Russian Information Retrieval Evaluation Seminar

Boris Dobrov[♣], Igor Kuralenok[♥], Natalia Loukachevitch[♣],
Igor Nekrestyanov[♥], Ilya Segalovich[♠]

[♣]Moscow State University, [♥]St.Petersburg State University, [♠]Yandex
Russia
romip@oasis.apmath.spbu.ru
http://romip.narod.ru

Abstract

This paper presents Russian information retrieval evaluation initiative and results obtained during first year. In particular, we describe first ROMIP seminar, used Cyrillic Web collection and search tasks as well as ongoing efforts on ROMIP'2004.

1. Introduction

Russian information retrieval evaluation initiative was launched in 2002 with purpose to increase communication and support community of researchers (from both academia and industry) in the area of text retrieval for Russian language collections by providing infrastructure necessary for evaluation of information retrieval methodologies.

In particular, series of Russian Information Evaluation Retrieval Seminars (ROMIP seminars) is planned to be held yearly. The first seminar was organized in 2003 with the final workshop attached to the Russian Conference on Digital Libraries (St. Petersburg, October 2003).

In many respects ROMIP seminars are similar to other world information retrieval events such as TREC, CLEF, NTCIR, etc. Initiation of the new one was motivated by several reasons:

- *absence of publicly available Russian test collections.*
  To the best of our knowledge ROMIP'2003 collection is the first publicly available large-scale Russian text collection for evaluation of information retrieval methods;
- *relatively low interest for the creation of Russian language tracks/collections within the framework* of the existing evaluation initiatives
  As far as we know only CLEF'2003 had Russian document collection but it was rather small (37 Mb of about 20,000 stories from Izvestia newspaper in 1995);
- *low rate of participation of Russian research groups in the existing evaluation initiatives.*
  Some of ROMIP'2003 participants have a wide experience in IR research and applications but this was their first experience with a public independent evaluation forum.

Similar to TREC ROMIP has cycle nature and is overseen by a program committee consisting of representatives from academia and industry. Given collection and tasks participants run their own system on the data and submit results to the organizing committee. Collected results are independently judged and the cycle ends with a workshop for sharing experience and discussing future plans.

However, we did not precisely copy TREC tasks and methodology. Indeed we adapt it to our circumstances and combined them with other recent approaches in the information retrieval evaluation domain.

In the rest of the paper we describe ROMIP'2003 collection, tracks, participants and evaluation methodology. Due to size limitations we only briefly outline most of things trying to highlight ROMIP specifics. Interested readers may consult full ROMIP proceedings (available at romip.narod.ru) for details.

2. Collection

Construction of large test collections possess number of problems to be solved (Cormack et all., 1998). For ROMIP'2003 we decided to concentrate on the Web collection. This was motivated by interest of participants and relative simplicity to obtain legal permissions to use data.

We used sites from the *Narod.Ru* domain as a source. Constructed collection consists of about 728,000 pages 7Gb in total (about 4.5 million unique words, over 130 millions document - unique word pairs).

Narod.ru is a popular Russian free web site hosting service similar to the *Yahoo! GeoCites*. It hosts wide variety of sites from small personal homepages and small companies cites to large 200M+ online newspapers. Diversity of content makes this collection to be a challenging ground for IR experiments.

Collection was formed as snapshot of random subset of over 22000 sites (about 20% of the whole domain). Only HTML files were taken.

Narod.Ru offers page templates for number of typical needs of the web site owner and we decided to omit sites extensively using these templates in order to better represent typical content of the Russian Web. Note, that no modifications to page content were performed.

We collected data by directly copying them from the web server harddrive (courtesy of the Yandex) instead of crawling them. Therefore for some web sites we have pages that are not accessible from outside by crawling. Moreover link structure of result dataset is rather sparse and does not represent link structure of the real Web.

