Анализ поисковых запросов. Часть 2

Павел Браславский

Санкт Петербург,
ноябрь-декабрь 2010

Я

План

- Повторение пройденного
- Обсуждение практики
- Пространственные характеристики запросов
- ВременнЫе характеристики запросов
- Перевод запросов

+ Академические программы Яндекса

ГЕОГРАФИЯ В ЗАПРОСАХ

Почему география?

- 1990-е: эйфория от отсутствия границ; свобода перемещения в информационном пространстве
- 2000-е: повседневная жизнь и потребности – в Вебе

Результат

- «Поиск на местности»: картографические сервисы как часть поисковых порталов
- местоположение пользователя, локальные аспекты ресурса и запроса → вклад в релевантность

География ресурса

- Расположение провайдера (владельца) веб-ресурса
- Локализация контента (чему посвящено)
- География предоставления сервиса

→ Разные методы для автоматического определения разных типов привязок

Wang et al., 2005

Данные для анализа

1. IP
2. URL
3. Содержание документов
4. Структура ссылок
5. Поведение пользователей

Внешние ресурсы

1. Газеттир (gazetteer)
2. Справочники (телефонные коды, почтовые индексы, население, …)
3. Интернет-каталоги
4. «Желтые страницы»

Положение владельца ресурса

1. Извлечение адресных блоков по шаблонам
2. Отделение адресов владельцев ресурса от прочих адресов (SVM, признаки: повторяемость на разных страницах, уровень URL, заголовок страницы, текст ссылки, положение на странице)

[Wang et al., 2005]

Локализация контента

- Система *Web-a-Where*: географическая привязка отдельных страниц на основании их содержания
- **Мировой** газеттир (40 000 мест, 75 000 имен)
- Разрешение двух типов неоднозначности: geo/non-geo, geo/geo
- До 4 географических «фокусов» страницы
- Тестирование: ODP/Regional
- точность: 92% на уровне страны, 38% - на уровне города

[Amitay et al., 2004]

География ресурса

1. Основан на парсировании регистрационных записей доменных имен + запросы к МП
2. Газеттир: IP, доменное имя, город, индекс, тел. код, географические координаты
3. Макетное приложение: страницы в домене *.edu*
4. Запрос к Google [link:URL site:edu]
5. Визуализация

[Buyukkokten et al., 1999]

Типы запросов

- локализуемые: ожидается ответ, «близкий» пользователю ~15%(?)
  купить холодильник, химчистка, пицца
- локально-специфичные (связаны с определенным местом)
  гостиницы Новосибирск, кинофестиваль кинотавр, салават юлаев

NB: присутствие географических имен: *банк москвы, париж техас, небо над берлином, мост ватерлоо*

Локализуемые запросы

- классификация локальных запросов на основе выдачи МП: признаки на основе присутствия геоимен в документах (Gravano et al, 2003)
- Отношение запросов с/без гео модификаторами; разнообразие гео
- Клики на «локальных» результатах; отношение CTR на запросах с/без гео
- частота, количество пользователей (Vadrevu et al, 2008; Welch&Cho, 2008)

Локально-специфичные

- Query dominant location (QDL) – ассоциируется ли с запросом место:
  - геоимя – отдельный сегмент? (на основе анализа сниппетов топа, ср. *New York Times*)
  - анализ лога: IP пользователей и клики на документы
  - анализ топа на присутствие геоимен

  (Wang et al, 2005)

Локально-специфичные – 2

- Географическое распределение интереса к теме (запросу)
- Модель: у запроса есть «центр», константа C (частота в центре) и коэффициент падения интереса по мере удаления от центра α, частота: $Cd^{-\alpha}$
- Типы запросов: имена больших городов, университетов, газет, бейсбольных команд, национальных парков, фамилии сенаторов, а также региональные компании
- данные: лог Yahoo! и база «IP – гео координаты»

(Backstrom et al.,2008)

