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

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

Olga Megorskaya,
Head of Crowdsourcing Department, Yandex

Yandex.Toloka is a service of Swiss company Yandex Services AG
Facebook chat for materials
wsdm2020crowd
Crowdsourcing

Specific way to design a business process

A big task → Cloud of performers → Result
## Crowdsourcing applications: examples

<table>
<thead>
<tr>
<th>Task type</th>
<th>Where is used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information assessment</td>
<td>Ranking of search results</td>
</tr>
<tr>
<td>Content categorization</td>
<td>Text and media moderation, data cleaning and filtering</td>
</tr>
<tr>
<td>Content annotation</td>
<td>Metadata tagging</td>
</tr>
<tr>
<td>Pairwise comparison</td>
<td>Offline evaluation, media duplication check</td>
</tr>
<tr>
<td>Object segmentation, including 3D</td>
<td>Image recognition for self-driving car</td>
</tr>
<tr>
<td>Audio and video transcription</td>
<td>Speech recognition for voice-controlled virtual assistant</td>
</tr>
<tr>
<td>Spatial crowdsourcing</td>
<td>Verify business information and office hours</td>
</tr>
</tbody>
</table>
Example: binary classification

Is this cat white?

Yes

No
Example: multi classification

"Real French restaurant"

If you are a gourmand, I can recommend you the "Real French restaurant", located in the historic cellar, with elements of antique design and quite interesting cuisine. The restaurant is small, but very cozy and romantic. The restaurant is very suitable for romance and even for business meetings.

Is it a feedback?

- Yes, it is
- No, it's other comment

- Personal information
- Swearing, vulgarity, insults, aggressive statements
- Spam, advertising
Example: multi classification with ordered labels
Examples: pairwise comparison

How to Make Perfect Pancakes

Food Network Magazine shows you how to make the best short stack, plus some tasty toppings.

April 24, 2015

From: Food Network Magazine

Query: how to make pancakes
Which one do you like better?

- Left
- Right

Please, comment your choice

Continue
Examples: transcription with textual answers

1. There is a speech on the record
2. No speech or inaudible

Annotation
4. The quick brown fox jumps over the lazy dog
Examples: object segmentation
Examples: spatial crowdsourcing
A crowdsourcing platform: two-sided market

Performers → Requesters

Performers

Requesters

Icons: shopping cart, building, globe, factory, government building
Crowdsourcing platforms: examples

› Amazon Mechanical Turk
› Yandex.Toloka
› Microworkers.com
› Gigwalk
› ClickWorker
› CloudFactory
› Figure Eight
› CrowdSource
› DefinedCrowd
› …
Pros of crowdsourcing platforms

- 24/7
- Variety of skilled performers
- Vast region coverage
- Ongoing processes
Crowdsourcing growth: Yandex experience

Active performers in Yandex.Toloka

- 2014: 9K
- 2015: 120K
- 2016: 270K
- 2017: 570K
- 2018: 1.1M
- 2019: 2.2M
Crowdsourcing growth: Yandex experience

Different projects in Yandex.Toloka
Everyday on Yandex.Toloka

500+ different projects  36K+ performers  12M+ tasks
<table>
<thead>
<tr>
<th>Task Description</th>
<th>Number of Items</th>
<th>Time to Complete</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Side-by-side object comparison</td>
<td>1,000 tasks</td>
<td>Done in 10 min</td>
<td>$2.4</td>
</tr>
<tr>
<td>Object classification</td>
<td>1,000 photos</td>
<td>Done in 15 min</td>
<td>$1.2</td>
</tr>
<tr>
<td>Object segmentation</td>
<td>about 1,000 objects in 100 photos</td>
<td>Done in 6 h</td>
<td>$3.6</td>
</tr>
<tr>
<td>Phrase generation for a chatbot</td>
<td>500 phrases</td>
<td>Done in 15 min</td>
<td>$1</td>
</tr>
<tr>
<td>Audio transcription</td>
<td>100 recordings 25 minute long</td>
<td>Done in 20 min</td>
<td>$6</td>
</tr>
<tr>
<td>Video rating</td>
<td>10,000 videos</td>
<td>Done in 2 h</td>
<td>$10</td>
</tr>
</tbody>
</table>
Tutorial overview
Why this tutorial?

Practice
Part I: 30 min

Main components of data collection via crowdsourcing

› Decomposition for effective pipeline
› Task instruction & interface: best practices
› Quality control techniques

Olga Megorskaya
Head of Crowdsourcing Department, Yandex
Part II: 25 min

Analysis of label collection projects to be done (practice session)

› Dataset and required labels
› Discussion: how to collect labels?
› Data labelling pipeline for implementation

Daria Baidakova,
Project Manager,
Crowdsourcing Department, Yandex
Part III: 10 min

Introduction to the crowdsourcing platform Yandex.Toloka for requesters

› Main types of instances
› Project: creation & configuration
› Pool: creation & configuration
› Tasks: uploading & golden set creation
› Statistics in flight and download of results

Evfrosiniya Zerminova
Head of Data Analysis and Research Group
Crowdsourcing Department, Yandex
Part IV: 60 min

Setting up and running label collection projects (practice session)

You
› create
› configure
› run on real performers

Data labelling projects in real-time

Daria Baidakova,
Project Manager,
Crowdsourcing Department, Yandex
Part V: 35 min

Interface & quality control

› Detailed examination of quality control techniques
› Comprehensive overview of best practices for creating a functional interface

Alexey Drutsa
Head of Efficiency and Growth Division
Crowdsourcing Department, Yandex
Part VI: 25 min

Theory on Aggregation

› Multiclass labels
› Pairwise comparisons

Valentina Fedorova
Researcher,
Research Department, Yandex
Part VII: 90 min

Setting up and running label collection projects cont. (practice session)

You
› create
› configure
› run on real performers

Data labelling projects in real-time

Daria Baidakova,
Project Manager,
Crowdsourcing Department, Yandex
Part VIII: 20 min

Theory on efficient incremental relabelling and pricing

› Incremental relabelling
› Performance-based pricing

Valentina Fedorova
Researcher,
Research Department, Yandex
Part IX: 10 min

Discussion of results from the projects & conclusions

› Results of your projects
› Extensions to work on after tutorial

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

Lunch break: 90 min

Part IV: 85 min  
Set & Run Projects

Part V: 35 min  
Interface & Quality control

Part VI: 25 min  
Theory on Aggregation

Coffee break: 30 min

Part VI: 60 min  
Set & Run Projects cont.

Part VII: 20 min  
Incremental relabeling and pricing

Part VIII: 10 min  
Results & Conclusions
Thank you!
Questions?

Olga Megorskaya
Head of Crowdsourcing Department

omegorskaya@yandex-team.ru

https://research.yandex.com/tutorials/crowd/wsdm-2020