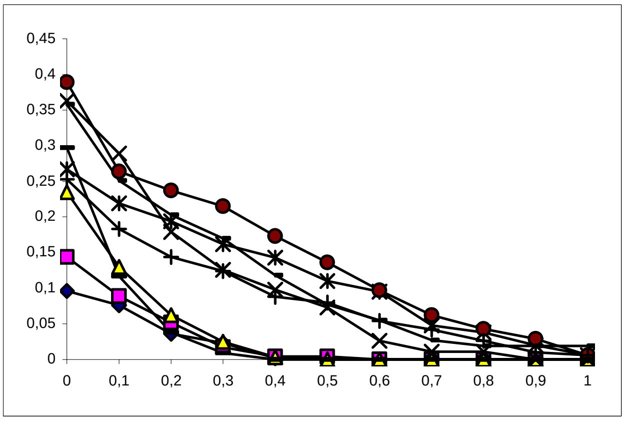

Figure 1. Sample results of the adhoc track (weak relevance)

3. Tracks

ROMIP'2003 had two tracks — "adhoc" retrieval from Web collection and Web-site classification. This selection was stipulated by interest of participants and availability if suitable text collections in given time constraints.

3.1. Adhoc retrieval

This track is similar to TREC adhoc retrieval track (1992-1998) and web retrieval (1998-2003). The main innovation here is the usage of large scale Russian-language test collection and tasks.

Queries for "adhoc" track were selected from the daily log of the popular Russian Web retrieval system Yandex (www.yandex.ru). We consider queries consisting of Russian words (at least two) without misspellings.

To prevent fine-tuning of results (which was not allowed in ROMIP'2003) participants were asked to perform 15000 queries and for each query submit the first 100 results to the organizing committee. Queries for evaluation (54) were selected by organisers after all the participants had submitted their results. At least 2 independent judges evaluate the relevance of the each answer document.

Average results were not very high (see figure 1 for example). We see three major reasons for this:

- The data was rather difficult due to very high diversity.
- "Broad" queries. Most of queries were rather short (3 words on average) and they allow multiple interpretations.
- "Narrow" judges. Judges used extended descriptions of queries (see section 5 for details) and while this potentially increased rate of the agreement between judges it may have caused rejection of some answers to initial "broad" question.

These points are indirectly supported by low level of agreement between judges (table 5).
Also, these results are in line with recent results of the topic distillation track of the TREC'2003 that also deals with "broad" queries.

3.2. Web site classification

To the best of our knowledge ROMIP'2003 Web site classification track is the first large-scale attempt to evaluate the effectiveness of web site classification algorithms in an independent way.

The training set for the classification track was based on the existing self-moderated Web catalogue for Narod.Ru sites. We selected about 170 categories from the second level of hierarchy. Each of the selected categories has at least 5 samples. Participants were asked to assign a list containing maximum 5 categories to each of 22000 web sites from the collection. At the evaluation stage all the assignments from 17 selected categories were judged by at least two judges.

To assign relevance score judge was expected to read all documents from this site even if they are not accessible from the main page via links and if any of viewed pages is relevant to the category the whole site assumed to be relevant.

Run	*Precision*	*Recall*	*F_1*
1	0.28	0.21	0.20
2	0.20	0.08	0.11
3	0.28	0.06	0.10
4	0.15	0.13	0.10
5	**0.38**	**0.55**	**0.42**

Table 1. Sample results of classification track (weak relevance).

Average results for this track are also relatively low (see table 1). In addition to possible explanations of low performance for adhoc track we must note that training set appears to be very noisy - about 33% of training set documents were judged not relevant during assessment. On the other hand this limitation came from the real-world and clean training sets likely to be typical for the Web domain.

4. Participants

This year we had nine applications for participation but only seven teams were able to complete tasks on schedule. Among them were several research teams from industry including two major players on the Russian web search market (see Table 2). Coordination and assessment of results was performed by research group from the St. Petersburg State University (ir.apmath.spbu.ru).

