Анализ поисковых запросов

Павел Браславский

Киев, 5-6 августа 2010 года

Я

План

- 1 день
  - Введение
  - Характеристики потока запросов
  - Данные для анализа
  - Сегментация
- 2 день
  - Машинное обучение (кратко)
  - Тематическая классификация
  - Близкие запросы
  - Академические инициативы Яндекса

ВВЕДЕНИЕ

Предварительные замечания

- Логи запросов – «опыт» и богатство МП
- «Меньше информации, больше контекста» (ср. Broder, 2007)
- Мало открытых данных (→ проблема для академических исследований)
- Проблемы с персональными данными (приватность)
- Очень короткие тексты – сложность анализа
- Недостаток информации компенсируется большими объемами данных (веб)

Статистика запросов к Яндексу

К началу графиков | CSV | +1 Добавить

День ▲	Посетители	Хиты	Хитов с посетителя
03.08.2010	11 826 883	74 087 506	6,26
02.08.2010	11 409 019	74 314 723	6,51
01.08.2010	8 145 747	51 549 726	6,33
31.07.2010	8 521 649	52 332 257	6,14
30.07.2010	11 497 052	69 706 604	6,06

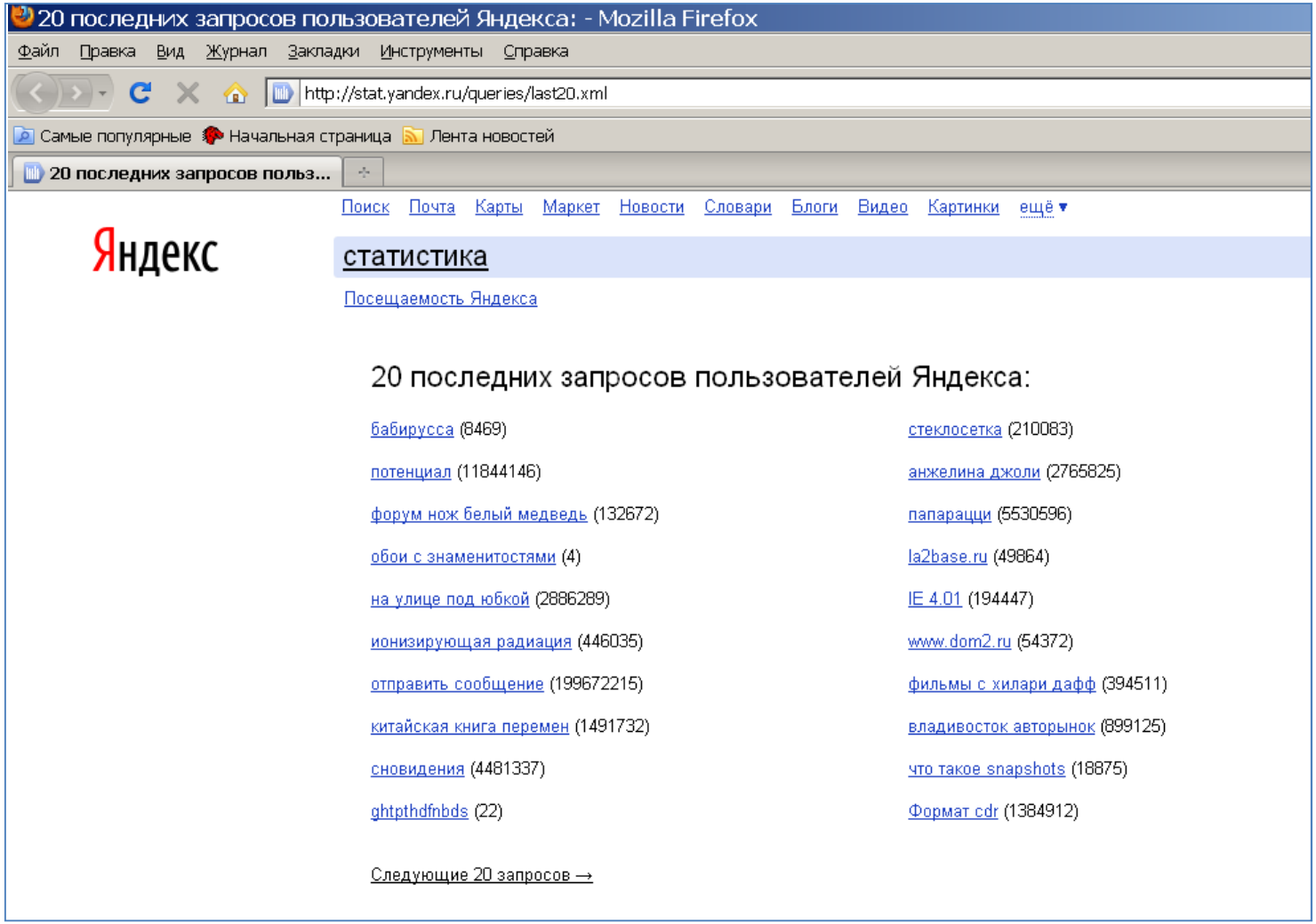
Россия

К началу графиков | CSV | +1 Добави

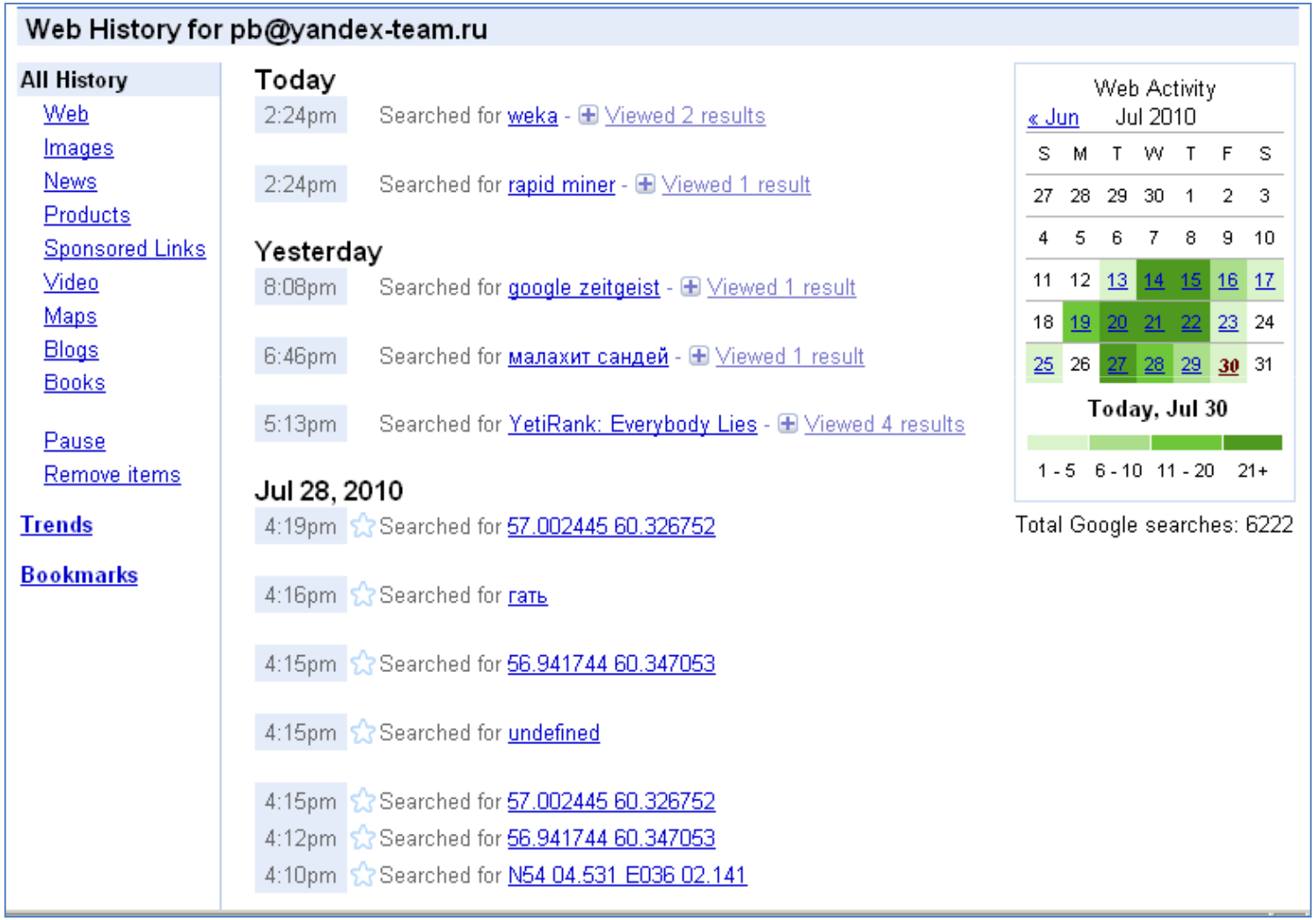
День ▲	Посетители	Хиты	Хитов с посетителя
03.08.2010	1 164 551	7 009 300	6,02
02.08.2010	1 208 839	7 148 688	5,91
01.08.2010	906 763	5 789 317	6,38
31.07.2010	927 704	5 800 029	6,25
30.07.2010	1 139 321	6 997 894	6,14

Украина

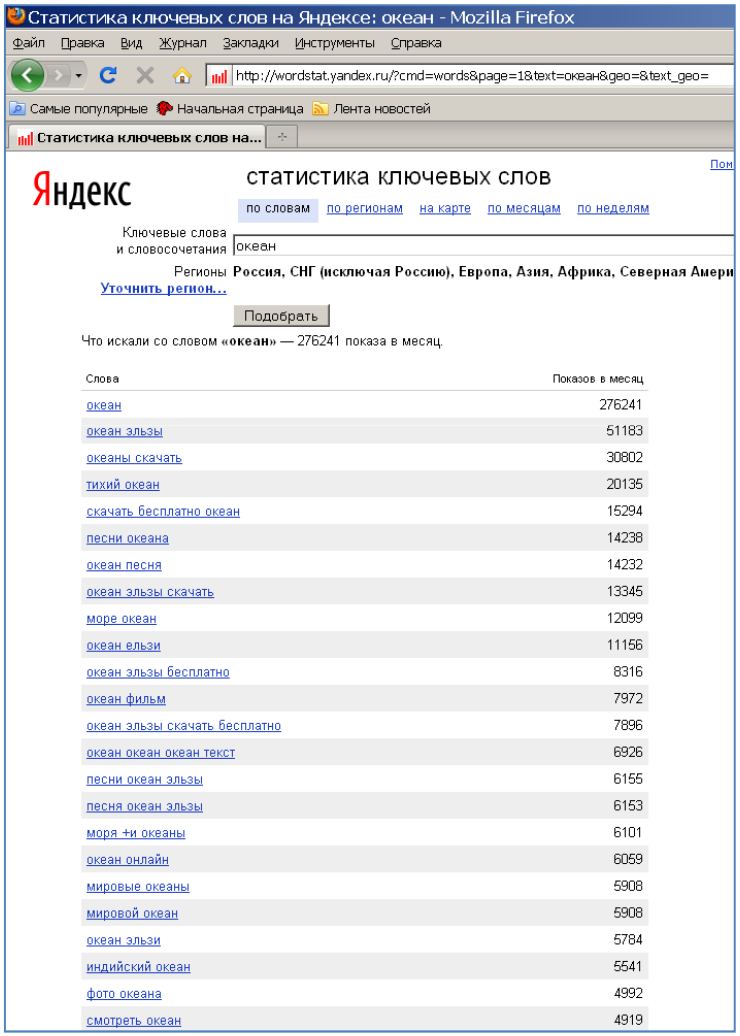
Прямой эфир


20 последних запросов пользователей Яндекса: - Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
http://stat.yandex.ru/queries/last20.xml
Самые популярные Начальная страница Лента новостей
20 последних запросов польз...
Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё
Яндекс
статистика
Посещаемость Яндекса
20 последних запросов пользователей Яндекса:
бабирусса (8469)
потенциал (11844146)
форум нож белый медведь (132672)
обои с знаменитостями (4)
на улице под юбкой (2886289)
ионизирующая радиация (446035)
отправить сообщение (199672215)
китайская книга перемен (1491732)
сновидения (4481337)
ghtpthdfnbds (22)
стеклосетка (210083)
анжелина джоли (2765825)
папарацци (5530596)
la2base.ru (49864)
IE 4.01 (194447)
www.dom2.ru (54372)
фильмы с хилари дафф (394511)
владивосток авторынок (899125)
что такое snapshots (18875)
Формат cdr (1384912)
Следующие 20 запросов →

