas on the form and correct where necessary. This is very inefficient; therefore other methodologies and tools were devised in order to enhance the productivity.

Years ago we developed an approach that we have called SmartKey [1] to boost the operator's productivity. In the following sections we shall explain how key challenges in crowdsourcing are addressed thanks to this approach and its extensions, therefore we precede by a short description of the method.

In SmartKey we collect all character-images that the OCR engine has identified as a certain character with a confidence level that is below a certain threshold, and present them in Character Sessions, or carpets. For example, Fig 1 shows a carpet for the digit 3.

F. Daniel and F.M. Facca (Eds.): ICWE 2010 Workshops, LNCS 6385, pp. 408–411, 2010.

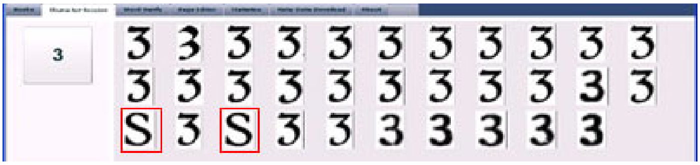

Fig. 1. Character Session for the digit "3"

The operator would reject the two S (we highlighted them in Fig. 1) that were misidentified as 3. Doing so all other characters are thus validated. For the sake of argument, suppose that the 32 characters in this session were collected from 32 different forms and in each form only this character was recognized with a low confidence. 30 Forms are thus validated with the handling of a single screen, versus the need to go over 30 screens in the naïve approach. This explains the huge productivity boost. As for the 2 rejected characters, they will move to a Word Session, where operators will see them in the context of the word to which they belong. (There is more in SmartKey, but this suffices for the follow-on discussion on crowdsourcing).

2 Virtual Service Delivery Centers

For many enterprises the document processing task is not a core business, so they may opt for outsourcing it to services delivery centers, on-shore or off-shore, provided the later can do the job with the required quality level. We have advocated [2] that the role of business services delivery centers will change, and they will basically turn into trusted Service Exchanges between enterprise-consumers and providers. The emerging Virtual Service Delivery Center, or VSDC for short, will be the broker that ties consumers to a providers pool, namely the crowd.

The consumers submit jobs that require both automatic and manual service. VSDC processes the job, performs as much as possible automatic processing, then segments the workload into subtasks suitable for the individual providers, matching tasks to skills. When the work is done, VSDC assembles and qualifies the completed subtasks and returns to requester, as well as handling the accounting.

We emphasize that VSDCs must establish "trust" in order to appeal to both consumers and providers. Consumers will look for guaranteed availability, quality, anonymity, privacy, and security. The small and medium businesses will also look for low entry barrier, and obviously every consumer is interested in low cost and fast setup. The providers would go for a trusted brokerage if they are guaranteed to be paid for their services, and would like the convenience of working over the internet (no commuting, flexible working time).

An efficient operation of VSDC relies on sophisticated technologies that introduce as much automation as possible into the process, and make a smart use of the manual work of the crowd. In the following sections we demonstrate how such technologies are applied to document processing, implementing the components of trust mentioned above.

3 Working with and Monitoring of the Crowd

We enlist key challenges in working with the crowd who provides the services, and show how they are addressed in our approach.

On-line quality control is maintained by introducing pseudo random errors in the Character Sessions (and other sessions) and measuring the percentage p_i of the missed characters by employee number i. This figure can be used for *fair rewarding* of the employees, with the compensation being a monotonically increasing function of $N_i*(1-p_i)$, where N_i is the amount of work done (e.g., number of screens validated by employee i). Further, *incentives plans* can be derived from this figure, and since it is being computed on-line it can also help urging an employee to take a break when the measured performance seems to drop due to fatigue.

A similar mechanism can be used for *training,* where screens that had already been processed are presented to new employees, along with instructions. While the employee is practicing, his or her performance is monitored, and feedback can be applied to maintain progress.

Privacy, or securing the *anonymity* of the individuals who filled the forms, is a key issue, which often limits the use of outsourcing, let alone crowdsourcing, when you do not know who will see the form. Clearly Character Sessions can be seen by anyone, since the operator has no idea if a digit comes from a social security number, date, $ amount, or some other field in a form.

For the more advanced sessions, rules can be set concerning the *authorization* of employees to see various pieces of information, or perform operations on them. For example, words can be seen and corrected by people who have been passed some screening procedures; At the other extreme, resolution of addresses versus some ID numbers, which call for examining a large portion of the form, but would rarely be needed (as most of the forms are corrected at lower level stages) would be done in-house (i.e., not outsourced).

4 Satisfying the Enterprise's Quality Requirements

The service delivery center, which handles the document processing for its enterprise customers, usually signs a service level agreement (SLA) with its customer. The keep up with the SLA the delivery center should assure the job quality, and its timely availability.

Crowdsourcing poses a challenge for both requirements, as the delivery center is not at full control of the crowd. People would join crowdsourcing operations for the convenience of flexible work, so they do not commit their work time, and they are an anonymous group with initially unknown skill level and performance.

The way we decompose the jobs in the document processing practice, most of the workforce is occupied at the low level correction and validation stages. Hence people can be pulled into the working force in a relative short time, after fast training sessions. In such a way we can keep a large pool, mitigating workforce availability issues.

For *quality assurance* we take advantage of the on-line monitoring of the crowd. Suppose the customer demands a quality level 1-p, meaning that, at most, a small fraction p of characters remains erroneous. It might well be that for each i, $p_i > p$, so no single employee can be assigned to this job. However, we can send the job to say employees number j and k, provided $p_{j*} p_k < p$. (We have made the plausible assumption that their errors are independent). Of course this technique can be generalized to the case where more than 2 providers are needed to get the expected error rate below the required threshold p. In other words, depending on the employee skill level, the same task may be performed either by one, two or three operators.

Another key advantage of our job decomposition methodology is that workloads are assigned to employees based on their skills. For example, in the processing of medical claims, the simple digit recognition will be done by basic level employees, while the validation of names of diseases will be done with people who have some medical background (either before their started to work, or via knowledge acquired by training).

5 Cooperative Correction System

COoperative eNgine for Correction of ExtRacted Text (CONCERT)[1] is a system that we develop for crowdsourcing workloads in book digitization, as part of the IMPACT project, one under the European 7th Program [3]. Some of the tools described above have been implemented, and the system is being used in a pilot. We shall describe the system in the Crowdsourcing Workshop.

References

1. US patent 5,455,875: System and method for correction of optical character recognition with display of image segments according to character data
2. Karnin, E., Walach, E.: Virtual Service Delivery Centers. Presented in Frontiers in Service 2007 conference (2007)
3. IMPACT Project, `http://www.impact-project.eu`

[1] The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-201 under grant agreement 215064.