Академические инициативы Яндекса

Павел Браславский

Я

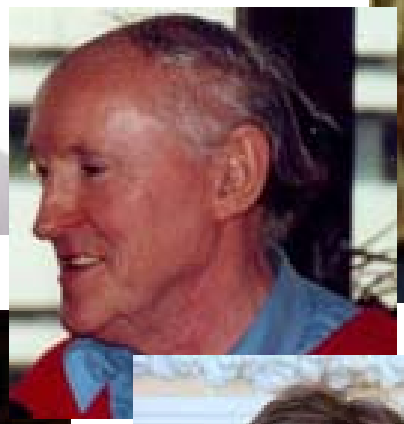
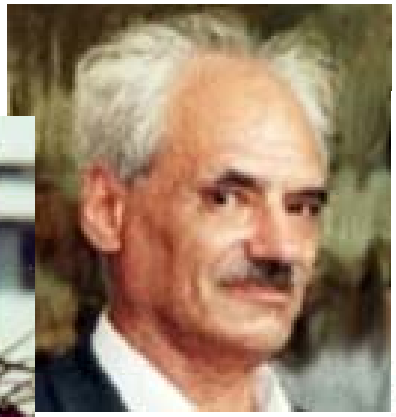
Академические инициативы

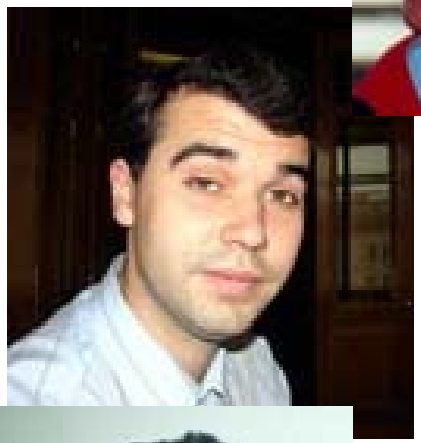
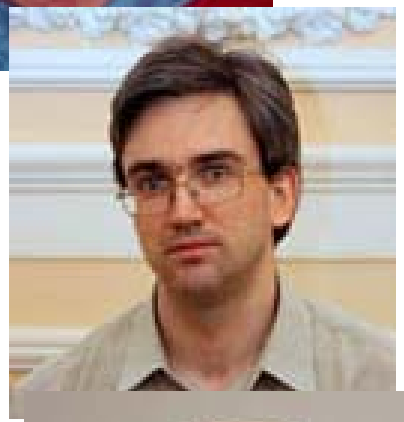
- Школа Анализа Данных
- Семинары Яндекса
- Интернет-математика
- РОМИП
- Школа по информационному поиску (RuSSIR)
- Книга «Введение в информационный поиск»

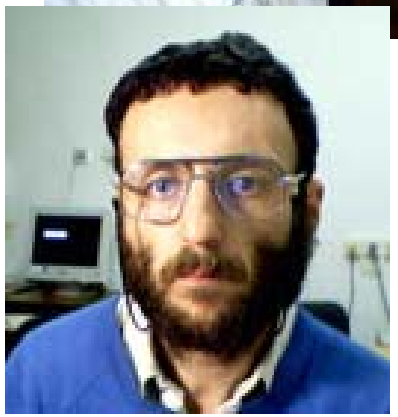
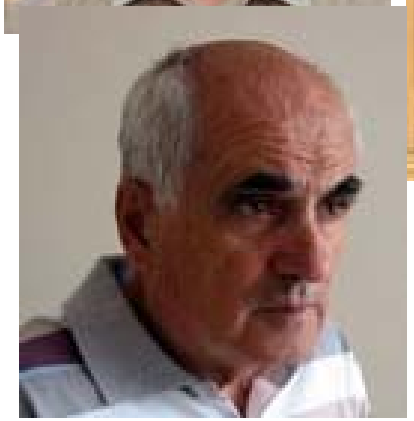
Yandex School of Data Analysis