Личная история запросов

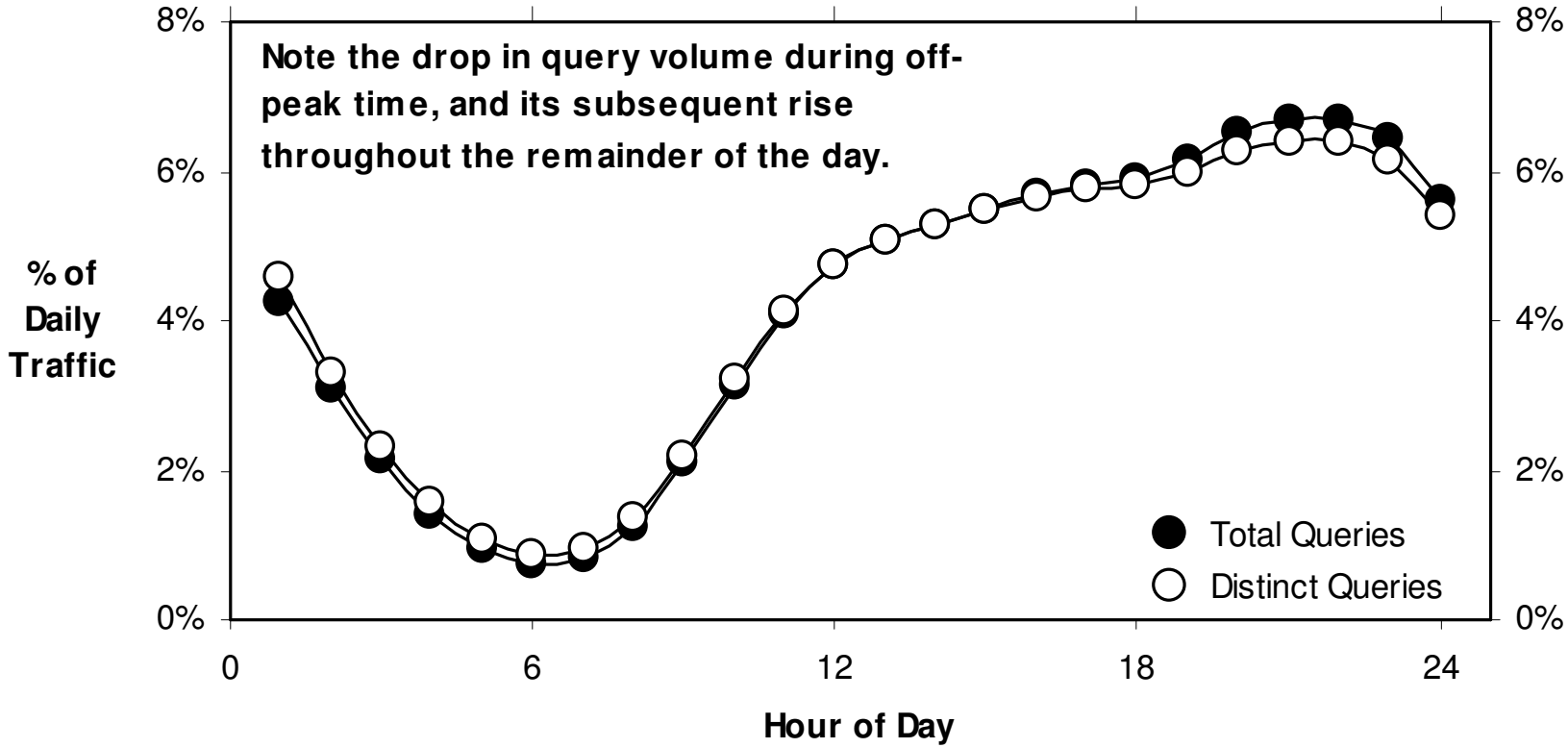


Статистика слов запросов



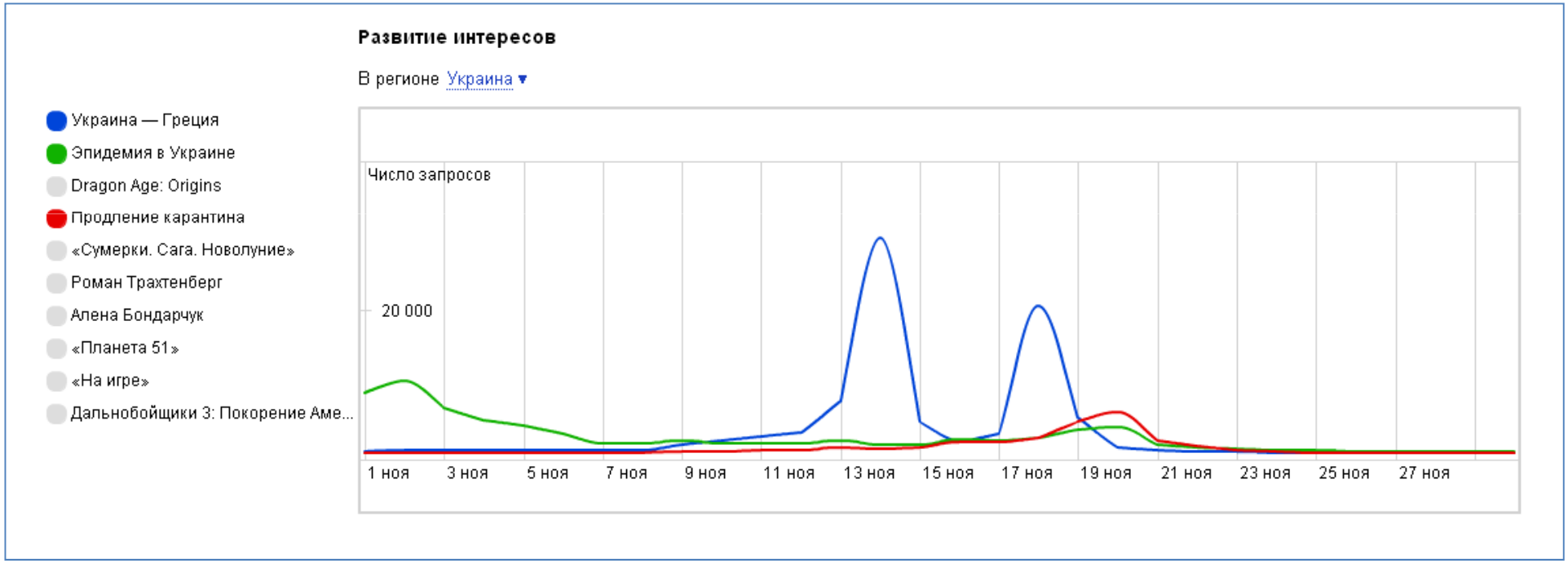
Слова	Показов в месяц
океан	276241
океан эльзы	51183
океаны скачать	30802
тихий океан	20135
скачать бесплатно океан	15294
песни океана	14238
океан песня	14232
океан эльзы скачать	13345
море океан	12099
океан ельзи	11156
океан эльзы бесплатно	8316
океан фильм	7972
океан эльзы скачать бесплатно	7896
океан океан океан текст	6926
песни океан эльзы	6155
песня океан эльзы	6153
моря +и океаны	6101
океан онлайн	6059
мировые океаны	5908
мировой океан	5908
океан эльзи	5784
индийский океан	5541
фото океана	4992
смотреть океан	4919

Traffic Volume Over a Day



[Beitzel]

Динамика запросов

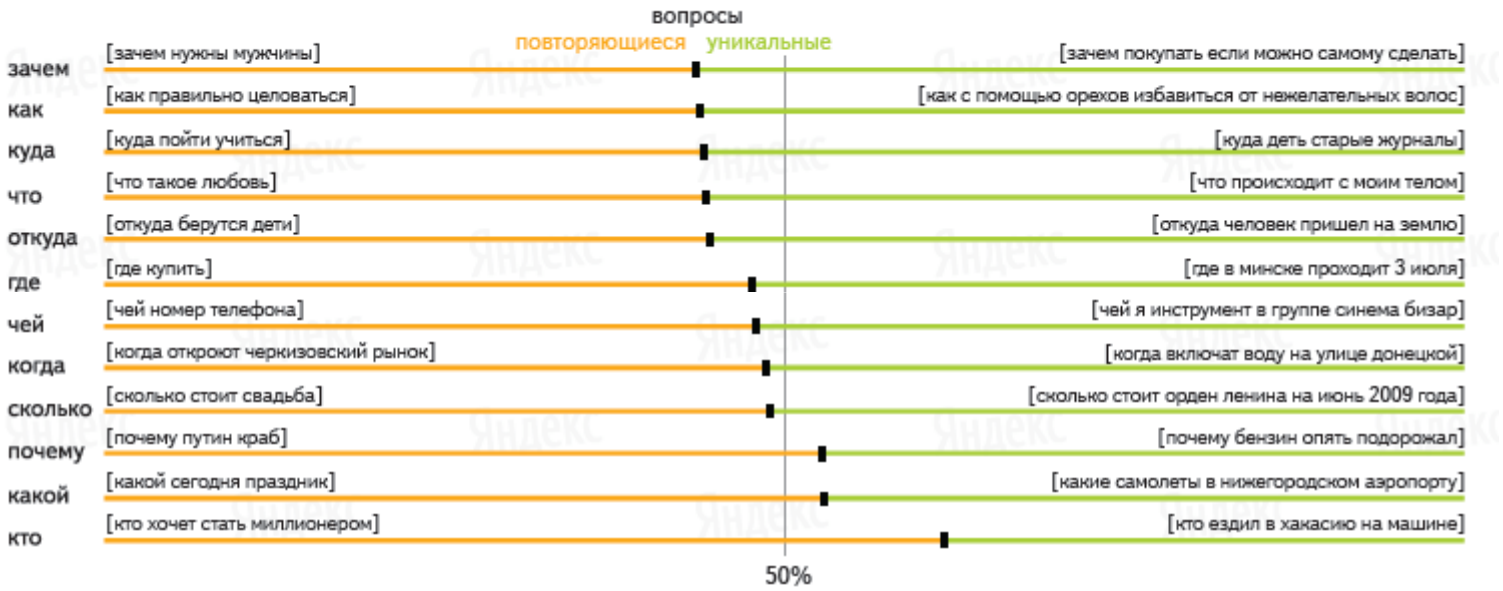


http://interes.yandex.ru/

Несколько цифр

- Длина запроса 2-3 слова
- Поисковая сессия в среднем 3 запроса
- 2-3% сформулированы как вопрос
- 12-15% запросов содержат опечатки

Запросы – вопросы



http://company.yandex.ru/facts/researches/ya_search_2009.xml

Классификация запросов / информационных потребностей (Broder, 2002)

- **Informational** – want to learn about something (~40% / 65%)

  Low hemoglobin

- **Navigational** – want to go to that page (~25% / 15%)

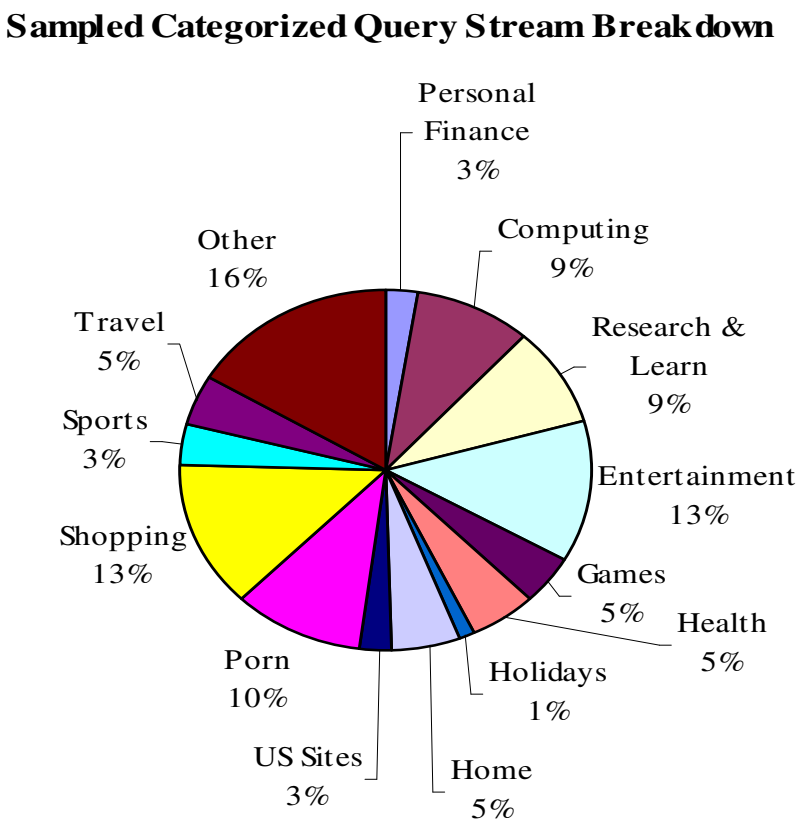
  Warsaw Airport

- **Transactional** – want to do something (web-mediated) (~35% / 20%)
  - Access a service — Barcelona weather
  - Downloads — Mars surface images
  - Shop — Canon S410
- Gray areas
  - Find a good hub — Car rental Poland
  - Exploratory search "see what's there"