Participation from academia was lower probably because research prototypes were not ready for scale of considered tasks and deadlines were tight.

Participants were allowed to submit results from more than one run. In total 9 runs for adhoc track and 5 runs for classification track (Table 3).

Name	Type	Typical Software
Kodeks www.kodeks.ru	Business	Legal Search Engine 10^6 documents
Moscow Medical Academy (Key-To-Texts) www.mmascience.ru/ktt/	Univ.	Search Engine Tools
Rambler www.rambler.ru	Business	Web Search Engine 10^8 documents
Research Computing Center of Moscow State Univ. & Center for Information Research www.cir.ru	Univ. & NCO	Corporation Search Engine 10^6 documents
Russian Context Optimizer www.rco.ru	Business	Corporation Search Engine 10^6 documents
Velton Soft www.soft.velton.net.ua	Business	Search Engine Tools
Yandex www.yandex.ru	Business	Web Search Engine 10^8 documents

Table 2. Participants of ROMIP'2003.

5. Evaluation Methodology

For the evaluation of results we used the TREC-like pooling mechanism. For every retrieval task all results from all participants were collected to single pool and then were judged by assessors for relevance. In total 19 assessors were involved into ROMIP'2003.

This approach works well for TREC for many years and has many useful features – e.g. results are judged anonymously (i.e. judge can not favour particular system) and it is possible to approximate recall using answers found by other systems.

5.1. Reusability

One of the key features in Cranfield (Jones, 1981) evaluation methodology is reusability of constructed test corpus. Because ROMIP'2003 employed this type of evaluation our result corpus can be used outside of the conference scope.

The ROMIP'2003 collection and tasks for both tracks are publicly available from organizing committee on request.

5.2. Human relevance judgments

For ROMIP'2003 we used triple relevance judgements (*relevant/not-relevant/can-not-judge*). Introduction of third mark was motivated by purely technical reasons – some of web pages use malformed HTML that can not be visualized by our assessment tool.

It is widely accepted that notion of relevance is highly subjective (Voorhees, 2000). Therefore, set of answers accepted by one human judge may not include many of results that are good for another judge. To improve recall approximation and decrease the influence of subjectivity we used multiple (at least two) human relevance assessments.

Name	Ad-hoc Task	Classification Task
Kodeks	1	
Moscow Medical Academy		1
Rambler		2
RCC of MSU & NCO CIR	3	
RCO	1	1
Velton Soft	1	1
Yandex	3	

Table 3. Distribution of the submitted runs vs. tracks.

According to Mizzaro model of relevance (Mizzaro, 1998) judge starting from written retrieval problem reconstruct original information need. This process naturally results in the discrepancy and information needs to be reconstructed by different judges may significantly differ. This is especially noticeable if formulated information need (i.e. written retrieval problem) is "broad", e.g. because it is short. And it is well known that typical user queries are rather short.

To minimize this discrepancy we introduced the "extended" version of the search problem specification. Note that this specification differs from extended query in TREC. It is formulated only for assessors and supposed to be more narrow than the short search problem specification.

An extended version of the search problem includes the native language description of expected results and was prepared during the selection of queries to be evaluated. The purpose of extended description is to clarify the information need and minimize the number of possible interpretations by assessors.

Note, that this means ROMIP'2003 assessors judge the relevance not to given query but to particular information need that could cause it.

Still merging multiple judgements for the same document query pair is a problem – if judges have different opinions then it is unclear how to deduce final judgement.

For the official ROMIP'2003 evaluation we used 2 alternative ways to merge judgements from different judges – *weak* and *strong* agreements. In first case document considered to be relevant if any of judges said it is relevant. According to later approach document is relevant only if all assessors agree on that.

These two approaches provide us two extremes – strongly relevant documents are important to evaluate precision and weakly relevant document are important for estimating recall.