two-year master program, http://shad.yandex.ru

Teachers

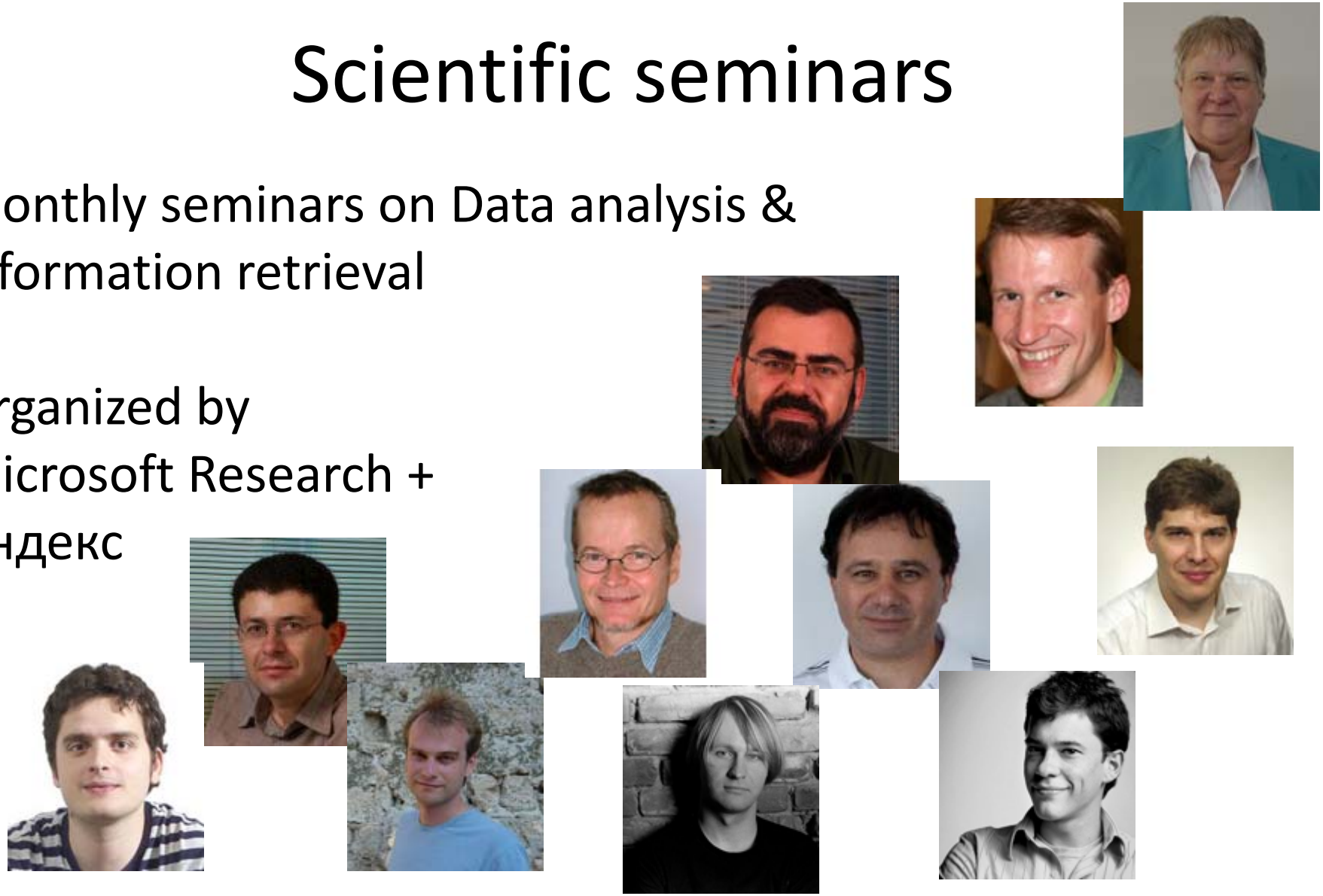




Scientific seminars

Monthly seminars on Data analysis & information retrieval

Organized by
Microsoft Research +
Яндекс



http://company.yandex.ru/public/seminars/schedule/

IMAT 2009

- Learning to rank
- 245 features for query-document pairs
- Graded relevance judgments (0..4)
- Pure numeric data (i.e. no original queries, documents or feature semantics)
- Learning set: 97 290 feature vectors (9 124 queries)
- Test set: 115 643 vectors (21 103 – public evaluation; 94 540 – final evaluation)
- Evaluation measure: DCG
- http://imat2009.yandex.ru

Рейтинг

Таблица объединяет финальный рейтинг конкурса (на 15.05.2009) и новые результаты. Подробнее о задаче и методике оценки с
о конкурсе и раздел Задачи и данные.