Baeza-Yates

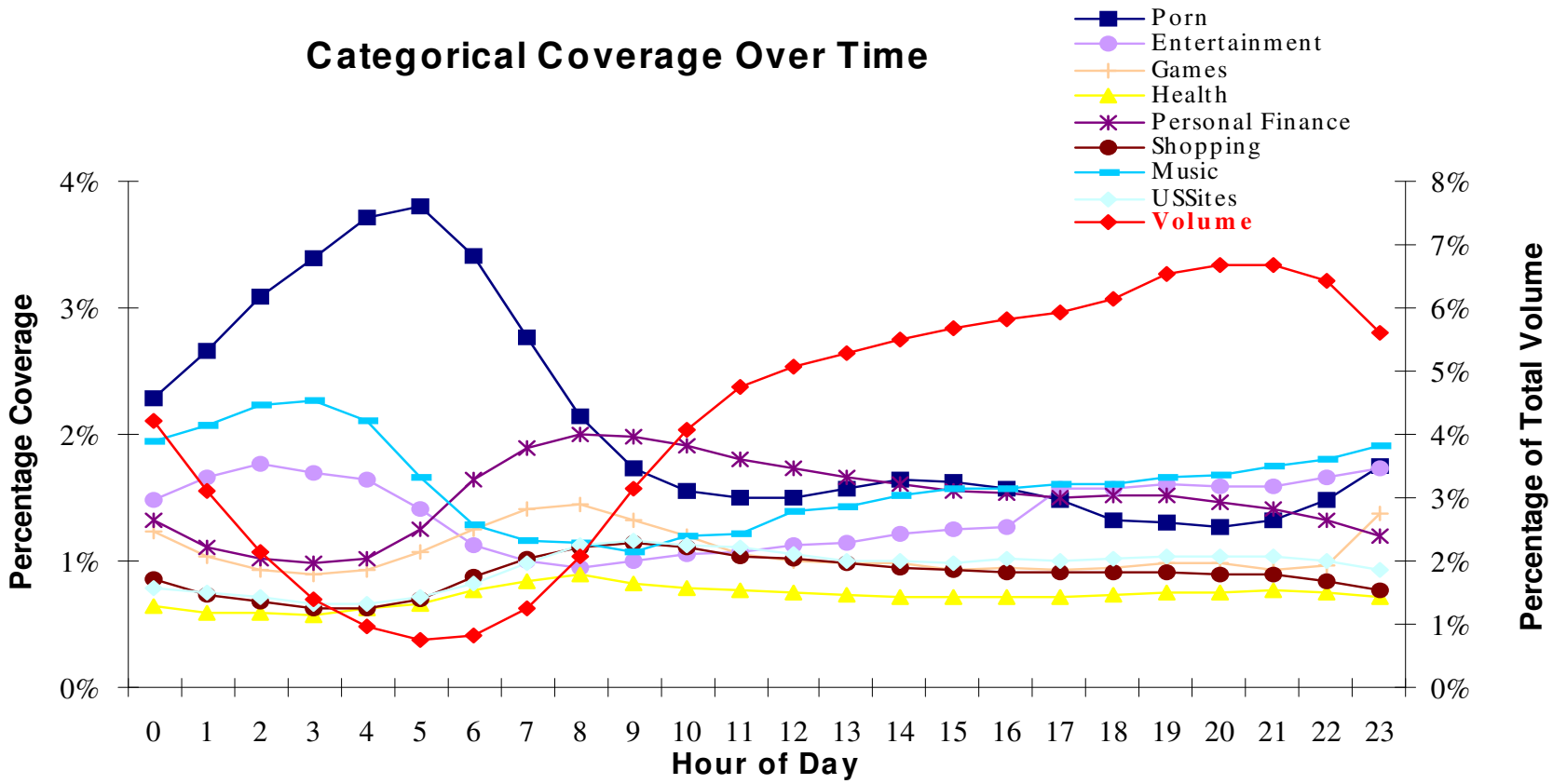
Category Breakdown

- Query lists for each category formed by a team of human editors
- Query stream classified by exactly matching each query to category lists



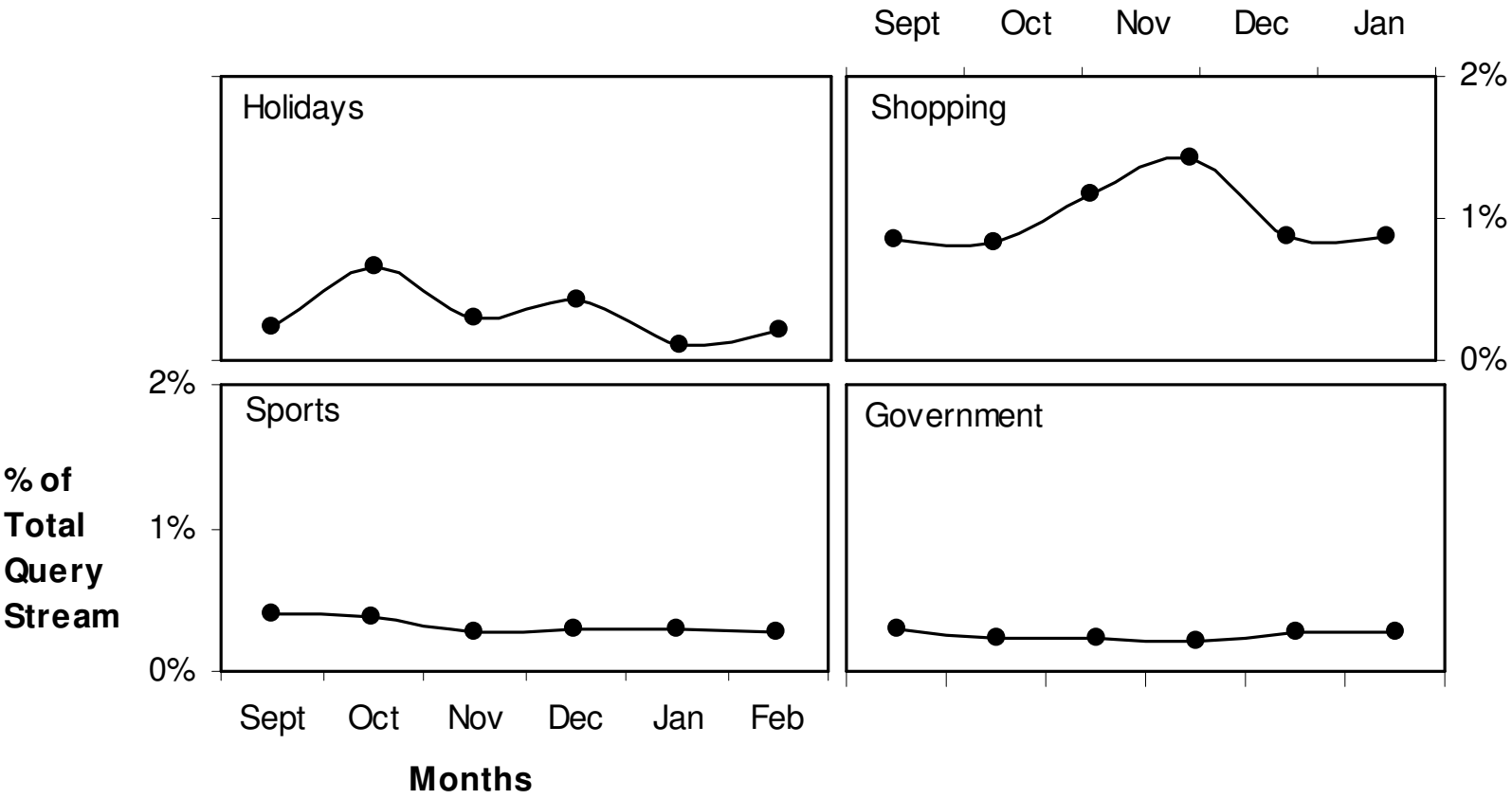
[Beitzel]

Category Popularity Over a Day



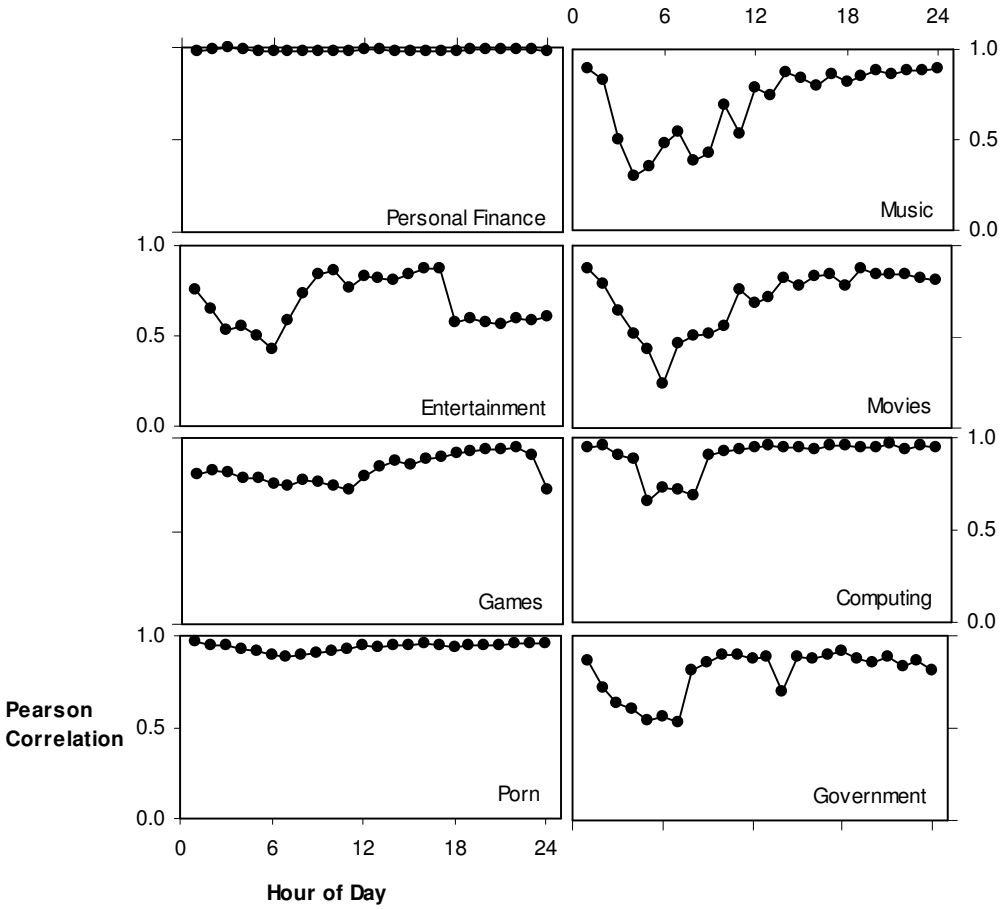
[Beitzel]

Category Popularity Over Six Months



[Beitzel]

Pearson Correlations for Selected Categories Over A Day



[Beitzel]

Какая информация у нас есть?

- текст запроса
- время
- IP → география
- Cookie → (уникальный) пользователь
- клики на результатах поиска
- (персональные данные, соцдем)

Данные (→ методы)

- отдельные запросы
- список запросов
- + время
- + сниппеты/документы
- + клики
- …

Яндекс: ~6 Кбайт/запрос + ~0,5 Кб/клик

Доступные данные

- Excite 1997, 1999, 2001
- AOL 2006
- ИМАТ 2004
- MSN Search query Log excerpt (RFP 2006 dataset)
- …

Excite 1997

- Запросы за один день (16 сентября 1997 г.)
- userId, timeStamp, query
- ~1M запросов (много повторов)

0C6B5395895CD808	970916125351	henri rousseau
0C6B5395895CD808	970916125511	henri rousseau+tiger
949946B881F137F0	970916115517	"pharmacy"
949946B881F137F0	970916115550	prescriptions
91A98BC9BEDCF053	970916075435	australian+chat+victoria
61305D2ADC74BC78	970916095742	dailyplanet
61305D2ADC74BC78	970916095846	dailyplanet
33D1A0D49E8DB2AB	970916144916	maizehighschool
33D1A0D49E8DB2AB	970916144951	maize high school
FCBB8401805D783F	970916212508	warez strata studio pro
FCBB8401805D783F	970916212541	warez mac
12FE04344578F249	970916202819	"midwife conference"
12FE04344578F249	970916202924	midwifery
477CC4190EF76EB4	970916165602	nrwmac
477CC4190EF76EB4	970916172706	npac
477CC4190EF76EB4	970916175242	nrwmac

ИМАТ 2004

- 7 дней * 10% от 5-10 миллионов запросов в день → 3,5 Гб

```
<UID1>
        <запрос1> <время> <найдено документов> <номер страницы>
                <URL1> <время выбора>
                <URL2> <время выбора>
                ...
        <запрос2> <время> <найдено документов> <номер страницы>
                <URL1> <время выбора>
                <URL2> <время выбора>
                ...
        ...
<UID2>
        ...
```

http://company.yandex.ru/academic/grant/datasets_description.xml

AOL 2006

- Большой скандал!!!
- ~20M web queries from ~650k users over three months

AnonID	Query	QueryTime	ItemRank	ClickURL
993	myspace.co	01.03.2006 12:13		
993	myspace.com	01.03.2006 12:13		
993	googl	01.03.2006 15:03		
993	chasebadkids.net	03.03.2006 16:55	1	http://www.chasebadkids.net
1268	ozark horse blankets	01.03.2006 17:39	8	http://www.blanketsnmore.com
1268	www.ghostrockranch.com	04.03.2006 13:58		
1268	openrangeht.zachsairforce.com	09.03.2006 22:38		
1268	sstack.com	11.03.2006 0:17		
1268	www.mecab.org	12.03.2006 18:59		
1268	www.raindanceexpress.com	18.03.2006 20:13		
1268	www.victoriacostumiere.com	19.03.2006 0:26		
1268	osteen-schaztberg.com	21.03.2006 17:55		
1268	osteen-schatzberg.com	21.03.2006 17:55	1	http://www.osteen-schatzberg.com
1268	osteen-schatzberg.com	21.03.2006 17:55	2	http://www.osteen-schatzberg.com

MSN Search query Log excerpt

- 15 million queries
- Sampled over one month
- Queries from the US site (mostly English)

Per query attributes included:

- Session ID
- Time-stamp
- Query string
- Number of results on results page
- Results page number

Data per query for each result clicked:

- URL
- Associated query
- Position on results page
- Time-stamp

СЕГМЕНТАЦИЯ ЗАПРОСОВ

Сегментация запросов

Сегментация:

1. поиск
2. дальнейшая обработка запросов

международный почтамт | москва
молодежный отдых | в турции
официальный сайт | автоваз
Купить | кроссовки | Nike Zoom BB
магазин | рыбачьте с нами
ирина круг | пусть сейчас я плачу| слушать
сбербанк россии | в алтайском крае
знак зодиака | близнецы
нино катамадзе | билеты
Смысл названия | рассказа | матренин двор
Сергей Тармашев | Корпорация | скачать | бесплатно

вечерняя москва vs пицца москва

банк москвы vs банки москвы

Сегментация на основе лога

$$conn(S) = freq(S)*I(w_1 \ldots w_{n-1}; w_2 \ldots w_n)$$

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}, \qquad SI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}.$$

```
msdn library visual studio

34259: (msdn library)[5110] (visual studio)[29149]
29149: msdn[47658] library[209682] (visual studio)[29149]
5110: (msdn library)[5110] visual[23873] studio[53622]
41: (msdn library visual studio)[41]
7: msdn[47658] (library visual studio)[7]
0: msdn[47658] library[209682] visual[23873] studio[53622]
```

Risvik et al. WWW2003

Сегментация на основе ML

нино | катамадзе | билеты

(0, 1, 0, 1, 1, 0) (1, 1, 0, 1, 0, 1)

Table 1: Indicator features.

Name	Description
is-the	token x = "the"
is-free	token x = "free"
POS-tags	Part-of-speech tags of pair $x_{L0}\ x_{R0}$
fwd-pos	position from beginning, i
rev-pos	position from end $N - i$

Table 2: Statistical features.

Name	Description
web-count	count of "x" on the web
pair-count	web count "$w\ x$"
definite	web count "the $w\ x$"
collapsed	web count "wx" (one word)
and-count	web count "w and x"
genitive	web count "w's x"
Qcount-1	Counts of "x" in query database
Qcounts-2	Counts of "$w\ x$" in database

Bergsma and Wang, 2007

Сегментация: «наивный подход»

На основе веб-корпуса n-грамм

$$score(S) = \sum_{s \in S, |s| \geq 2} |s|^{|s|} \cdot count(s)$$

Hagen at al. SIGIR2010

Сегментация: подходы

- похоже на выделение устойчивых словосочетаний
- например: генерация и взвешивание всех сегментов (частота + связность по логу); сегментация запроса = сумма весов всех сегментов
- + микросинтаксис
- лог vs корпус текстов
- ML (больше признаков, более богатое описание)
- language models
- внешние ресурсы (Wikipedia)
- эвристики ([дима билан]←[димабилан])

План

- 1 день
  - Введение
  - Характеристики потока запросов
  - Данные для анализа
  - Сегментация
- 2 день
  - Машинное обучение (кратко)
  - Тематическая классификация
  - Поиск похожих
  - Академические инициативы Яндекса

Машинное обучение

- См., например, материалы на www.informationretrieval.org

ТЕМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ

Тематическая классификация

- список запросов
- коллекция текстов
- веб
- клики

Selectional Preferences: Step 1

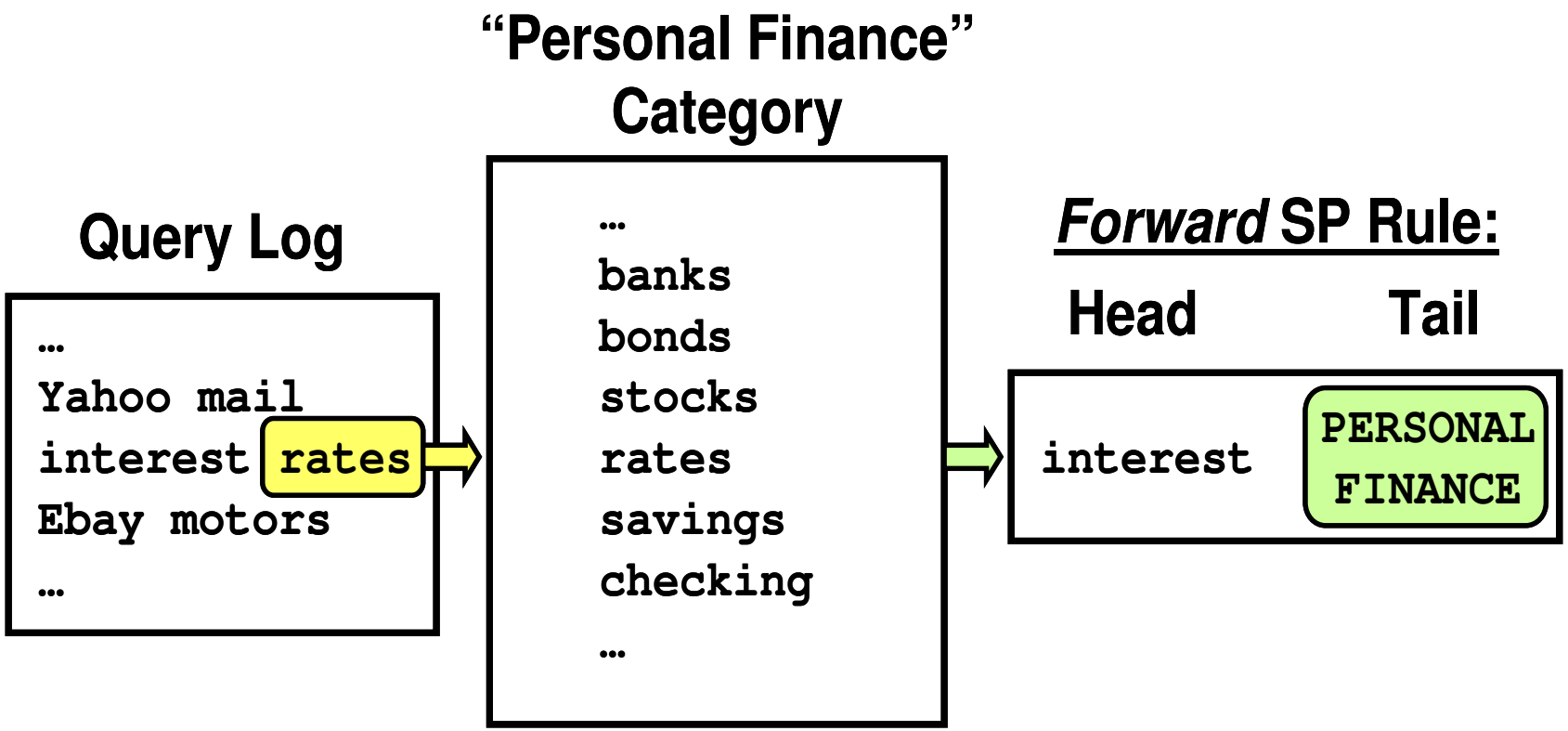
- Obtain a large log of unlabeled web queries
- View each query as pairs of lexical units:
  - <head, tail>
  - Only applicable to queries of 2+ terms
  - Queries with *n* terms form *n-1* pairs
  - Example: “directions to DIMACS” forms two pairs:
    - <directions, to DIMACS> and <directions to, DIMACS>

[Beitzel]

Selectional Preferences: Step 2

- Obtain a set of manually labeled queries
- Check the heads and tails of each pair to see if they appear in the manually labeled set
- Convert each <head, tail> pair into:
  - <head, CATEGORY> (*forward* preference)
  - <CATEGORY, tail> (*backward* preference)

Selectional Preferences: Step 2



[Beitzel]

Selectional Preferences: Step 3

- Score each preference using Resnik's Selectional Preference Strength formula:

$$S(x) = D(P(U|x) \,||\, P(U))$$
$$= \sum_{u} P(u|x) \log_2 \left(\frac{P(u|x)}{P(u)} \right)$$

- Where u represents a category, as found in Step 2.
- S(x) is the sum of the weighted scores for every category associated with a given lexical unit

Selectional Preferences: Step 4

- Use the mined preferences and weighted scores from Steps 3 and 4 to assign classifications to unseen queries

Incoming Query

`interest only loan`

***Matching* SP Rule:**

Head	Tail
`interest`	`PERSONAL FINANCE`

"`interest only loan`" **is classified as a** `PERSONAL FINANCE` **query.**

[Beitzel]

Selectional Preference Rule Examples

Forward Rules

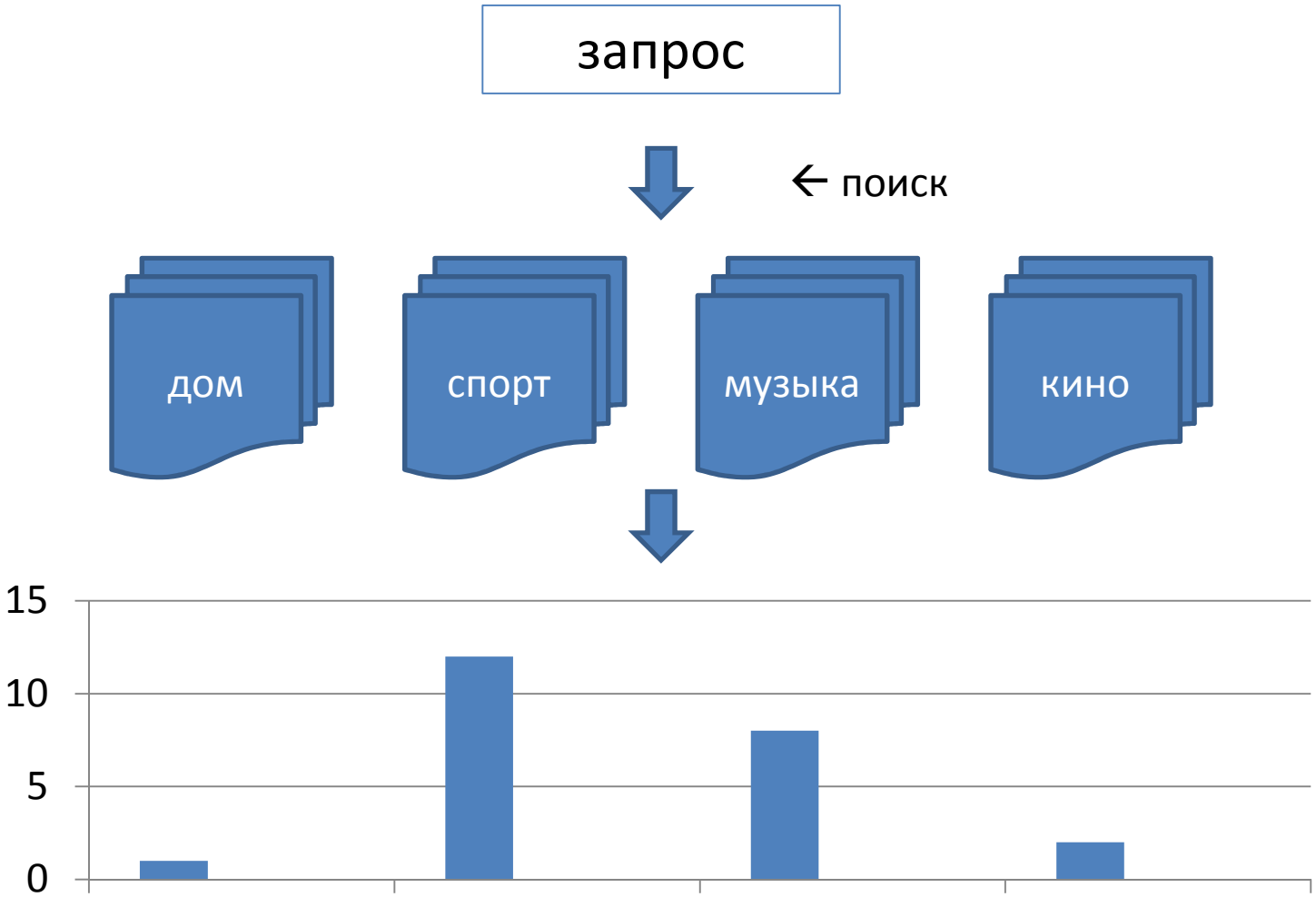
- harlem club ***X***
  - ENT->0.722
  - PLACES->0.378
  - TRAVEL->1.531
- harley all stainless ***X***
  - AUTOS->3.448
  - SHOPPING->0.021
- harley chicks with ***X***
  - PORN->5.681

Backward Rules

- ***X*** gets hot wont start
  - AUTOS->2.049
  - PLACES->0.594
- ***X*** getaway bargain
  - PLACES->0.877
  - SHOPPING->0.047
  - TRAVEL->0.862
- ***X*** getaway bargain hotel and airfare
  - PLACES->0.594
  - TRAVEL->2.057

[Beitzel]

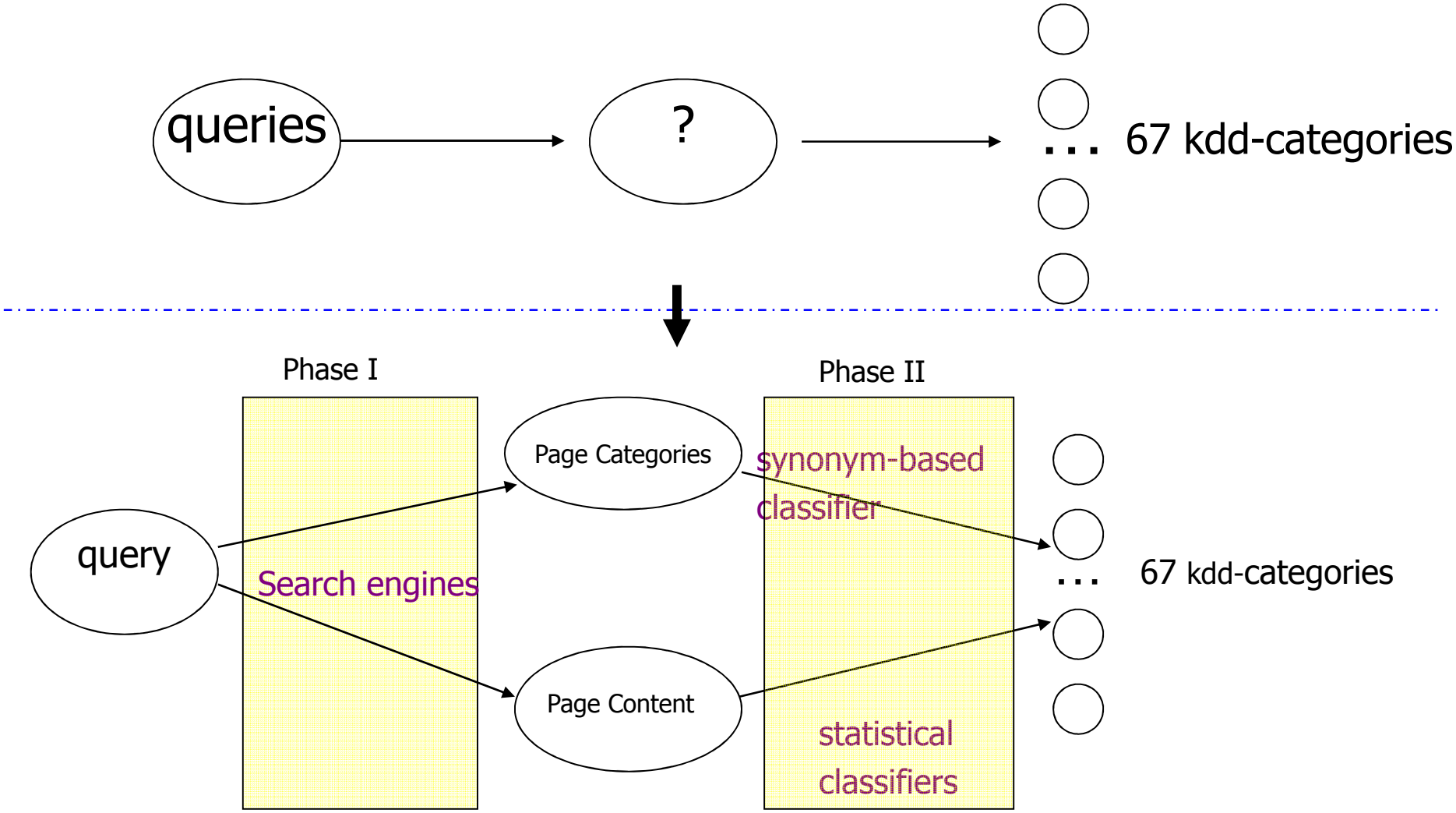
На основе коллекции текстов



KDD Cup 2005

- Классификация запросов
- 800,000 запросов, 67 категорий
- есть примеры, нет обучающего множества
- нет подробного описания категорий
- ответ системы: до 5 категорий
- оценка: 800 запросов оцениваются тремя асессорами
- метрика: F1

Nature of Problem



[Shen]

Phase I: From queries to pages and categories

- Input:
  - A query: Q_i,
- Output:
  - <Page $list_i$, Category $list_i$ >
- Approach:
  - through Search Engines (SE)

Computerworld
Category: Computers > News and Media > Magazines and E-zines > Magazines
News and product coverage for information technology managers.
www.computerworld.com/ - Cached - Similar pages

Computer Associates
Category: Computers > Security > Authentication > Single Sign-On
Makers of eTrust Single Sign-On. Automates access to authorized Web services and enterprise applications.
www.ca.com/ - Cached - Similar pages

- We collected
  - 40 million entries
  - 50GB
- Search engines
  - Lumur (CMU open source)
  - Google
  - ODP
  - Looksmart

[Shen]

Phase II.a Synonym-based Classifier: using directories

- 67 KDD-categories in KDDCUP

↕

- 172,565 in ODP/Google, 272,405 in Looksmart

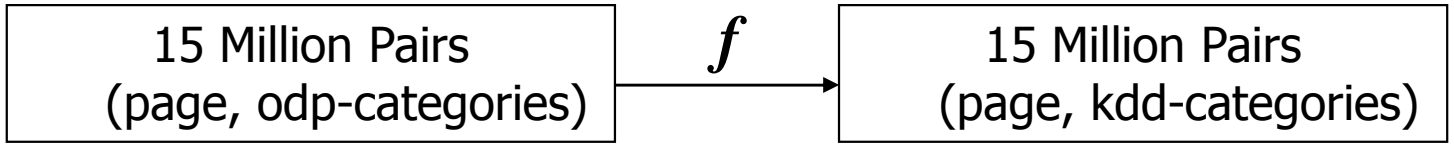
- For each of the KDDCUP category
  - Apply Wordnet to find the corresponding synonyms in the categories of ODP (Google) and Looksmart, respectively
- This produces one mapping function f for each directory
  - Also returns a rank by matching frequency

- Advantage
  - Fast,
  - Precise
- Disadvantage
  - Many of the 172K and 272K categories from ODP/Google and Looksmart do not map to KDDCUP categories
  - This may result in low recall

[Shen]

Phase II.b: Statistical Classifiers

- Statistical Classifiers
  - Support Vector Machine (SVM): mapping pages to KDDCUP categories
- Training Data
  - 15 million pages with categories from ODP Directory
  - Apply the mapping f from Phase II.a, to build training data.



- Application of the classifier
  - Construct a virtual document for each query by combining the snippets from the returned pages given in Phase I.
  - Classifier returns category and rank

Component Classifier Integration

- We follow an *ensemble learning* approach
  - Each classifier returns the category and rank
  - The two kinds of classifiers have the similar performance.
  - We integrate the different classifiers together by a weighted sum of the ranks
    - Weights can be determined by validation data set:
      - Based on the performance on the 111 sample data;
      - Assign different weight values for a classifier on different categories
      - The higher the precision, the higher the weight value
    - We have also tried to use equally weighted component classifiers

[Shen]

Final Result Generation

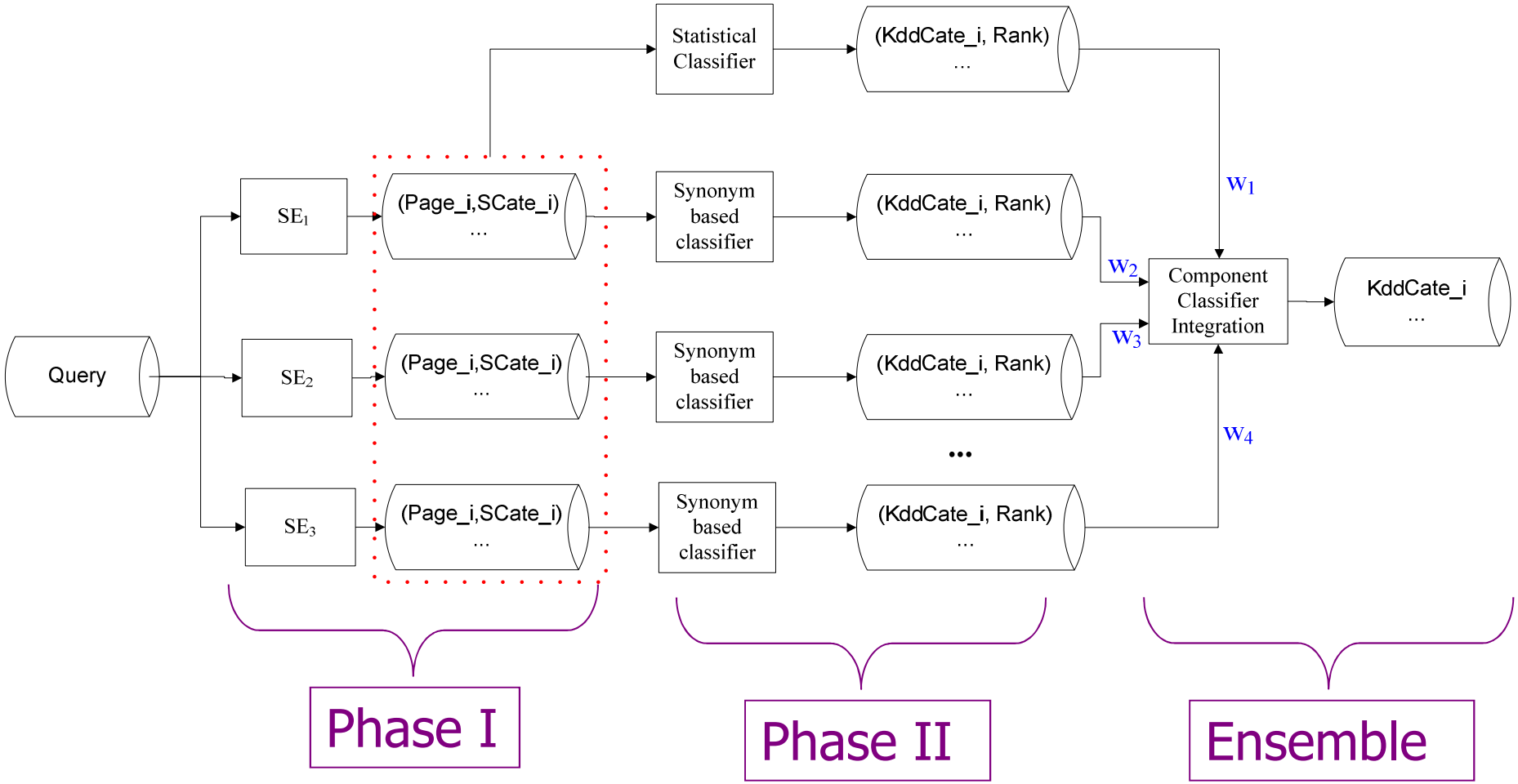
- Two Solutions: One for each evaluation criteria
  - ***S1:*** Using the validation data set is expected to achieve better precision measure
    - Since each component classifier is highly weighted on the classes where it achieves high precision.
  - ***S2:*** Equally weighted combination is expected to achieve higher F1 performance
    - Since the recall is relatively high
  - Evaluation Results (http://www.acm.org/sigs/sigkdd/kdd2005/kddcup.html)

	Submission ID	Precision	F1
S1	37	0.423741	0.426123
S2	22	0.414067	0.444395

- The Results are generated automatically.

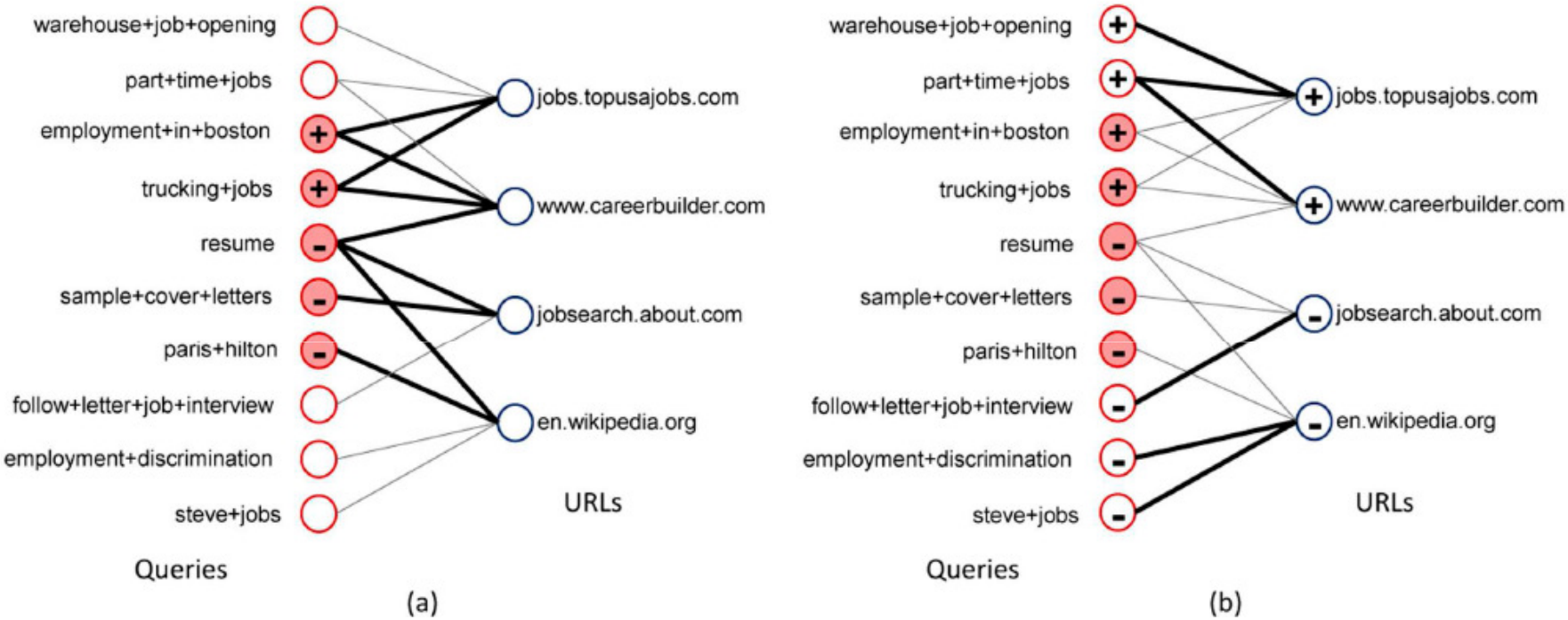
[Shen]

Putting them together



[Shen]

Классификация на основе кликов



1. Расширение обучающего множества
2. Классификация на основе лексических признаков

Li et al. SIGIR2008

БЛИЗКИЕ ЗАПРОСЫ

михаил булгаков

мастер и маргарита

мастер и маргарита фильм

владимир бортко

тарас бульба фильм

богдан ступка

михаил боярский

д'артаньян

три мушкетера

александр дюма

Подробнее

1. Уточнение:
   золотое кольцо → золотое кольцо с бриллиантом
2. Снятие неоднозначности:
   ягуар → ягуар животное
3. Расширение:
   золотое кольцо → ювелирные украшения
4. Синоним:
   японская вишня → сакура

Еще подробнее

5. Другой запрос на ту же тему:
   мерседес → ауди
6. Смена поисковой цели:
   купить санки → детский мир
7. Перевод:
   коралловый клуб → coral club

Близкие, но «про другое»

- Опечатки: *курсовая р**о**бота – курсовая работа*
- Транслитерация: *золото – zoloto*
- Раскладка: *lbvf ,bkfy – дима билан*
- Реникса: *otbeptka – отвертка*

Хороший запрос – это непросто

Назовите глагол из вопроса, помещенного на борту транспортного средства подопечных Фатиха Терима на первенстве континента?

IX Кубок Яндекса по поиску (2008)

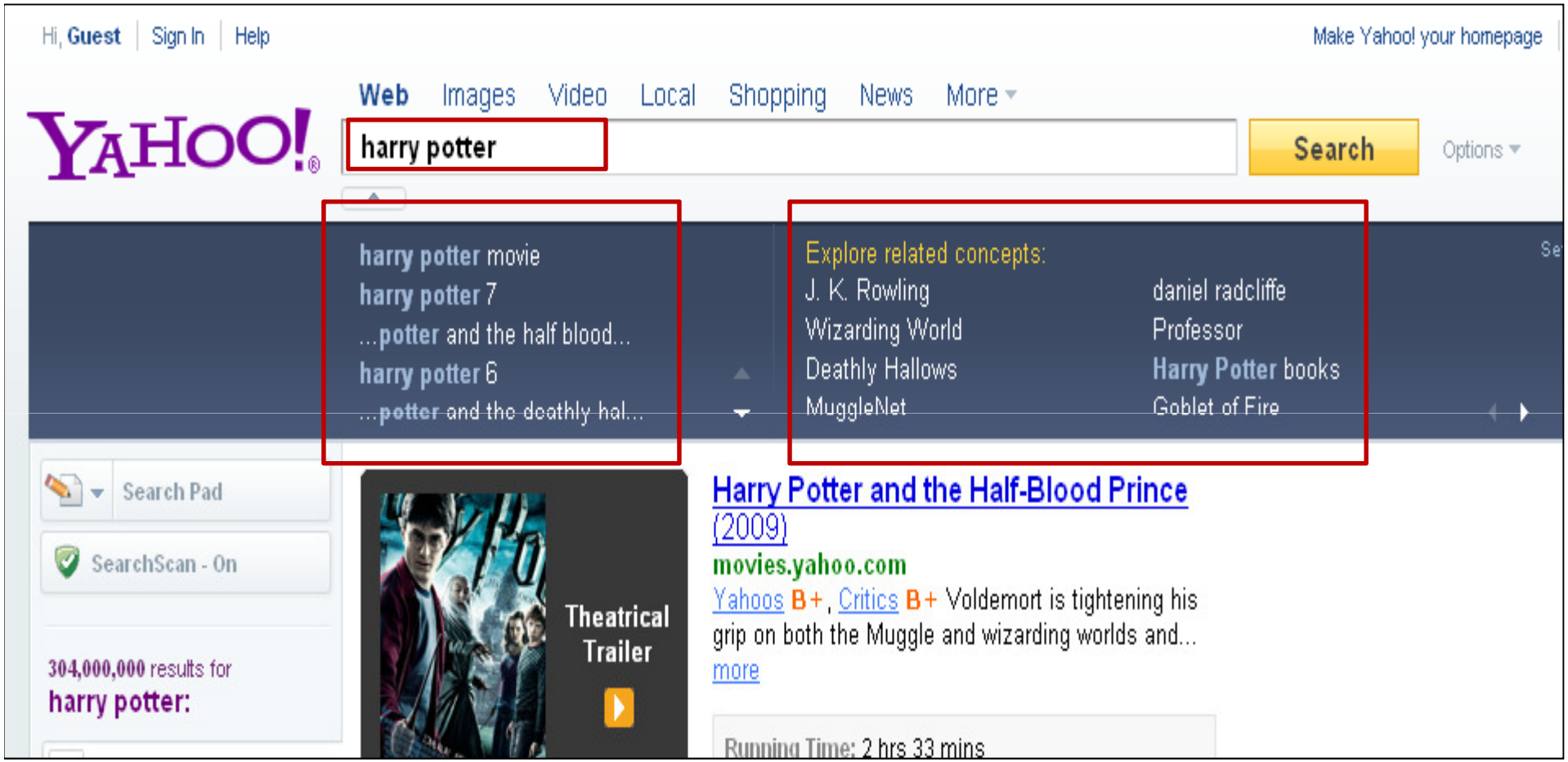
http://kubok.yandex.ru

Работа мысли

- Фатих Терим
- сборная турции на борту самолета
- сборная турции "на борту" самолета
- сборная турции надпись "на борту" самолета
- сборная турции надпись "на борту" автобуса
- сборная турции надпись на автобусе
- футбол "сборная турции" надпись на автобусе
- чемпионат европы футбол "сборная турции" надпись на автобусе

Ответ

Вместит ли автобус всю страсть Турции?


Hi, Guest
Sign In
Help
Make Yahoo! your homepage
Web
Images
Video
Local
Shopping
News
More
YAHOO!
harry potter
Search
Options
harry potter movie
harry potter 7
...potter and the half blood...
harry potter 6
...potter and the deathly hal...
Explore related concepts:
J. K. Rowling
Wizarding World
Deathly Hallows
MuggleNet
daniel radcliffe
Professor
Harry Potter books
Goblet of Fire
Search Pad
SearchScan - On
304,000,000 results for
harry potter:
Theatrical Trailer
Harry Potter and the Half-Blood Prince (2009)
movies.yahoo.com
Yahoos B+, Critics B+ Voldemort is tightening his grip on both the Muggle and wizarding worlds and...
more
Running Time: 2 hrs 33 mins

chevrolet - Ask.com Search

AskEraser Set

chevrolet

Search

Advanced Search

Web | Images | News | Deal$ | Videos | Q&A Beta | More

Showing 1-10 of 7,870,0

Today's best deals for **chevrolet**

Chevrolet

Chevrolet sells an impressively wide range of vehicles, from subcompact hatchbacks to huge vans and SUVs. If you're looking for a... More »

See Also: Official Site · Chevrolet Models · Reviews · History

Source

Search For: Chevrolet Accessories · Images

Popular Models: Silverado 1500 · Camaro · Corvette · Malibu Hybrid

Продаешь Шевроле?

Sponsored Result

Размести бесплатное объявление На новой торговой площадке Рунета

Avito.ru

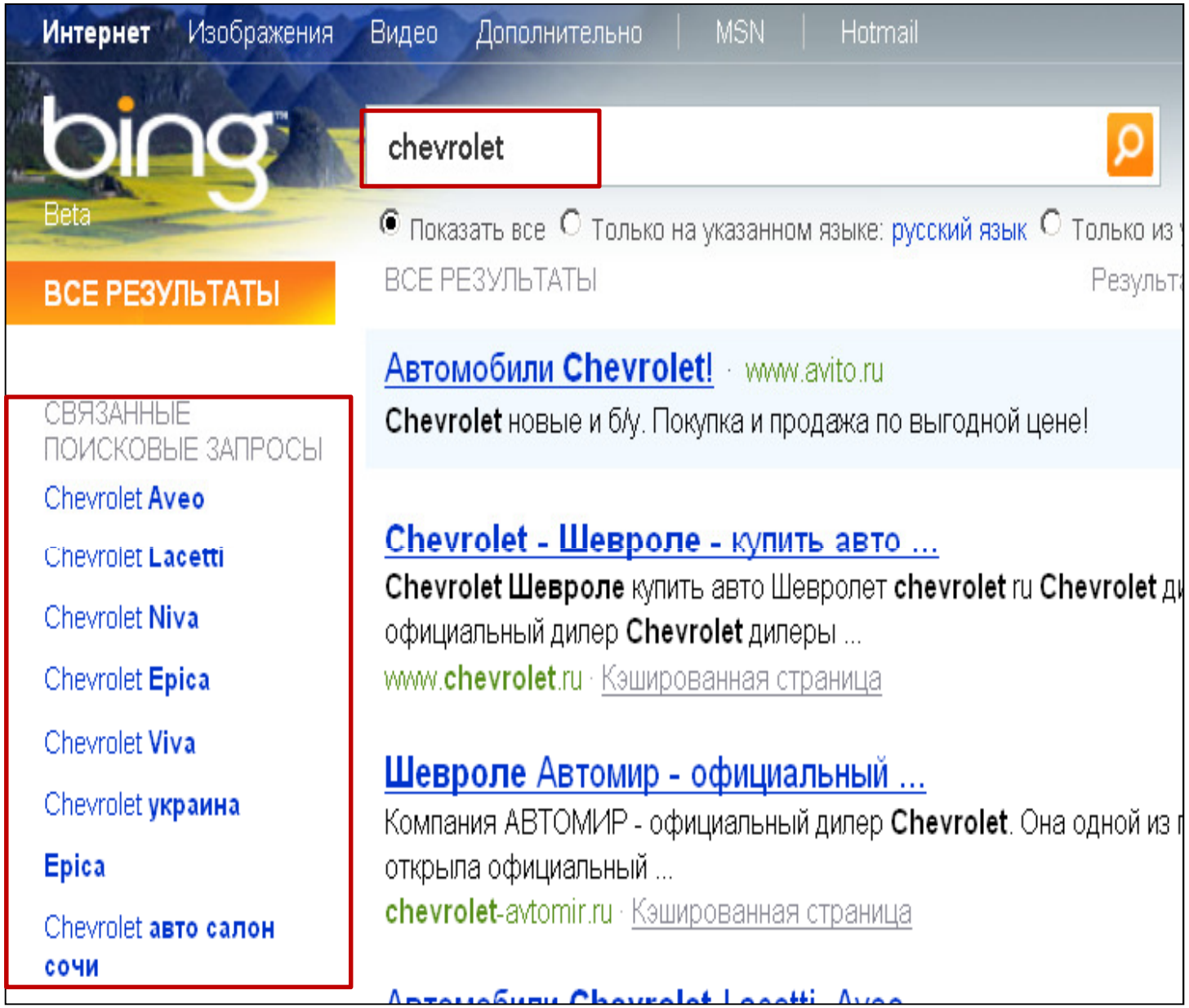
2010 Chevy Corvette | Performance Cars | Corvette Cars

Chevrolet Home ... Experience **Chevrolet** ... Close **Chevrolet** Worldwide...

www.chevrolet.com/corvette/ - Cached

Related Searches

- **Chevrolet** Camaro
- **Chevrolet** History
- **Chevrolet** Parts
- **Chevrolet** History Beginning
- **Chevrolet** Logo
- **Chevrolet** History of Cars
- Who Invented **Chevrolet**
- **Chevrolet** Truck Parts
- GMC
- Pontiac
- General Motors
- Louis **Chevrolet**


Интернет
Изображения
Видео
Дополнительно
MSN
Hotmail
bing
Beta
chevrolet
Показать все
Только на указанном языке: русский язык
Только из
ВСЕ РЕЗУЛЬТАТЫ
ВСЕ РЕЗУЛЬТАТЫ
СВЯЗАННЫЕ ПОИСКОВЫЕ ЗАПРОСЫ
Chevrolet Aveo
Chevrolet Lacetti
Chevrolet Niva
Chevrolet Epica
Chevrolet Viva
Chevrolet украина
Epica
Chevrolet авто салон сочи
Автомобили Chevrolet! · www.avito.ru
Chevrolet новые и б/у. Покупка и продажа по выгодной цене!
Chevrolet - Шевроле - купить авто ...
Chevrolet Шевроле купить авто Шевролет chevrolet ru Chevrolet
официальный дилер Chevrolet дилеры ...
www.chevrolet.ru · Кэшированная страница
Шевроле Автомир - официальный ...
Компания АВТОМИР - официальный дилер Chevrolet. Она одной из
открыла официальный ...
chevrolet-avtomir.ru · Кэшированная страница

Video results for hermione granger
Hermione Granger
3 min 23 sec - 4 May 2006
www.youtube.com
Emma Watson/Hermione Granger *Lucky*
3 min 26 sec - 28 Jan 2008
www.youtube.com
Searches related to hermione granger
hermione granger pictures
ginny weasley
harry potter quiz
hermione granger costume
ron weasley
draco malfoy
harry potter
emma watson
Goooooooooogle
1 2 3 4 5 6 7 8 9 10
Next
hermione granger
Search within results - Language Tools - Search Help - Dissatisfied? Help us in

Поиск Почта Карты Маркет Новости Словари Блоги Видео **Картинки** ещё ▾

Яндекс
картинки

гарри поттер

в найденном

Любые Обои Большие Средние Маленькие Портреты

Поищите также:
гермиона harry potter джинни уизли
рон уизли дамблдор драко малфой
хогвартс снейп ден рэдклифф

Яндекс.Видео

Гарри Поттер 3
Еще по запросу: «гарри поттер» 71145

Гарри Поттер и Орден феникса (**Harry Potter** and the Order of the...
240×215 ... 699×631
www.kino-govno.com

73. **Harry Potter** 5 / **Гарри Поттер** 5Просмотров
62×100 ... 850×1360
gallery.galaxy-portal.ru

Гарри Поттер и орден Феникса / **Harry Potter** and the Order of the...
120×90 ... 1500×1125
relax.grad66.ru

Поиск Почта Карты Маркет Новости Словари Блоги Видео **Картинки** ещё ▾

Яндекс
картинки

эмма уотсон

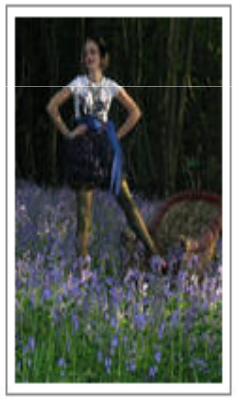
в найденном

Любые Обои Большие Средние Маленькие Портреты

Поищите также:
руперт гринт
ден рэдклифф
том фелтон
бонни райт
эмма робертс
эванна линч
emma watson
эмма уотсон 2009

Яндекс.Фотки

Еще по запросу:
«эмма уотсон» 1441



Эмма Уотсон (**Emma** Watson) in Teen Vogue (фотосессия.
383×511 ... 500×666
www.port2all.dp.ua

Эмма Уотсон (**Emma** Watson) in Teen Vogue (фотосессия.
511×383 ... 690×518
www.port2all.dp.ua

Эмма Уотсон (**Emma** Watson) in Teen Vogue (фотосессия.
511×391 ... 700×535
www.port2all.dp.ua

Источники данных

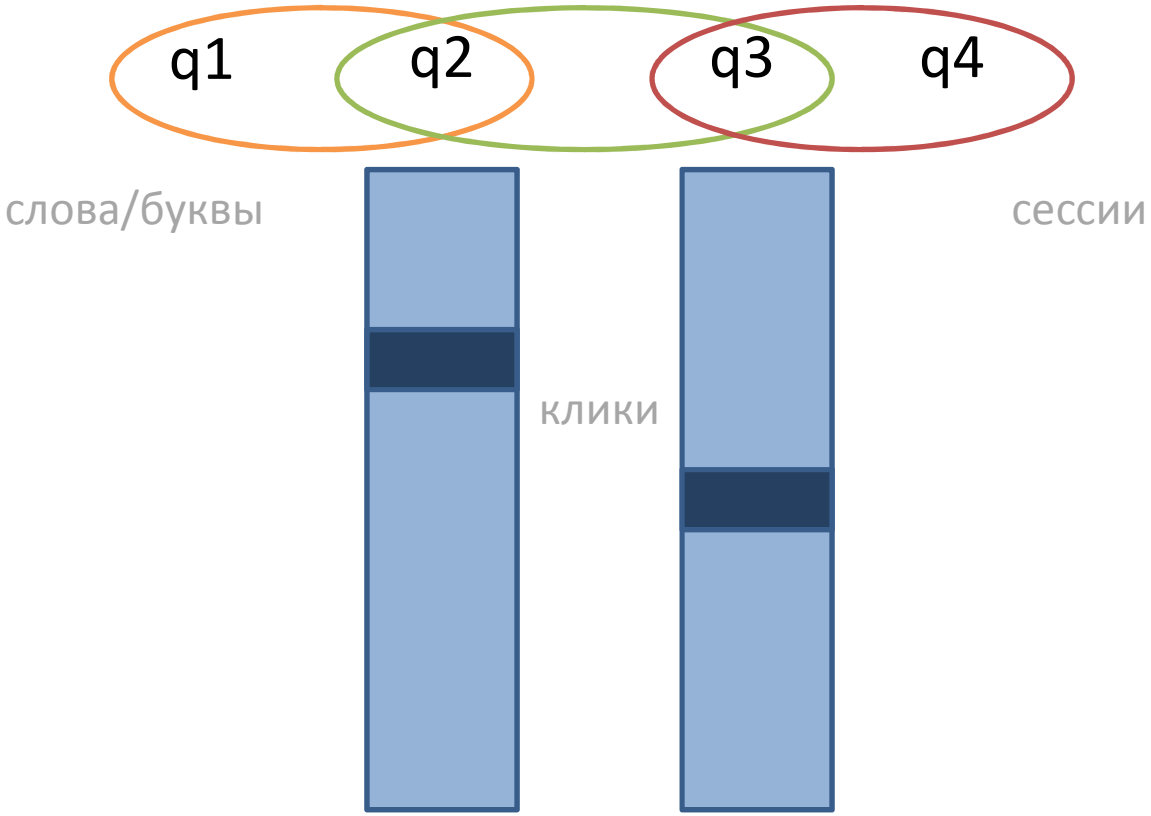
- Лог запросов
- Текст ссылок
- Корпус текстов

ford → ford focus, ford fusion, ford mondeo

карта → карта памяти, карта города

машина → стиральная машина, швейная машина

Близость запросов

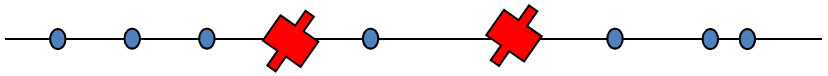


Метод

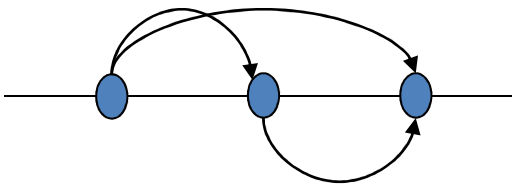
запрос=<timeStamp, userId, queryText, ckicks>

0. чистка лога (порно, Яндекс, «подсказки», язык запросов, опечатки)

1. выделение сессий (по времени)



2. выделение пар (в.ч. транзитивно)



Метод – 2

3. нормализация запросов (стоп-слова, капитализация, лемматизация, сортировка слов + выбор лучшего обратного преобразования)

4. борьба со флешмобами и «событийными» ассоциациями

5. матрица частоты переходов «запрос-запрос» (пороги для пользователя, ограничение на абсолютную частоту)

6. weight(q1 → q2)= $f(freq_{q1}, freq_{q2}, freq_{q1q2})*f(freq_{q2q1})$

7. ранжирование, отсечение по порогу

8. индекс: q → q1, q2, q3… (оригинальные запросы)

Оценка

- «оценка глазами»
- «классы объектов»
- кластеры запросов Яндекс.Директа
- кликабельность

Результат

«**манго**» — связанные запросы

mango | манго **одежда** | **авокадо**
папайя | манго **фрукт** | **инжир**
финик | **магазин** манго | **помело**

«**умные фразы**» — связанные запросы

умные **мысли** | умные **слова** | **смешные** фразы
умные **цитаты** | умные **статусы** | фразы **о жизни**
умные **вопросы** | фразы **о любви** | **мудрые** фразы

«**хочу похудеть**» — связанные запросы

хочу **любви** | хочу **умереть** | хочу **замуж**
хочу **ребенка** | хочу **денег**

«**смайлики**» — связанные запросы

смайлы для qip | **анимашки** | **smiles**
улыбка | **анимации** | **иконки**
позитив

«**люблю**» — связанные запросы

любимому **смс** | **скучаю** | **сердце**
поцелуй

«**утконос**» — связанные запросы

ехидна | **коала** | **кенгуру**
выхухоль | **эму** | **муравьед**
вомбат | **динго** | **броненосец**

«**лук**» — связанные запросы

рыцарь вампир | **лукашенко** | **мечи**
капуста | **морковь** | **свекла**
огурец | **картофель** | **арбалет**

«**ива**» — связанные запросы

дуб | **тополь** | **береза**
липа | **сосна** | **клен**
ель | **каштан** | **рябина**

АКАДЕМИЧЕСКИЕ ИНИЦИАТИВЫ

Академические инициативы

- Школа Анализа Данных
- Семинары Яндекса
- ИМАТ 2009, ИМАТ 2010
- РОМИП
- RuSSIR
- IIR book

Yandex School of Data Analysis

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

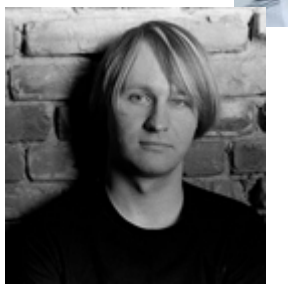
Teachers

VC-dimension (Vapnik-Chervonenkis)

Scientific seminars

Monthly seminars on Data analysis & information retrieval

Organized by
Microsoft Research + Яндекс



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

ИНТЕРНЕТ-МАТЕМАТИКА

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

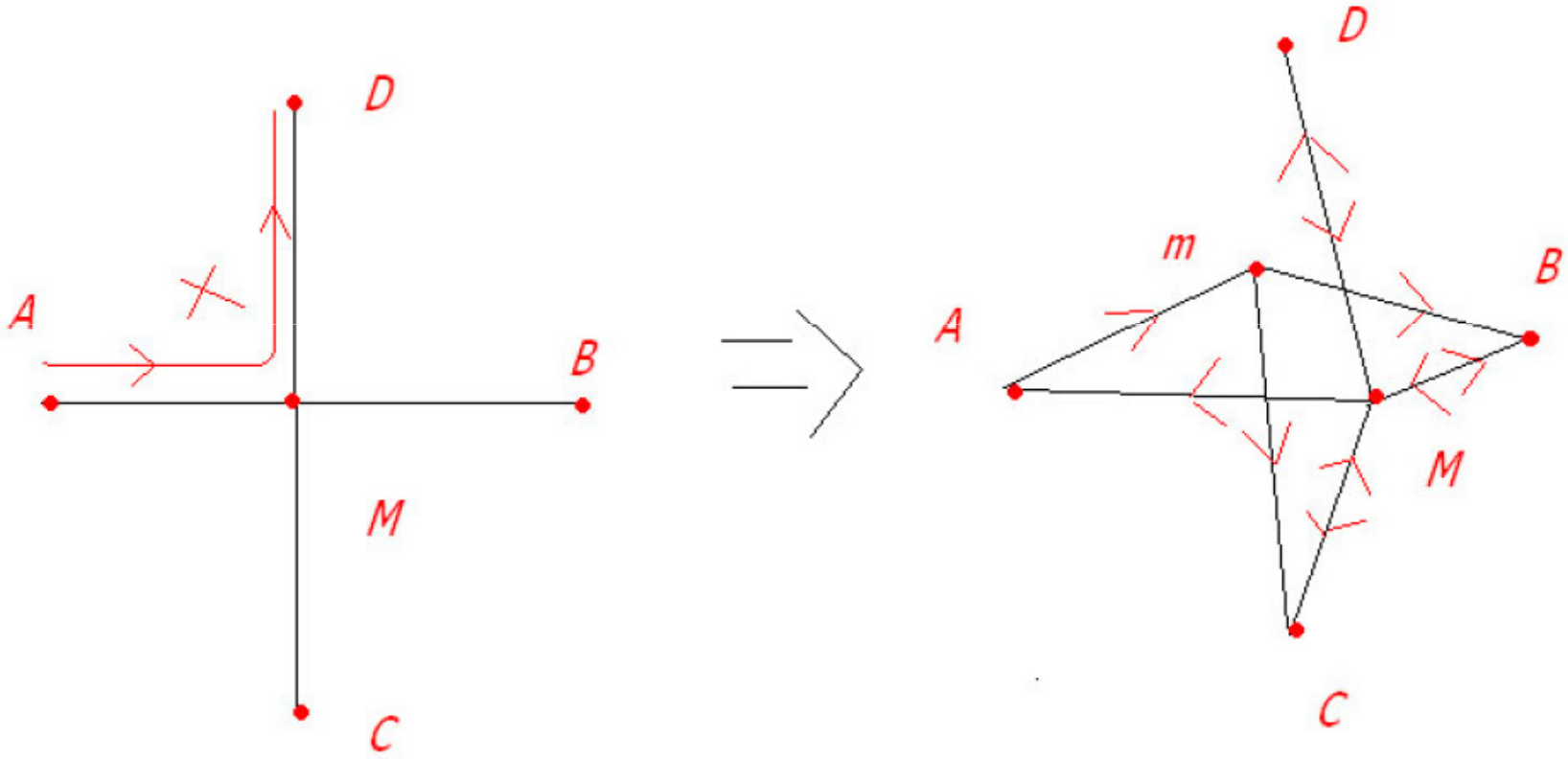
Таблица объединяет финальный рейтинг конкурса (на 15.05.2009) и новые результаты. Подробнее о задаче и методике оценки конкурсе и раздел Задачи и данные.

Команда	Время последней загрузки	Количество попыток	Последний результат (публичная оценка)	Финальный результат
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
Simple	21.02.2010 (21:18 GMT+03)	531	4.282597	4.135317
-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
depechemode	29.10.2009 (01:19 GMT+03)	30	4.278378	4.142855
Победа	17.03.2009 (16:25 GMT+03)	3	4.276001	4.139854
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
RelevanceDoesMatter	22.02.2010 (17:54 GMT+03)	378	4.272542	—
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
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
MysteriousGuest	18.01.2010 (16:00 GMT+03)	8	4.259078	4.134784
Euclid	08.05.2009 (21:46 GMT+03)	40	4.257802	4.122558

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

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

Данный рейтинг составлен на основе оценки небольшой части результатов, поданных участниками («публичная оценка»). Финальная оценка результатов будет проведена после 16 мая 2010 года. Подробнее о методике оценки см. Задача и данные. Правила участия описаны в Положении о конкурсе.

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)
test	05.05.2010 (15:09 GMT+03)	126	59.188577
View	05.05.2010 (20:01 GMT+03)	82	59.469988
SlowMotion	03.05.2010 (23:37 GMT+03)	60	59.712130
Уууй	06.05.2010 (10:26 GMT+03)	328	59.865502
Kejpa	13.04.2010 (06:40 GMT+03)	71	60.728057
kek_ksu	05.05.2010 (16:08 GMT+03)	11	60.758554
Сергей Гуда и Денис Рябов (ЮФУ)	06.05.2010 (14:49 GMT+03)	269	61.375202
Denis Ryabov	06.03.2010 (00:02 GMT+03)	6	61.487828
deas	28.04.2010 (14:28 GMT+03)	16	62.511518
Dimatsi	18.04.2010 (11:08 GMT+03)	79	62.832792
Паровоз	03.05.2010 (23:06 GMT+03)	83	63.094360
lex	05.05.2010 (12:35 GMT+03)	43	63.169360
Иван	03.05.2010 (14:18 GMT+03)	103	63.491430
timelimit	16.04.2010 (01:51 GMT+03)	67	63.593053
A1234	03.05.2010 (14:23 GMT+03)	36	63.679937
GoGo	05.05.2010 (17:29 GMT+03)	25	64.111546
metro	27.04.2010 (00:25 GMT+03)	15	64.125849
nihao	23.03.2010 (03:00 GMT+03)	22	64.170135

РОМИП

ROMIP

- TREC-like Russian initiative
- Started 2002
- Several data collections
- 15-20 participants every year
- Remote participation + live meeting
- ~3000 man-hours of evaluation (2009)

http://romip.ru

Tracks

- Ad-hoc text retrieval
  - Web collection, legal collection
- Text categorization
  - Web collection, legal collection
- Query-dependent summarization
- Image retrieval
  - by image query
  - near-duplicates detection
- News clustering
- QA
- Fact extraction from news collection

Text retrieval

collection	*size*	*tasks*	*pool depth*	*labeled*
Legal	*60 000*	*14 794*	*35*	*75 (220)*
By.Web	*1 524 676*	*29 231*	*20*	*550 (1060)*
KM.RU	*3 010 455*	*29 231*	*50*	*50 (170)*

Image retrieval

- query by image
  - Collection: 20 000 images from flikr
  - 2 000 tasks, 250 labeled
- Near-duplicate detection
  - Collection: 37 800 frames from video

RUSSIR

RuSSIR 2007

Yekaterinburg,
5-12 September 2007

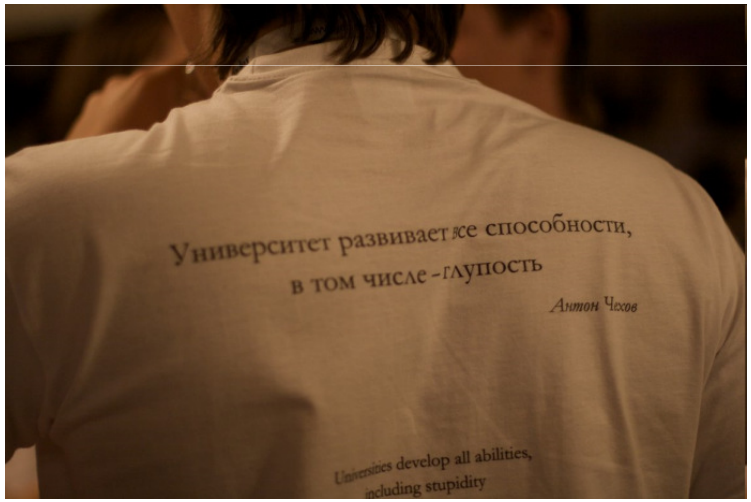
co-located with CSR 2007

~60 participants

Video+slides:
http://romip.ru/russir2007/schedule.html

RuSSIR 2008

- Taganrog, 1-5 September 2008
- 4 courses + a short course + a lecture
- Young scientists' conference
- ~100 participants
- All courses in English
- Materials:
  - http://videolectures.net/
  - http://romip.ru/russir2008/program.html



RuSSIR 2009

- Petrozavodsk, Karelia
- 4 full courses, 1 short course, 1 lecture
- Young scientists' conference
- All courses in English
- ~110 participants
- Co-located with ROMIP meeting
- Video lectures
  - http://intuit.ru/
  - http://videolectures.net
- Presentations
  - http://romip.ru/russir2009/program.html

RuSSIR 2010

- Voronezh, 13-18 September 2010
- Courses:
  - Web Data Mining - Ricardo Baeza-Yates, Yahoo! Research
  - Multimedia Information Retrieval - Stefan Ruger, The Open University
  - XML Information Retrieval - Mounia Lalmas, University of Glasgow
  - Graph-based Methods for Social Search - Alexander Troussov, IBM Ireland
  - Distributed Information Retrieval - Fabio Crestani & Ilya Markov, University of Lugano
- http://romip.ru/russir2010/

Информационный поиск по-русски

Оригинальная английская версия: http://informationretrieval.org

Литература

- Broder A. A Taxonomy of Web Search. SIGIR 2002.
- Risvik K. M. et al. Query Segmentation for Web Search. WWW2003.
- Bergsma S. & Wang Q. I. Learning Noun Phrase Query Segmentation. EMNLP-CoNLL 2007.
- Hagen M. et al. The Power of Naïve Query Segmentation. SIGIR 2010.
- Beitzel, S.M. et al. Temporal analysis of a very large topically categorized web query log, JASIST, vol. 58, no. 2, 2007.
- Beitzel, S.M., et al. Automatic classification of web queries using very large unlabeled query logs. ACM Trans. Inf. Syst., 25(2):9, 2007.
- Shen D. et al. Q2C@UST: Our Winning Solution to Query Classification in KDDCUP 2005, SIGKDD Explorations 7(2).
- Li X. et al. Learning Query Intent from Regularized Click Graphs. SIGIR 2008.

Ссылки

- История про лог AOL – см. http://en.wikipedia.org/wiki/AOL_search_data_scandal
- Query Log Analysis Workshop @ WWW2007, http://querylogs2007.webir.org/
- KDD Cup 2005, http://www.sigkdd.org/kdd2005/kddcup.html
- WSCD09: Workshop on Web Search Click Data 2009, http://research.microsoft.com/users/nickcr/wscd09/
- Jiang D. Web Search/Browse Log Mining: Challenges, Methods, and Applications, http://research.microsoft.com/en-us/people/djiang/web_search_and_browse_log_mining.pdf
- Демонстрация - определение пола по запросам: http://adlab.microsoft.com/Demographics-Prediction/

Яндекс

Павел Браславский
pb@yandex-team.ru