The following table presents the summary of judged ROMIP'2003 results. Comparison of numbers of results judged as weakly and strongly relevant shows how big is discrepancy between judges (for most of results 2 judges were involved).

	Weakly relevant	Strongly relevant	Total judged
Adhoc retrieval	1187	391	10084
Web site classification	906	338	3060

Table 5. Summary of the ROMIP'2003 judgments.

It is possible to apply different approaches to merge judgements (e.g. 'majority' rule). We varied set of assessors evaluating same answers (even for the same retrieval problem) and therefore it is possible to measure their agreement and even assign confidence weight to their scores. However, it is still an open question whether this will help to obtain better merged scores (in reliable way) and this is the topic of ongoing research.

5.3. Tools

To simplify assessment task we developed tool for collection of judgments. Same tool was used by all assessors for both tracks.

Development of such tool is mostly engineering task but there are few requirements affecting results of evaluation. In particular, we were collecting relevance judgments for HTML pages and our assessors were using different operating systems. Visual representation of ill-designed or malformed HTML documents in different browsers may significantly vary, thus introducing additional source of discrepancy. To avoid this we implemented cross-platform tool using java means to render HTML pages.

5.4. Metrics

As official measures or retrieval quality in ROMIP'2003 we mostly used widely known metrics (Rijsbergen,79) which are summarized in the following table:

Adhoc retrieval	Web site classification
• Precision • Precision at level • Average precision • Recall • 11pt TREC precision-recall graph	• Recall • Precision • F_1 (micro and macro averaged)

Table 6. Official measures of the ROMIP'2003.

Conclusion and future plans

The main achievement of ROMIP initiative in 2003 is the fact that first seminar was actually successfully held.

As the material outcome of the first year activity of ROMIP we have several useful resources:

- First large scale publicly available Cyrillic collection for evaluation of IR systems;
- First *reproducible* data on performance of several IR methods for Cyrillic that could be used as baseline for further research;
- Software for collection of relevance judgments for individual pages and Web-sites.

In future we plan to include new tracks into ROMIP program. This process is expected to be regulated by interest of participants.

In particular for ROMIP'2004 we plan to repeat two tracks from ROMIP'2003 using same collection but different tasks, in addition several new tracks are being introduced including question answering track and adhoc retrieval for non-Web collection.

ROMIP is an open initiative and we welcome participation of researchers interested in information retrieval for Russian language collections from all parts of the world.

Acknowledgements

We would like to thank members of ROMIP organizing committee and all others who helped to implement ROMIP'2003.

References

Cormack, G. V., Palmer, C. R. & Clarke, C. L. A. (1998). Efficient Construction of Large Test Collections. In *Proc. of the SIGIR'98* (pp. 282-289).

Rijsbergen, C.J. (1979). Information Retrieval, Butterworths, London

Ellen M. Voorhees (2000) Variations in relevance judgements and measurement of retrieval effectiveness. *Information Processing and Management*, 36:697-716

Harman, D. (2000). What we have learned, and not learned, from TREC. In *Proc. of the BCS IRSG'2000*, (pp. 2-20).

Jones, K. S. (1981). Information Retrieval Experiment. Butterworths, London.

Kuralenok I. & Nekrestyanov, I. (2002). Evaluation of Text Retrieval Systems. *Programming and Computing Software*, 28(4), (pp. 226-242).

Mizzaro, S (1998) How many relevancies in information retrieval? *Interacting with Computers*, 10(3), (pp. 303-320)

Voorhees, E. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proc. of the SIGIR'98* (pp. 315-323).

Wilbur, J. W. (1998). The knowledge in multiple human relevance judgments. *TOIS*, 16(2). (pp. 101-126).

Zobel, J. (1998). How reliable are large-scale information retrieval experiments? In *Proc. of the SIGIR'98,* (pp. 308-315).

TREC 2003 Web Distillation Results. In *Proc. of the TREC'2003*. http://trec.nist.gov/pubs/trec12/t12_proceedings.html