Practice of Efficient Data Collection via Crowdsourcing: Aggregation, Incremental Relabelling, and Pricing

Alexey Drutsa, Valentina Fedorova, Dmitry Ustalov, Olga Megorskaya, Evfrosiniya Zerminova, Daria Baidakova
Part IX

Discussion of the projects’ results

Conclusion

Alexey Drutsa,
Head of Head of Efficiency and Growth Division, Yandex

Yandex.Toloka is a service of Swiss company Yandex Services AG
Tutorial outline

Part I: 40 min
Main Components

Coffee break: 30 min

Part II: 25 min
Brainstorming pipeline

Part III: 10 min
Introduction to Crowd Platform

Part IV: 85 min
Set & Run Projects

Lunch break: 90 min

Part V: 35 min
Interface & Quality control

Part VI: 60 min
Set & Run Projects cont.

Coffee break: 30 min

Part VII: 20 min
Incremental relabeling and pricing

Part VIII: 10 min
Results & Conclusions
Reminder: we implemented this pipeline

Tasks for performers:

Project #1: Does a photo contain a specific item?  
- If yes: photos with items
- If no: photos with incorrect similar items

Project #2: Find a similar item in the online store
- If no: photos with incorrect similar items
- If yes: photos with correct similar items

Project #3: Is the found item (project 2) similar to the initial one? (post-verification)
- If no: photos with incorrect similar items
- If yes: photos with correct similar items

Project #4: Which of the found items is more similar to the initial one?
- OR: photos w/ best substitute

If yes: Accept the result in Project #2 Pay for it
If no: Reject the result in Project #2 Do not pay for it
Project #1: Filter out photos without objects

**Task**
› Does a photo contain an item of desired type?

**Our results**
› Amount: 30 photos
› Overlap: 3
› Time: 5 min
› Cost: $0.09 + Toloka fee
Project #2: Searching for similar items on the online store

Task
› Find a similar item on the internet

Our results
› Amount: 25 photos
› Overlap: 3
› Time: 25 min
› Cost: $1.74 + Toloka fee
Project #3: Accept correctness of items found

**Task**
- Is the found item (project 2) similar to the initial one?

**Our results**
- Amount: 75 photos
- Overlap: 3
- Time: 3 min
- Cost: $0.20 + Toloka fee
Project #4: Decide which substitute works best

Task
› Which of the items is similar to the initial one?

Our results
› Amount: 62 photos
› Overlap: 3
› Time: 10 min
› Cost: $0.10 + Toloka fee
Statistics over the whole pipeline

- 30 photos processed to find the substitute items and evaluate their similarity
- within 45 min on real performers
- total cost: $2.15 + Toloka fee
Afterparty: upgrade your pipeline

To obtain more comprehensive data
› Use more item types at the same time

To reduce costs
› Use incremental relabeling aka Dynamic overlap

To improve quality
› Use dynamic pricing
› Add more Golden Sets and hints
› Experiment with aggregation methods
› Add training for performers
API of Yandex.Toloka

- Allows you to automate all steps of our pipeline

Discover at: https://yandex.ru/dev/toloka/
Crowdsource all types of data

Search Relevance

Generation of content

Moderation

Speech Technologies

Computer vision
References: Aggregation

[5] Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y, Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks, EMNLP 2008
References: Aggregation

[15] Chen, X. and Bennett, P. N and Collins-Thompson, K. and Horvitz, E., Pairwise ranking aggregation in a crowdsourced setting, WSDM 2013
References: Incremental relabeling & Pricing

[20] Lin, C. H, Mausam, M., Weld, D. S, To Re(label), or Not To Re(label), HCOMP 2014

References: Tutorials

[27] Crowdsourcing: Beyond Label Generation, Vaughan, J. W. KDD 2017
[33] Truth Discovery and Crowdsourcing Aggregation: A Unified Perspective, Gao, J., Li, Q., Zhao, B., Fan, W., Han, J., VLDB 2015
[34] Data-Driven Crowdsourcing: Management, Mining, and Applications, Chen, L., Lee, D., Milo, T., ICDE 15
[35] Practice of Efficient Data Collection via Crowdsourcing at Large-Scale, Drutsa A., Fedorova V., Megorskaya O., Zerminova E., KDD 2019
Thank you!
Questions?

Alexey Drutsa
Head of Head of Efficiency and Growth Division

adrutsa@yandex-team.ru
https://research.yandex.com/tutorials/crowd/wsdm-2020