Berlin : Springer, 2010. S. 421 - 424. ISBN 978-3-642-16984-7.

Quality Assurance for Human-based Electronic Services: A Decision Matrix for Choosing the Right Approach

Robert Kern, Hans Thies, Cordula Bauer, Gerhard Satzger
Karlsruhe Institute of Technology (KIT)

Abstract

Crowdsourcing in the form of human-based electronic services provides a powerful way of outsourcing so called human intelligence tasks (HITs) to a large workforce of people over the Internet. Because of the limited control over that workforce, it is challenging to ensure the quality of the work results. Several approaches have been proposed that can be applied to specific types of HITs. However, it is difficult to identify a suitable quality management approach for any given type of HIT. This paper aims to provide a first sketch of a decision matrix.

Keywords: Crowdsourcing, quality control, human computation

1 Introduction

The idea of human-based electronic services is that they look like Web services but they are not performed by a computer, instead they use human workforce out of a crowd of Internet users. The success of Amazon's Mechanical Turk (www.mturk.com) platform and the growing number of companies that build their business model entirely on that platform demonstrate the potential of this approach. The MTurk platform acts as a broker between requesters who publish human intelligence tasks (HITs) and workers who perform those tasks in return for a small amount of money. Kern et al. proposed the term *people services* (pServices) for this type of human-based electronic services and define it as *"Web-based software services that deliver human intelligence, perception, or action to customers as massively scalable resources"* [3]. As there is limited control over the individual contributors, particular attention has to be paid to the quality of the work results. Several quality assurance (QA) strategies for pServices have been proposed [6, 5, 1, 2]. However, it is not obvious, which approach can be and should be applied to which type of pServices. The aim of this paper is to provide a sketch of a decision matrix as a basis for discussion.

2 Related Work

Sorokin and Forsyth distinguish between two strategies, that leverage the crowd for QA of pServices [6]: the collection of multiple results and the performing of a separate review task (they call it grading task). In order to establish a common terminology we use the term **majority vote** for those approaches that *introduce redundancy by passing the same task to multiple workers and aggregating the results in order to compute the result with the highest probability for correctness* while we propose the term **review** for such approaches that *leverage individuals of the crowd for reviewing (e.g. validating) the results delivered by others* [2]. Majority vote mechanisms are already widely used in pServices scenarios today, e.g. on the MTurk platform. Initial research has shown that an amazing level of result quality can be achieved for basic tasks like natural language annotation, image labeling and data labeling: Sorokin and Forsyth identify objects on images by combining the drawings of silhouettes of distinct persons [6]. Snow et al. have ten distinct workers rate natural language expressions and compare their aggregated results to gold-standard labels given by experts [5]. Barr et al. mention the application of review for quality control on MTurk [1].

3 Decision Criteria

This chapter motivates and proposes a set of decision criteria for the selection

Criterion	Characteristic
Determinacy of the task results	***Deterministic:*** There is a well defined optimal result for the task i.e. two workers who perfectly meet the task objectives will pass exactly the same result, or the results can be automatically transformed (normalized) into a well defined optimal result.
	Non-deterministic: There is no well defined optimal result for the task.
Execution versus validation effort	***Similar:*** Task can be executed as quickly as its result can be validated.
	High: Execution effort for the task clearly exceeds validation effort.
Required level of quality	***Low:*** A single qualified worker is able to meet the quality needs.
	High: Quality needs exceed the capabilities of a single worker. The "wisdom of the crowd" must be leveraged.
Number of equivalent quality relevant entities	***Low:*** Result is only made of one or few equivalent quality relevant entities.
	High: Result is made of many equivalent quality relevant entities.

Figure 1: Decision criteria for selection of an adequate quality assurance mechanism for a given type of pService

In order to motivate the criteria we underline their relevance based on a series of examples:

3.1 Determinacy of the task results

For a *deterministic task* there is a well defined optimal result i.e. two workers who perfectly meet the task objectives will pass exactly the same result [2]. An example is the transcription of a speech recording if spelling and punctuation does not matter. In general, the majority vote (MV) approach can only be used for deterministic tasks. For non-deterministic tasks (e.g. creation of creative designs, text authoring and language translation), it won't work, because multiple results provided by different workers cannot be compared or aggregated in order to derive a single correct result. However, the *review* approach can obviously even be applied to non-deterministic results because a human reviewer is capable of dealing with variety.

3.2 Execution versus validation effort

Depending on the type of task, the effort for executing the task can be much *higher* than the effort for validating the result of the task. One example is text authoring. It is usually much simpler to decide, whether a given text meets certain quality criteria than to write the text. Another example is an image research tasks: it is much harder to find a picture which shows five bananas than to validate that there are indeed five bananas shown on the picture. For other tasks, the execution effort is *similar* to the validation effort. Examples are basic classification tasks and such research tasks for which a result validation is only possible by performing the research again.

3.3 Required level of quality

Depending on the required level of quality, the *wisdom of the crowd* [7] might be required in order to produce an acceptable result. Consider for example a one page language translation: If the quality needs are *low*, a single qualified translator might be able to deliver an acceptable result. However, if the requester is intolerant to mistakes (*high* quality needs) it is very hard for a single worker to meet the quality needs.

3.4 Number of equivalent quality relevant entities

If a result consists of many equivalent quality relevant entities, the result quality will be proportional to the percentage of those entities that meet the quality objectives. For example, a speech transcript will be the better the more words have been transcribed correctly. On the other hand, a painting consists of several graphical elements but its quality does not simply scale with the number of "good" elements.