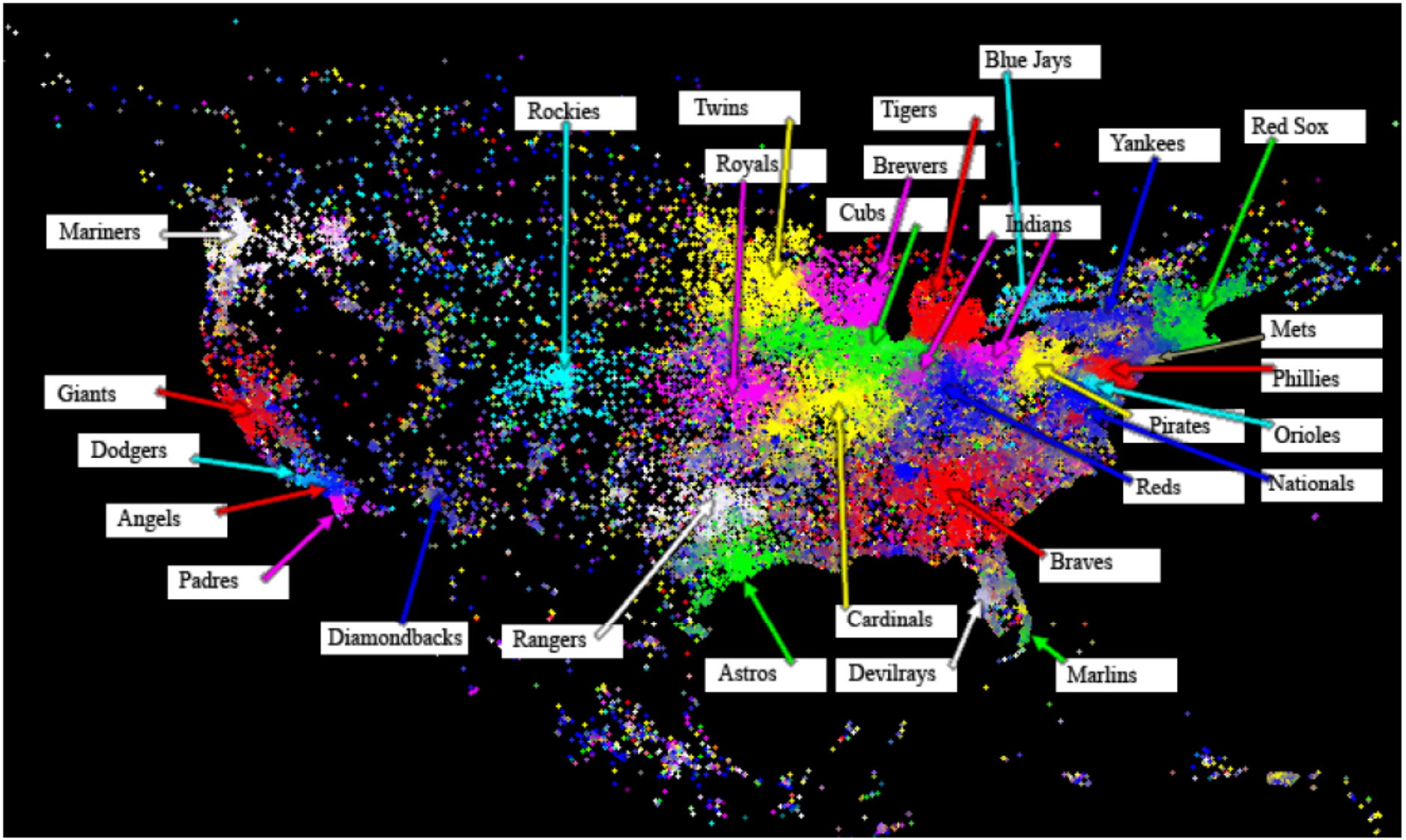


Figure 8: Spheres of influence of baseball teams.

