Russian Information Retrieval Evaluation Seminar (ROMIP)


Igor Nekrestyanov, Pavel Braslavski

CLEF 2010
ROMIP at a glance

- TREC-like Russian initiative
- Started 2002
- Several text and image collections
- 10-15 participants per year (total 50+)
  - Academia and industry, students support
- \( \sim 3\ 000 \) man-hours of evaluation (2009)
- Remote participation + live meeting
- Collections are freely available
- Popular testbed for IR research in Russia
- Related activities: summer school in IR
Why?

• Russia specifics
  • Strong IR industry
  • Limited research in academia
  • Participation in global events considered complicated for Russian groups (language barrier, costs, etc.)
  • Russian language was not covered in international campaigns

• Objectives
  • Consolidate IR community
  • Stimulate research in the area
  • Independent evaluation
Evaluation methodology

- Similar to TREC approaches
- What’s special?
  - Russian language collections
  - Some tasks are unique
    - E.g. news clustering, snippet generation, etc.
  - Mix of widely used and custom metrics
    - E.g. snippet informativeness/readability
  - Typically 2+ assessors (agreement 80-85%)
  - Domain experts for legal-related tracks
  - Rules and methodology are adjusted yearly
## Largest text collections

<table>
<thead>
<tr>
<th>Collection</th>
<th>Documents</th>
<th>Size (compressed)</th>
<th>Topics</th>
<th>Evaluated within ad-hoc search track</th>
</tr>
</thead>
<tbody>
<tr>
<td>Legal</td>
<td>~300 000</td>
<td>2 Gb</td>
<td>14 794</td>
<td>220</td>
</tr>
<tr>
<td>By.Web</td>
<td>1 524 676</td>
<td>8 Gb</td>
<td>~ 60 000</td>
<td>1 500+</td>
</tr>
<tr>
<td>KM.RU</td>
<td>3 010 455</td>
<td>13 Gb</td>
<td>~ 60 000</td>
<td>~250</td>
</tr>
</tbody>
</table>

21.09.2010
Text documents tracks

- Classic tracks run for years
  - Ad-hoc text retrieval
  - Text categorization (Web pages & sites, legal)
- Experimental tracks every year
  - Snippet generation
  - QA and fact extraction
  - News clustering
  - Search by sample document
Snippets evaluation
Image collections

- Photo collection: 20 000 images from Flickr
- Dups collection: 15 hrs video → 37 800 frames
Image tracks

- Content based image retrieval (started 2008)
  - 750 tasks labeled
- Near-duplicate detection (started 2008)
  - ~1500 clusters
- Image annotation (started 2010)
  - ~ 1000 labeled images
ROMIP timeline

- Systems applied
- Systems participated
- # of tracks

- News snippets
- News+ ROMIP
- Image tracks
- 3000 man-hours eval.
- QA image tagging

- Search classification
- Legal
- Legal 2007 BY.Web KM.RU


21.09.2010
Thank you!
Questions?

Pavel Braslavski
pb@yandex-team.ru

Igor Nekrestyanov
romip@romip.ru
RuSSIR

Annual event
100+ participants
4th RuSSIR: Voronezh
13 – 18 September