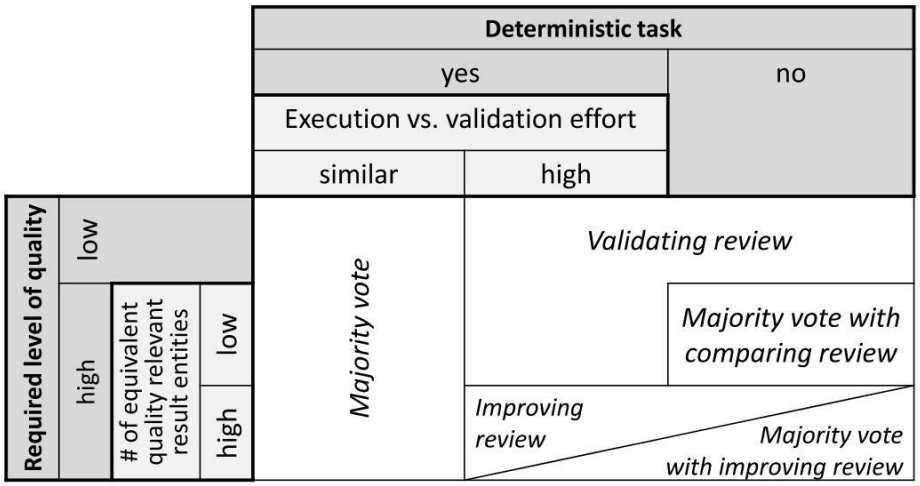

Figure 2: Decision matrix for choosing adequate quality assurance approaches for pServices

4 Decision Matrix

The proposed decision matrix links the decision criteria to a set of five quality QA for pServices which are variations of the simple MV and review approaches:
Majority vote (MV): The basic majority vote mechanism as described above.
Validation review (VR): Simple review mechanism for which a reviewer or a group of reviewers provides a binary rating whether to accept or reject a result. VR does not improve the quality of a result but only acts as a filter which can be passed by a result or not. When combining it with a feedback loop, VR can help to raise the skill level of workers.
Majority vote with comparing review (MVCR): An extension of MV in which a reviewer compares results delivered by multiple workers and groups them into sets of equivalent results. The reviewer basically performs the aggregation which is done automatically for deterministic tasks in case of MV.
Improving review (IR): A review approach in which the reviewer not only rates the result delivered by another worker, but spends additional effort on improving the result: A chain of reviewers improve the result step by step until it meets the requirements of the requester. A similar approach is used by CastingWords (www.castingwords.com) for speech transcription.
Majority vote with improving review (MVIR): A combination of MV and IR where multiple workers are delivering results which are then improved by a chain of multiple workers, each seeing the results delivered by the previous ones. The approach is a pService implementation of a Delphi study which is known to have the potential to deliver high quality forecasting results [4].

At the top of the decision matrix (Figure 2) we differentiate between deterministic and non-deterministic tasks. MV can only be applied to deterministic

tasks but depending on the execution effort for the task, MV might even not be recommended for those. For low quality needs, the basic recommendation is to use MV if the execution effort similar to the validation effort and to use the VR approach in all other cases. If the quality needs are high and the task is deterministic, the MV approach can be used and is recommended as long as the execution effort is not too high. If it is high, the IR approach is recommended which can even be applied to non-deterministic tasks. For those, also the MVIR approach can be used which can be assumed to achieve an even higher quality, because it massively increases the interaction between the workers.

5 Conclusion and Future Work

We have provided an initial sketch for a decision matrix which aims to identify an adequate quality assurance approach for a given type of human intelligence task (HIT). The matrix is based on four simple criteria that can be easily determined based on the characteristic of the HIT type, effort estimations as well as the quality needs. The decision matrix maps those criteria to a set of five general quality assurance mechanisms. In a next step, we plan to evaluate and concretize our concept based on a detailed analysis and comparison of the proposed mechanisms.

References

[1] Barr, J., Cabrera, L.F.: AI gets a brain. ACM Queue 4(4), 24–29 (May 2006)

[2] Kern, R., Bauer, C., Thies, H., Satzger, G.: Validating results of human-based electronic services leveraging multiple reviewers. In: Proceedings of the 16th Americas Conference on Information Systems (AMCIS). Lima, Peru (2010), (forthcoming)

[3] Kern, R., Zirpins, C., Agarwal, S.: Managing quality of Human-Based eServices. In: Feuerlicht, G., Lamersdorf, W. (eds.) Service-Oriented Computing - ICSOC 2008 Workshops, ICSOC 2008 International Workshops, Sydney, Australia, December 1st, 2008, Revised Selected Papers. Lecture Notes in Computer Science, vol. LNCS 5472, pp. 304–309. Springer (2009)

[4] Rowe, G., Wright, G.: The delphi technique as a forecasting tool: issues and analysis. International journal of forecasting 15(4), 353–375 (1999)

[5] Snow, R., OConnor, B., Jurafsky, D., Ng, A.Y.: Cheap and fastbut is it good? evaluating non-expert annotations for natural language tasks. In: EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 254–263. ACL, Stroudsburg, USA (2008)

[6] Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. In: CVPRW '08: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8. IEEE Computer Society, Washington, WA, USA (Jun 2008)

[7] Surowiecki, J.: The Wisdom of Crowds. Doubleday, New York, NY, USA, 1st edition edn. (2004)