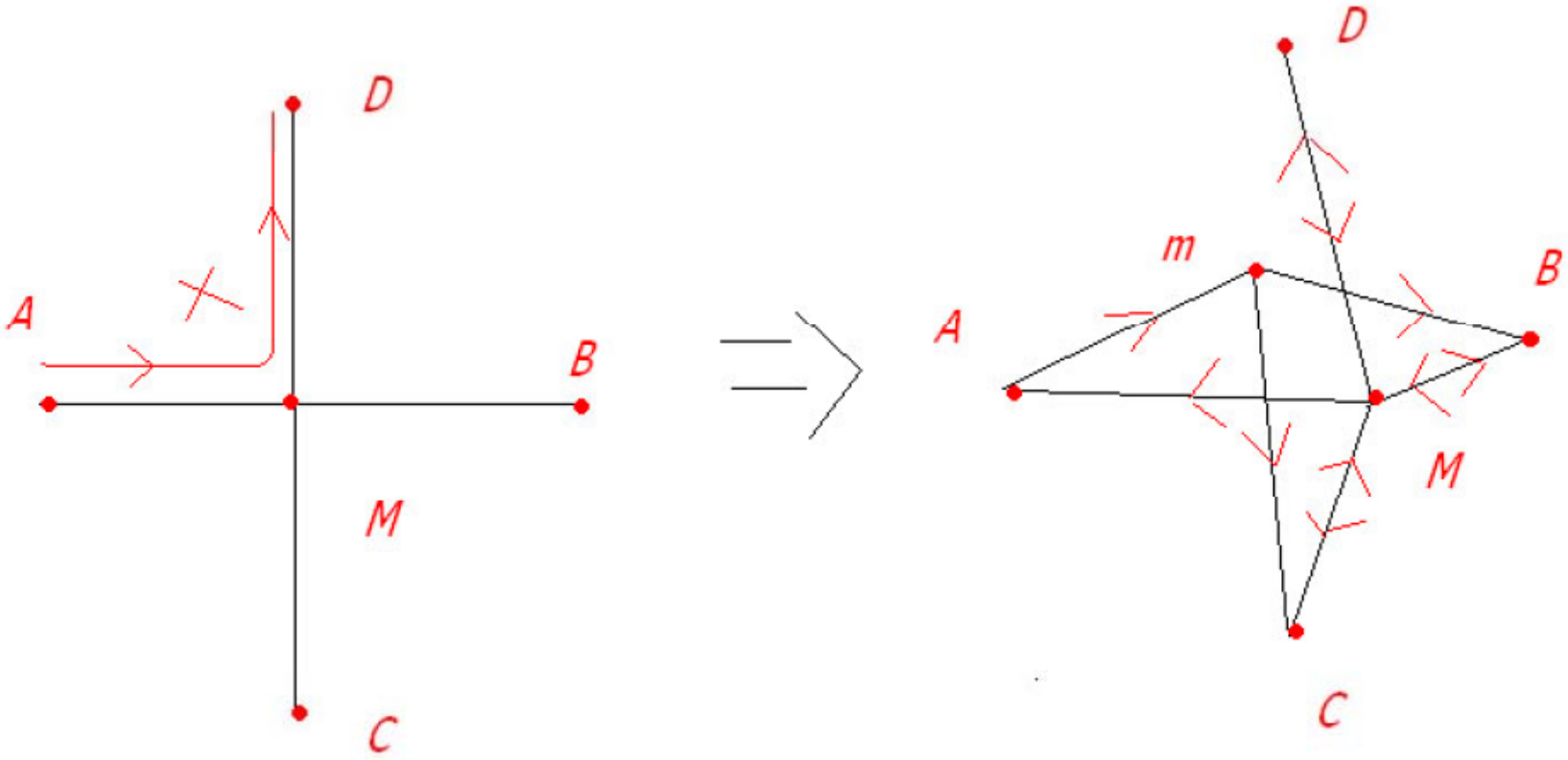
Команда	Время последней загрузки	Количество попыток	Последний результат (публичная оценка)	Финальный результат
alexeigor1	07.04.2010 (16:08 GMT+03)	1	4.284426	*4.156340*
Всем чмоке в этом чате :)	14.02.2010 (19:03 GMT+03)	4	4.283924	*4.133886*
Joker	05.09.2009 (05:07 GMT+03)	2	4.283317	*4.151528*
Сам	21.02.2010 (20:57 GMT+03)	187	4.282790	*4.147628*
-F	02.12.2009 (16:44 GMT+03)	2	4.281325	*4.145202*
alexeigor	07.05.2009 (17:02 GMT+03)	118	4.280676	*4.141230*
AG	03.05.2010 (02:13 GMT+03)	1	4.279520	*4.148574*
depechemode	29.10.2009 (01:19 GMT+03)	30	4.278378	*4.142855*
RelevanceDoesMatter	24.09.2010 (18:58 GMT+03)	448	4.276658	*4.135519*
Победа	17.03.2009 (16:25 GMT+03)	3	4.276001	*4.139854*
Simple	24.09.2010 (19:17 GMT+03)	581	4.275371	*4.132883*
ACGT	15.05.2009 (14:03 GMT+03)	21	4.274666	*4.128807*
stohastic	25.10.2009 (23:37 GMT+03)	819	4.274414	*4.129173*
Злобный Терминатор	03.09.2010 (15:52 GMT+03)	14	4.269719	*4.132461*
aport	13.09.2010 (10:45 GMT+03)	14	4.269432	*4.123704*
ПушистаяПуська	06.05.2010 (11:26 GMT+03)	1	4.268557	*4.128603*
WoodWeb	22.04.2009 (23:09 GMT+03)	12	4.267894	*4.127512*
Nordic	15.05.2009 (23:37 GMT+03)	4	4.266904	*3.857102*
stohastic	15.05.2009 (23:43 GMT+03)	176	4.266712	*4.118830*
Михаил	11.11.2010 (11:59 GMT+03)	2	4.266247	—
Euclid	27.02.2010 (05:40 GMT+03)	124	4.265264	*4.139924*
Test	15.05.2009 (23:45 GMT+03)	58	4.264024	*3.859052*
ZENIT	15.05.2009 (23:20 GMT+03)	206	4.259964	*4.117877*

