Efficient Ground Truth Generation with Crowdsourcing Competitions

STSM Scientific Report – COST-STSM-IC1302-33405

STSM Applicant
Markus Rokicki, Leibniz Universität Hannover, Hannover (DE), rokicki@l3s.de

Host:
Dr. Sergej Zerr, University of Southampton, Southampton S017 1BJ (UK),
s.zerr@soton.ac.uk

Period: 2016-03-20 to 2016-04-02

Purpose of the STSM

Evaluation of information retrieval methods requires reliable and extensive ground truths. However, they require human input and therefore are costly to acquire. This is a serious limitation, especially for emerging search problems and applications. Crowdsourcing offers efficient and established means to obtain human input. Our previous joint work on crowdsourcing competitions specifically has been successful in eliciting human input at scale [1,2] and to variably modulate throughput in crowdsourcing competitions for streaming scenarios [3]. However, it remains to be verified whether it will be applicable to the information retrieval domain in practice.

The goal of the proposed short term scientific mission was to prepare an extensive experimental evaluation of the ability of the crowd to provide real valued judgments in our framework.

Work Carried out during STSM

During the visit, we worked on the experimental design for our study. We considered a range of suitable data sets to include in the study and decided on a) tasks requiring truly real valued judgments and b) a difficult task aiming at evaluating crowd ability, leading to two experimental evaluations. Both directions are designed to provide insights and methods on a more general level to inform more traditional relevance judgment designs as a next step.

For the first direction, our experimental design presents the workers with regression type tasks and expects real valued input. For the feedback design this leads to a generalized problem of taking degree of correctness (individual judgment quality) into account. To this end, we considered solutions that award scores based on absolute error, as well as rank based feedback where judgments are ranked according to their quality – rewarding only the most useful judgments.

For the second direction, users are presented with difficult tasks, such as computation tasks, aimed at evaluating crowd performance. The aim of these experiments is to design and evaluate collaboration and evaluation methods to improve annotation quality. Our experimental design will evaluate various forms and levels of collaboration among workers in groups of different sizes, combined with aggregation methods ranging from simple averaging to machine learning algorithms.

Main Results

To summarize, as a first step we want pursue two directions that address two aspects that are specific to relevance judgments:

1. real valued (and ordinal) input
2. challenging or ambiguous tasks

The target of the first experimental evaluation is CIKM 2016. Design of the second study is in a preliminary state.

Further Collaboration

We will further collaborate on preparing and carrying out the studies designed during the visit. The future work and collaboration will first follow the two directions described above: the first direction will focus on system design for obtaining real valued crowd input; the second direction covers collaboration and input aggregation.

References

[1] Markus Rokicki, Sergiu Chelaru, Sergej Zerr, Stefan Siersdorfer. *Competitive Game Designs for Improving the Cost Effectiveness of Crowdsourcing*. 23rd ACM Conference on Information and Knowledge Management (CIKM), Shanghai, China, 2014

[2] Markus Rokicki, Sergej Zerr, Stefan Siersdorfer. *Groupsourcing: Team Competition Designs for Improving the Cost Effectiveness of Crowdsourcing*. 24th International Conference on World Wide Web, WWW 2015, Florence, Italy

[3] Markus Rokicki, Sergej Zerr, Stefan Siersdorfer. *Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions.* 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada