Анализ поисковых запросов

Павел Браславский

Санкт Петербург,
ноябрь-декабрь 2010

План на ноябрь

- Введение
- Характеристики потока запросов
- Данные для анализа
- Сегментация
- Тематическая классификация
- Близкие запросы
- Практические задания

ВВЕДЕНИЕ

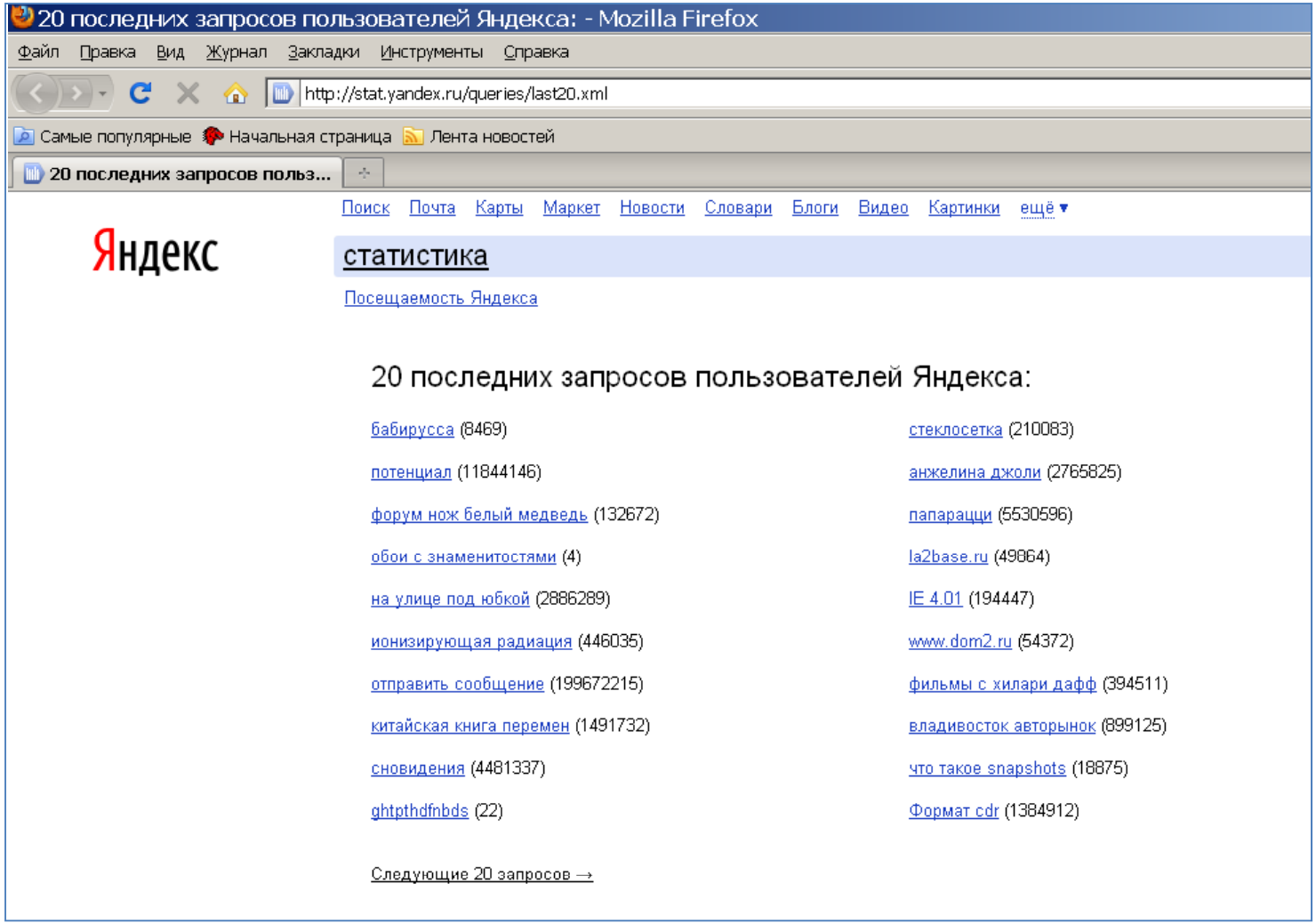
Предварительные замечания

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

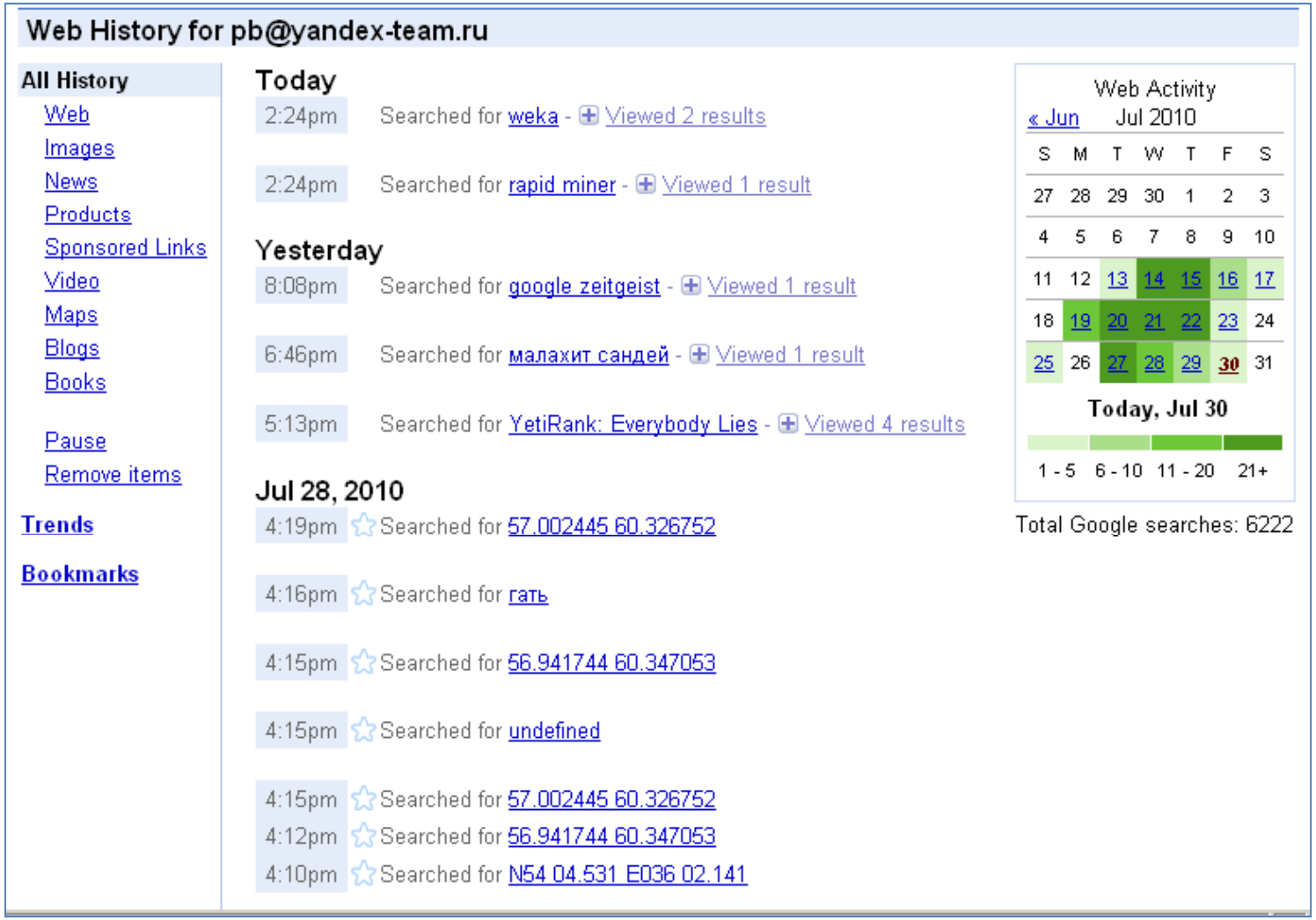
Статистика запросов к Яндексу

День ▲	Посетители	Хиты	Хитов с посетителя
11.11.2010	14 248 496	103 547 538	7,27
10.11.2010	15 506 565	122 113 251	7,87
09.11.2010	15 164 865	122 395 405	8,07
08.11.2010	15 198 131	134 154 973	8,83
07.11.2010	12 173 980	106 138 000	8,72
06.11.2010	11 434 498	95 857 719	8,38
05.11.2010	11 517 047	95 532 682	8,29
04.11.2010	11 647 878	95 698 951	8,22

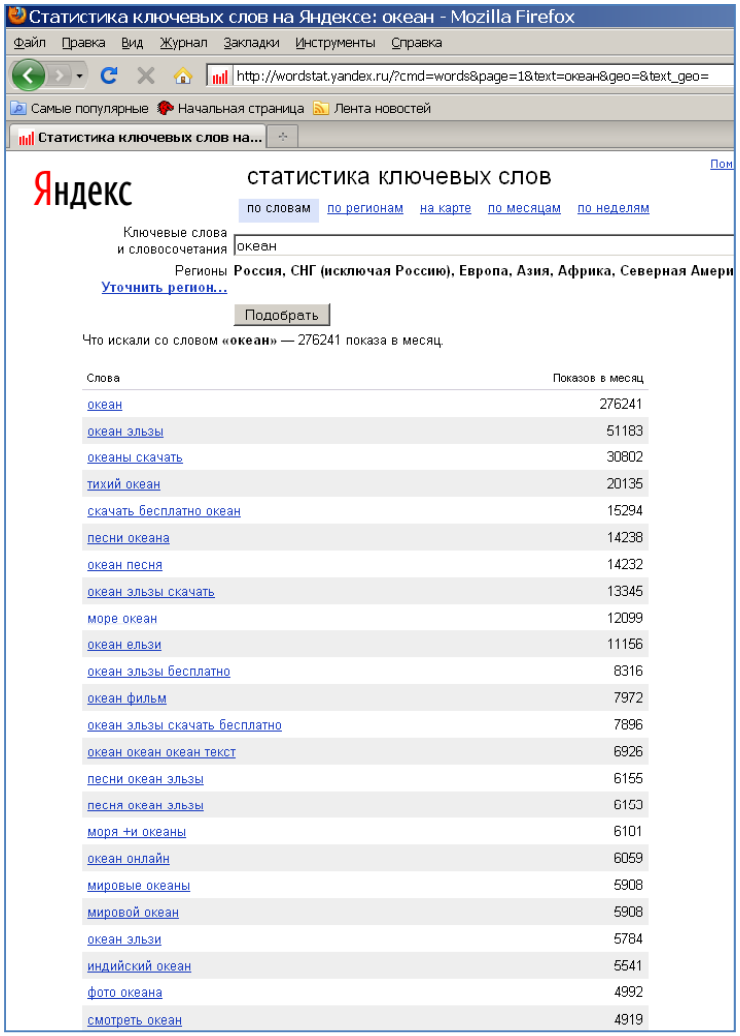
Прямой эфир

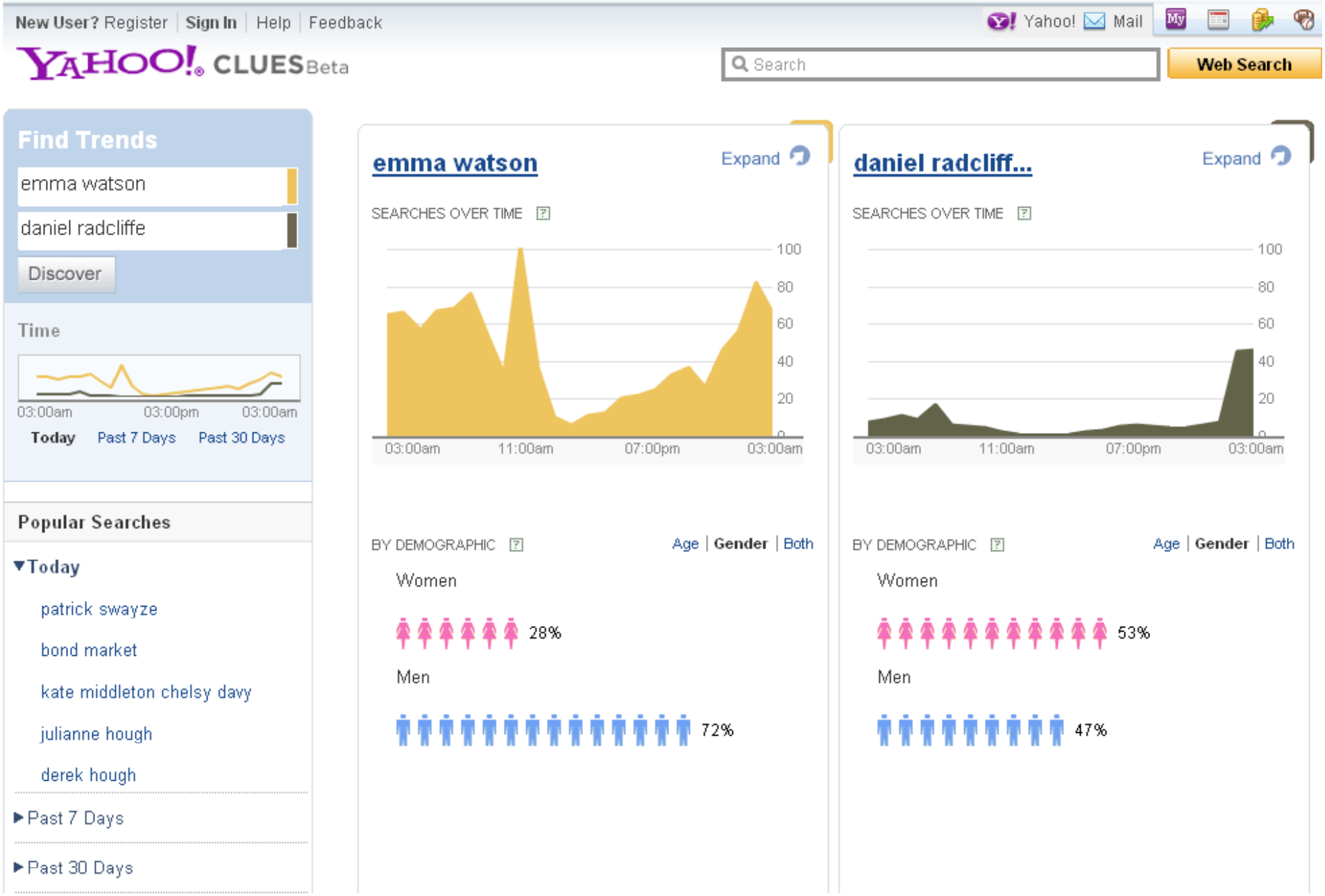

20 последних запросов пользователей Яндекса: - Mozilla Firefox
Файл Правка Вид Журнал Закладки Инструменты Справка
http://stat.yandex.ru/queries/last20.xml
Самые популярные Начальная страница Лента новостей
20 последних запросов польз...
Поиск Почта Карты Маркет Новости Словари Блоги Видео Картинки ещё
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




New User? Register | Sign In | Help | Feedback
Yahoo! Mail
YAHOO! CLUES Beta
Search
Web Search
Find Trends
emma watson
daniel radcliffe
Discover
Time
03:00am 03:00pm 03:00am
Today Past 7 Days Past 30 Days
Popular Searches
Today
patrick swayze
bond market
kate middleton chelsy davy
julianne hough
derek hough
Past 7 Days
Past 30 Days
emma watson
Expand
SEARCHES OVER TIME
100 80 60 40 20 0
03:00am 11:00am 07:00pm 03:00am
BY DEMOGRAPHIC
Age | Gender | Both
Women
28%
Men
72%
daniel radcliff...
Expand
SEARCHES OVER TIME
100 80 60 40 20 0
03:00am 11:00am 07:00pm 03:00am
BY DEMOGRAPHIC
Age | Gender | Both
Women
53%
Men
47%

New User? Register | Sign In | Help | Feedback

Yahoo! Mail

YAHOO! CLUES Beta

Search

Web Search

Find Trends

snowboard

ski

Discover

Time

Today Past 7 Days Past 30 Days

Popular Searches

▼**Today**

- patrick swayze
- bond market
- kate middleton chelsy davy
- julianne hough
- derek hough

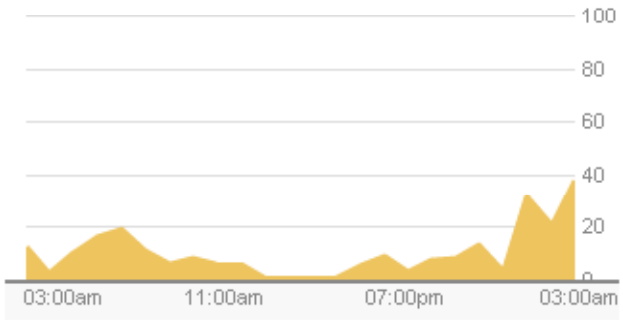
▶Past 7 Days

▶Past 30 Days

snowboard

Expand

SEARCHES OVER TIME



BY DEMOGRAPHIC **Age** | Gender | Both

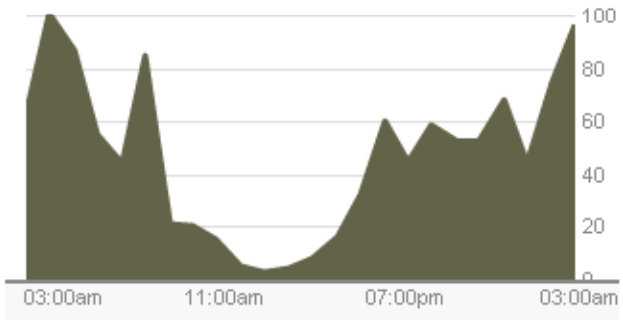
Age

Below 24	21%
25 to 34	56%
35 to 44	13%
45 to 54	8%
55 to 64	2%
Above 65	0%

ski

Expand

SEARCHES OVER TIME

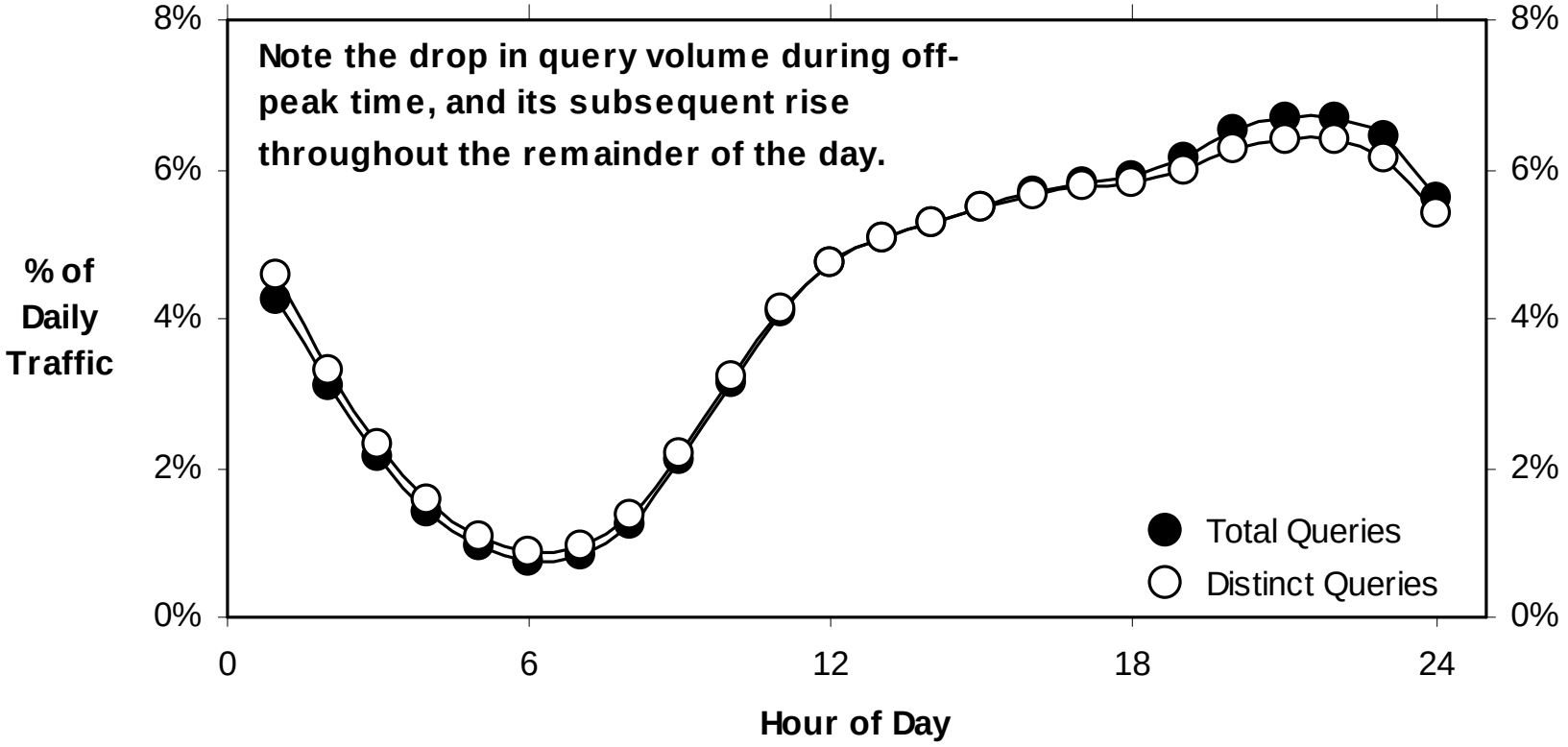


BY DEMOGRAPHIC **Age** | Gender | Both

Age

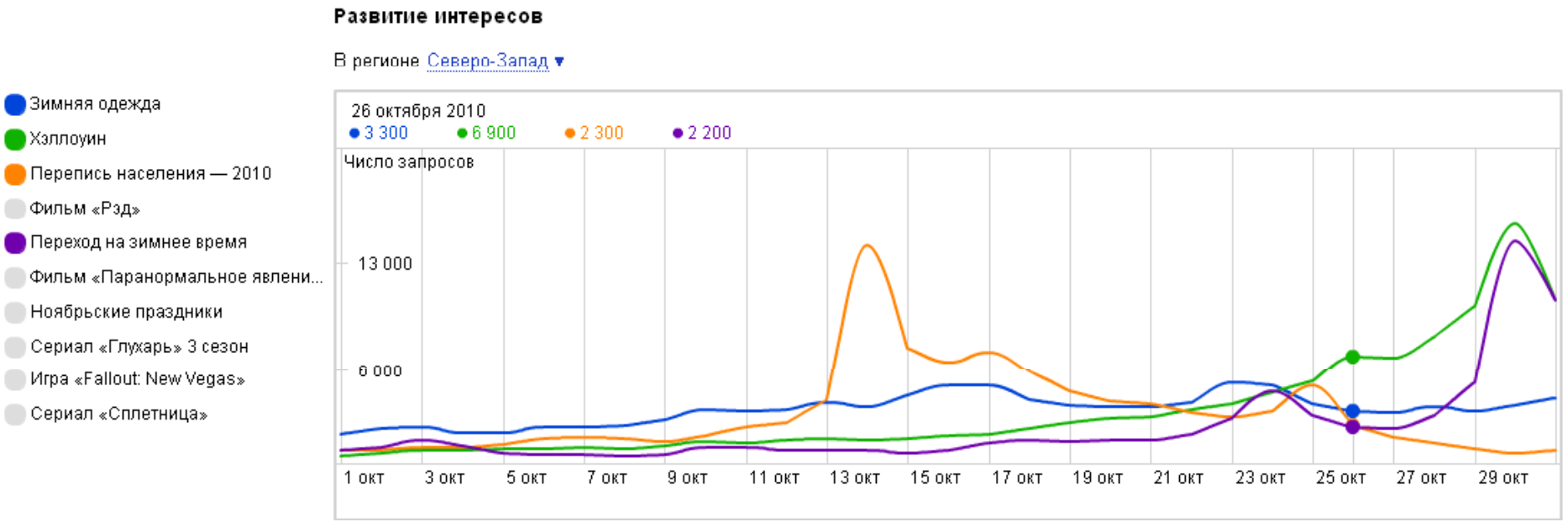
Below 24	8%
25 to 34	25%
35 to 44	23%
45 to 54	24%
55 to 64	13%
Above 65	5%

Traffic Volume Over a Day



[Beitzel]

Динамика запросов

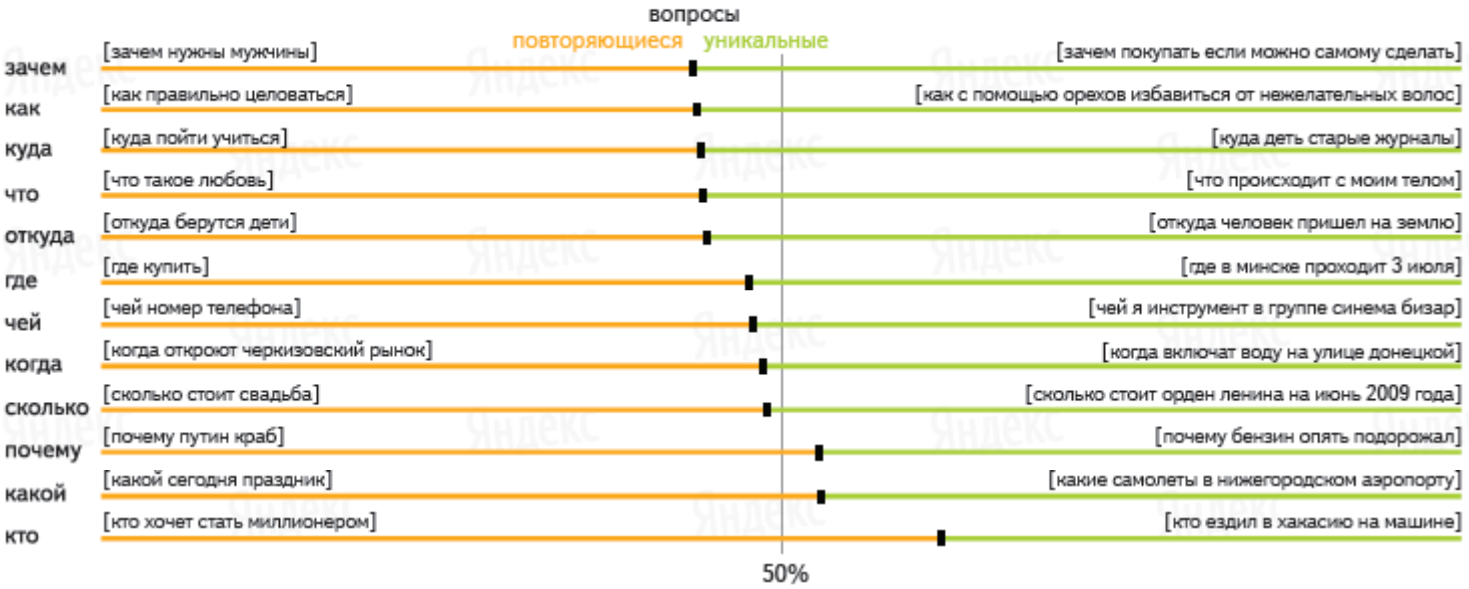


http://interes.yandex.ru/

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`

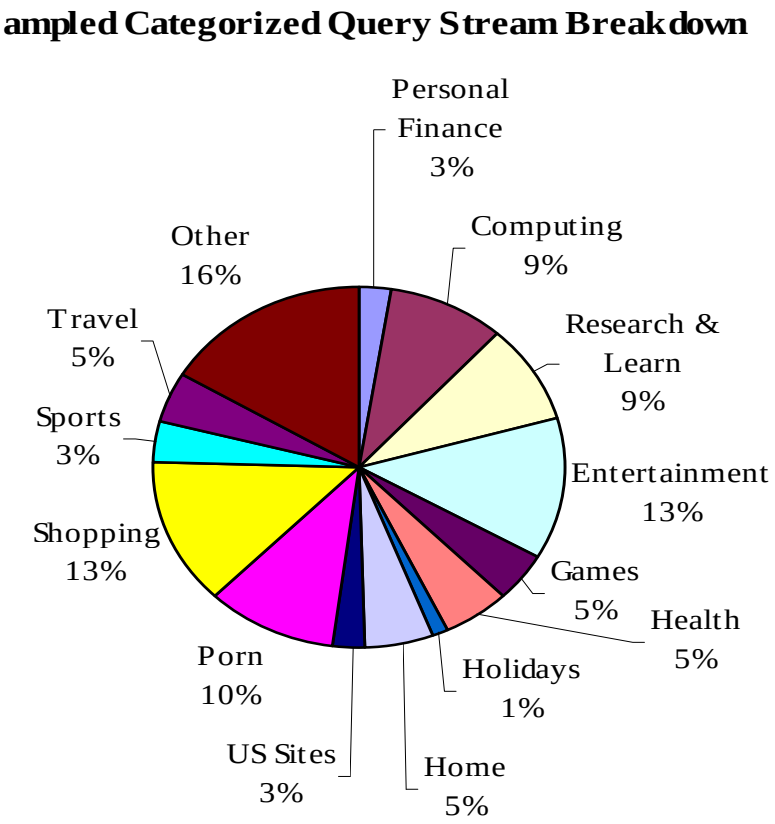
- **Transactional** – want to do something (web-mediated) (~35% / 20%)
  - Access a service — `Barcelona weather`
  - Downloads — `Mars surface images`
  - Shop — `Canon S410`

- Gray areas
  - Find a good hub — `Car rental Poland`
  - Exploratory search "see what's there"

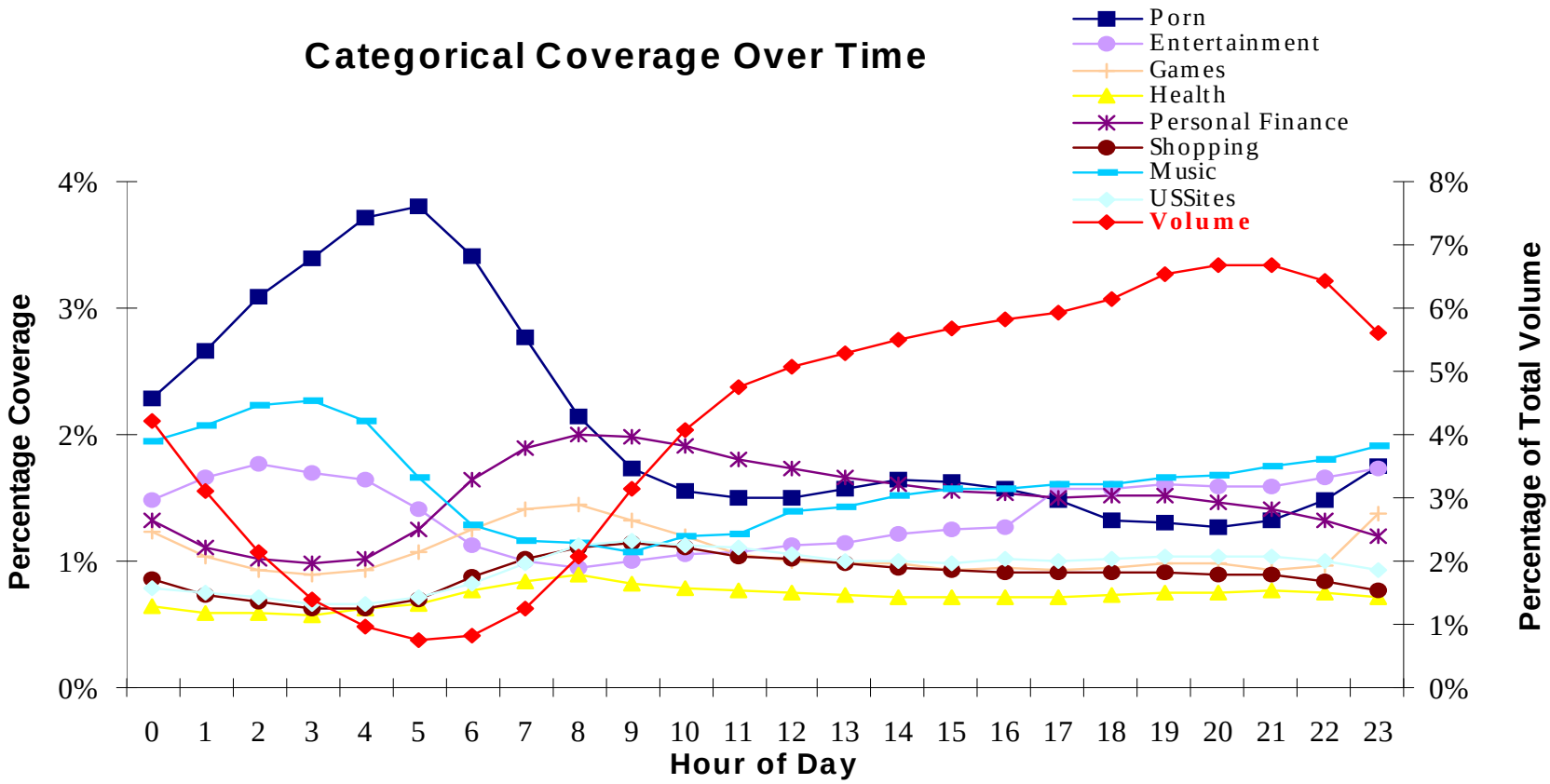
Baeza-Yates

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

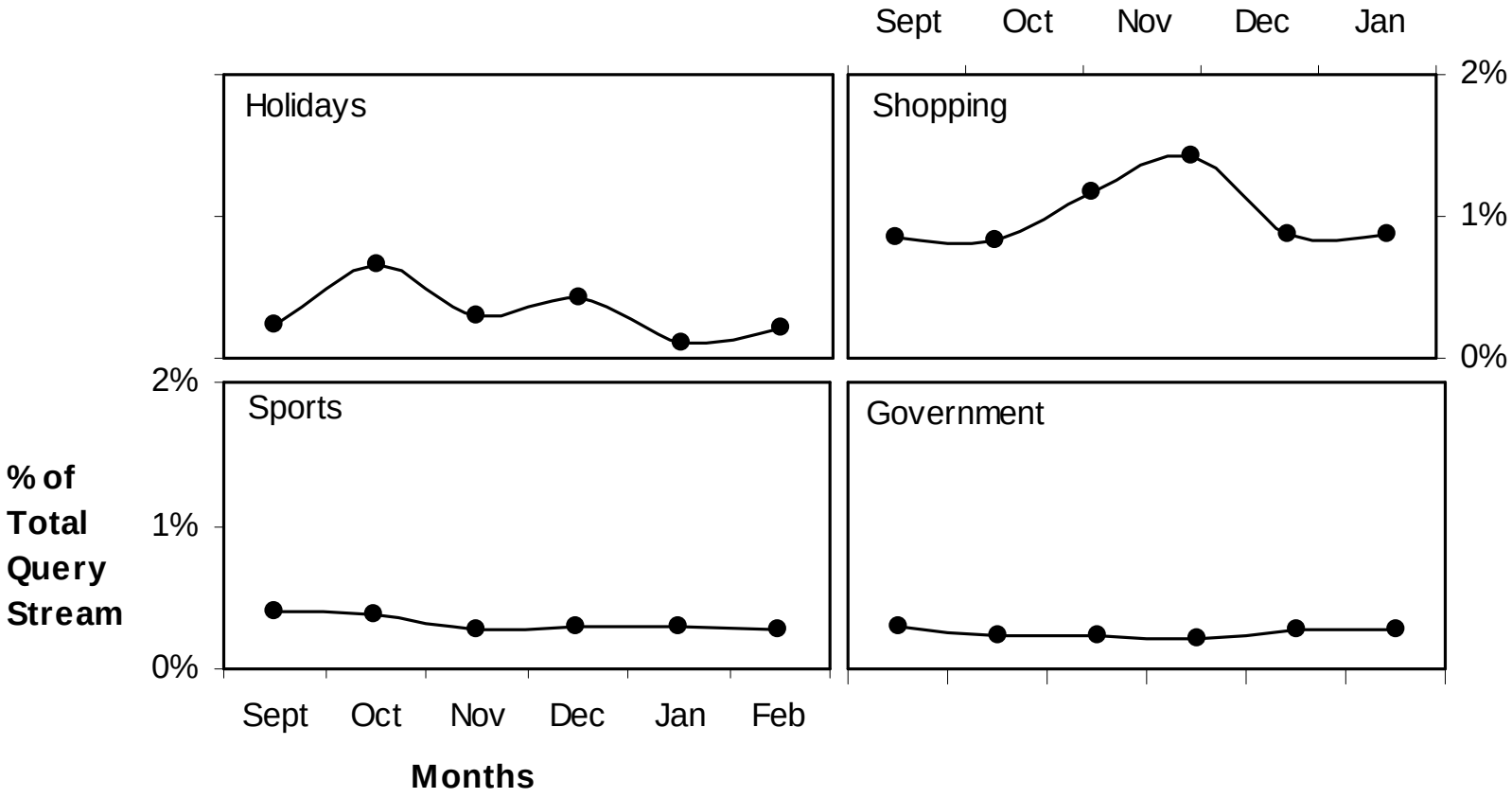


Category Popularity Over a Day



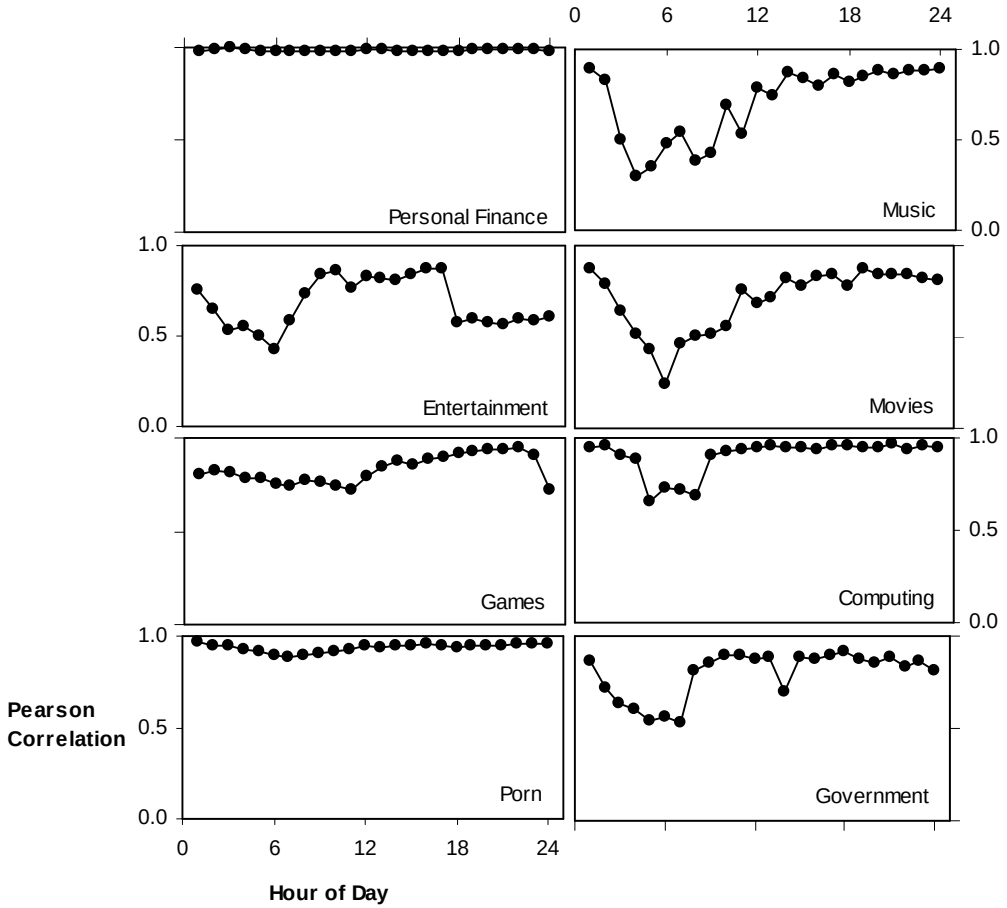
[Beitzel]

Category Popularity Over Six Months



[Beitzel]

Pearson Correlations for Selected Categories Over A Day



[Beitzel]

Источники и типы данных

Jiang et al. 2010

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: подходы

- похоже на выделение устойчивых словосочетаний
- + микросинтаксис
- лог vs корпус текстов
- ML (больше признаков, более богатое описание)
- внешние ресурсы (Wikipedia)
- эвристики ([дима билан]←[димабилан])

Сегментация на основе лога

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

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

```
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