IMAT 2010

- Traffic congestion prediction
- (Rough) data:
  - Modified graph of Moscow streets
  - Observed traffic speed 4-10 pm (4-min intervals) for 30 subsequent days + 4-6 pm on the 31st day
- Task: predict traffic speed 6-10 pm of the 31st day
- public/final evaluation
- http://imat2010.yandex.ru

Я

Modified graph of streets



IMAT 2010 Data

- Graph: vertices (139 241/33 029) and edges (206 260/86 249)
  - <id_vertex> <id_group>
  - <id_edge> <id_edge_group> <start_vert> <end_vert>
  - <id_edge_group> <length> <avg_speed>
- Observations (learning set, 29 226 208 lines)
  - <id_edge_group> <day> <time> <speed>
- Task (691 641 lines)
  - <id_edge_group> <day> <time> ??
- Evaluation

$$Err(a) = \frac{1}{n} \sum_{i=1}^{n} l_i t_i |a_i - v_i|$$

Рейтинг

Таблица объединяет финальный рейтинг конкурса (на 16.05.2010) и новые результаты. Подробнее о задаче и методике оценки см. Задача и данные.

Baseline — это «простая скептическая оценка»: средняя скорость для дуги по всем дням месяца для этого момента времени; если данных нет, то считаем, что скорость 0 км/ч (пробка).

Команда	Время последней загрузки	Количество попыток	Последний результат (публичная оценка)	Финальный результат
Сергей Гуда и Денис Рябов (ЮФУ)	16.05.2010 (23:21 GMT+03)	487	57.252870	*58.924831*
Паровоз	16.05.2010 (23:40 GMT+03)	252	58.137167	*59.039501*
malk	16.05.2010 (22:16 GMT+03)	9	58.352243	*59.450132*
RomanL	16.05.2010 (23:49 GMT+03)	252	60.117079	*60.021672*
УрГУ	16.05.2010 (23:50 GMT+03)	49	58.355595	*60.156530*
test	16.05.2010 (23:16 GMT+03)	246	58.427409	*60.165762*
Здравый смысл	31.07.2010 (15:27 GMT+03)	2	59.955410	*60.229683*
SlowMotion	16.05.2010 (20:51 GMT+03)	138	59.026925	*60.264464*
View	16.05.2010 (19:05 GMT+03)	87	59.470813	*60.349985*
Здравый смысл	16.05.2010 (01:20 GMT+03)	59	59.823316	*60.372594*
ММП	16.05.2010 (15:36 GMT+03)	241	59.996702	*60.745338*
GoGo	16.05.2010 (20:46 GMT+03)	49	60.911306	*61.392298*
kek_ksu	16.05.2010 (23:18 GMT+03)	28	60.452621	*61.556647*
DmitryS	16.05.2010 (23:38 GMT+03)	130	61.238659	*61.744422*
Kejpa	16.05.2010 (11:52 GMT+03)	73	60.728057	*61.907533*
Евгений Краско	16.05.2010 (14:13 GMT+03)	133	60.823800	*61.977751*
deas	15.05.2010 (23:27 GMT+03)	19	63.136674	*62.022574*
prognoz1	16.05.2010 (18:40 GMT+03)	21	62.245653	*62.334491*

ИМАТ 2011

Старт конкурса – февраль 2011

Задача интересная, победителю – приз ☺

ROMIP

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

ROMIP largest text collections

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

Image collections

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

RuSSIR

- Yekaterinburg, 5-12 September 2007
  http://romip.ru/russir2007
- Taganrog, 1-5 September 2008
  http://romip.ru/russir2008/
- Petrozavodsk, 11-16 September 2009
  http://romip.ru/russir2009/
- Voronezh, 13-18 September 2010
  http://romip.ru/russir2010/
- Saint Petersburg, 15-19 August 2011
  http://romip.ru/edbt-russir2011/

RuSSIR

Информационный поиск по-русски

Оригинальная английская версия: http://informationretrieval.org

Павел Браславский
pb@yandex-team.ru