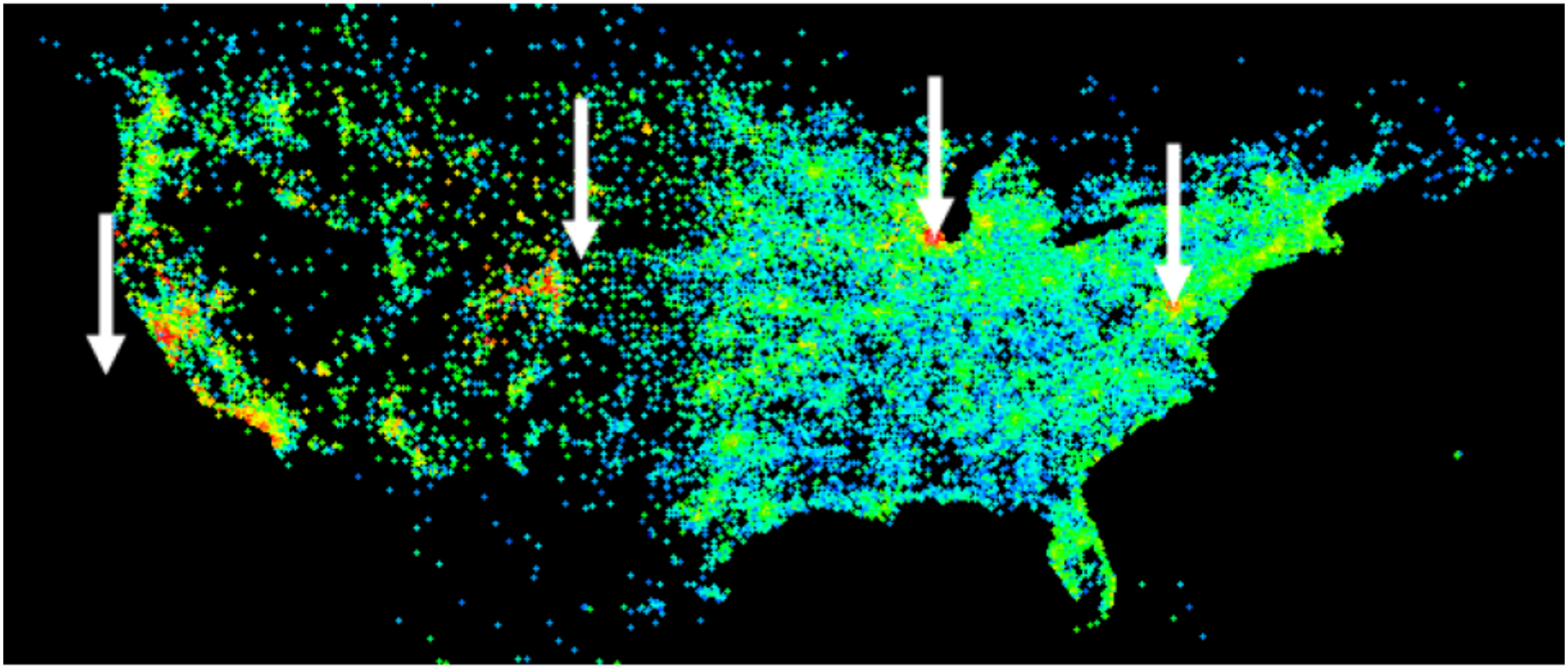
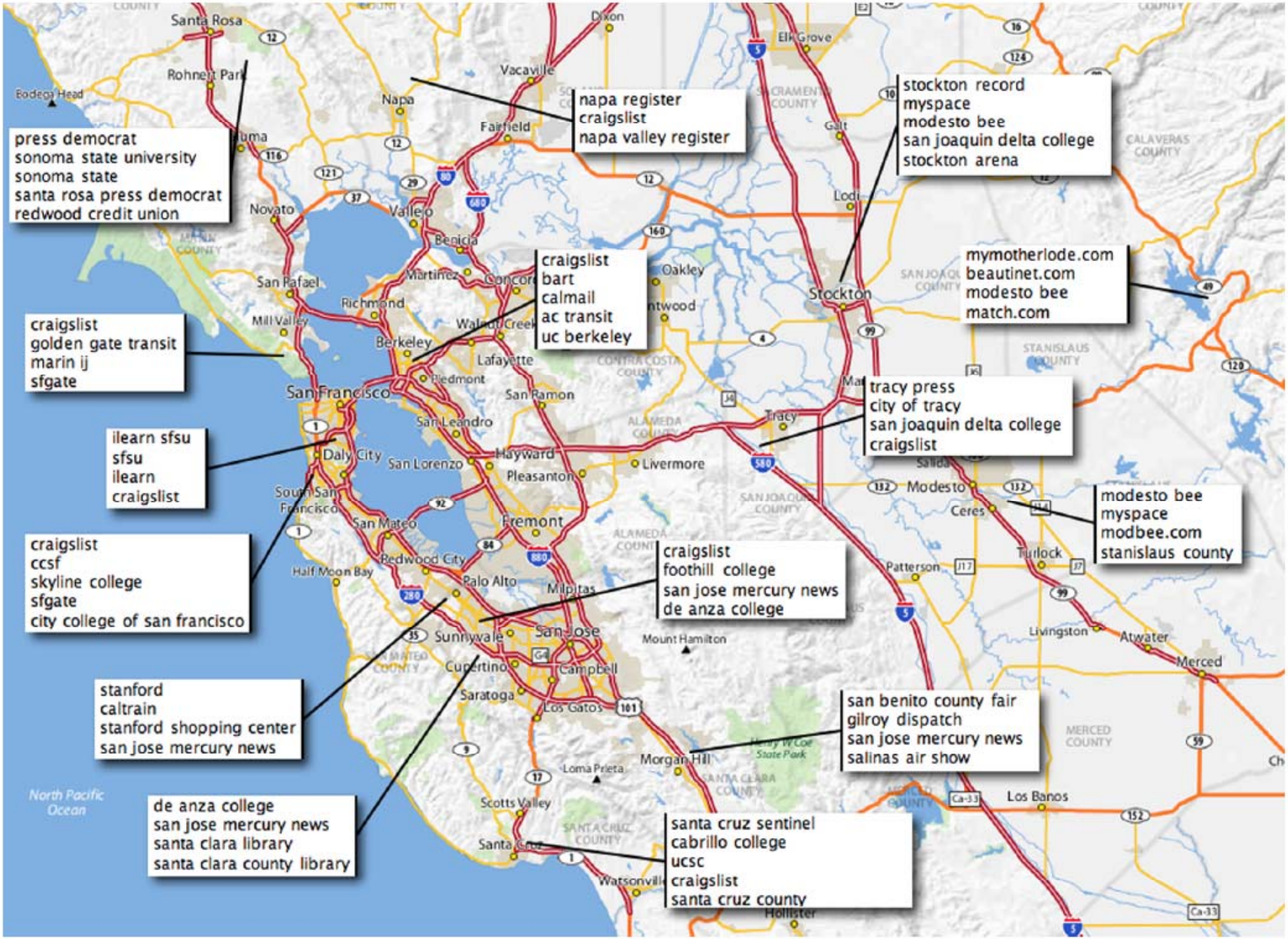
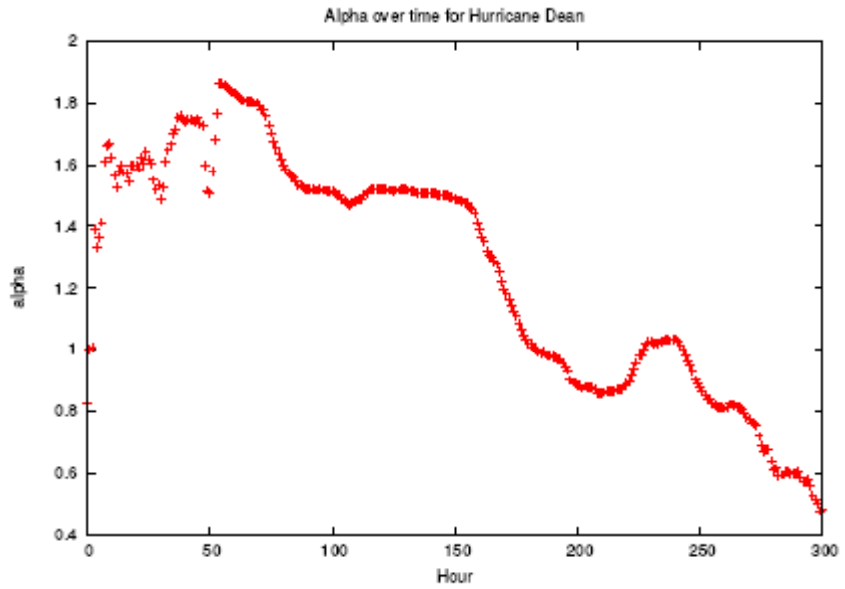


Figure 6: Multiple centers for the query “United Airlines.”


press democrat
sonoma state university
sonoma state
santa rosa press democrat
redwood credit union
napa register
craigslist
napa valley register
stockton record
myspace
modesto bee
san joaquin delta college
stockton arena
craigslist
bart
calmail
ac transit
uc berkeley
mymotherlode.com
beautinet.com
modesto bee
match.com
craigslist
golden gate transit
marin ij
sfgate
tracy press
city of tracy
san joaquin delta college
craigslist
ilearn sfsu
sfsu
ilearn
craigslist
modesto bee
myspace
modbee.com
stanislaus county
craigslist
ccsf
skyline college
sfgate
city college of san francisco
craigslist
foothill college
san jose mercury news
de anza college
stanford
caltrain
stanford shopping center
san jose mercury news
san benito county fair
gilroy dispatch
san jose mercury news
salinas air show
de anza college
san jose mercury news
santa clara library
santa clara county library
santa cruz sentinel
cabrillo college
ucsc
craigslist
santa cruz county


Smoothed
query-center path
Actual path of
storm's center


Alpha over time for Hurricane Dean
alpha
Hour
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0
50
100
150
200
250
300

GeoCLEF

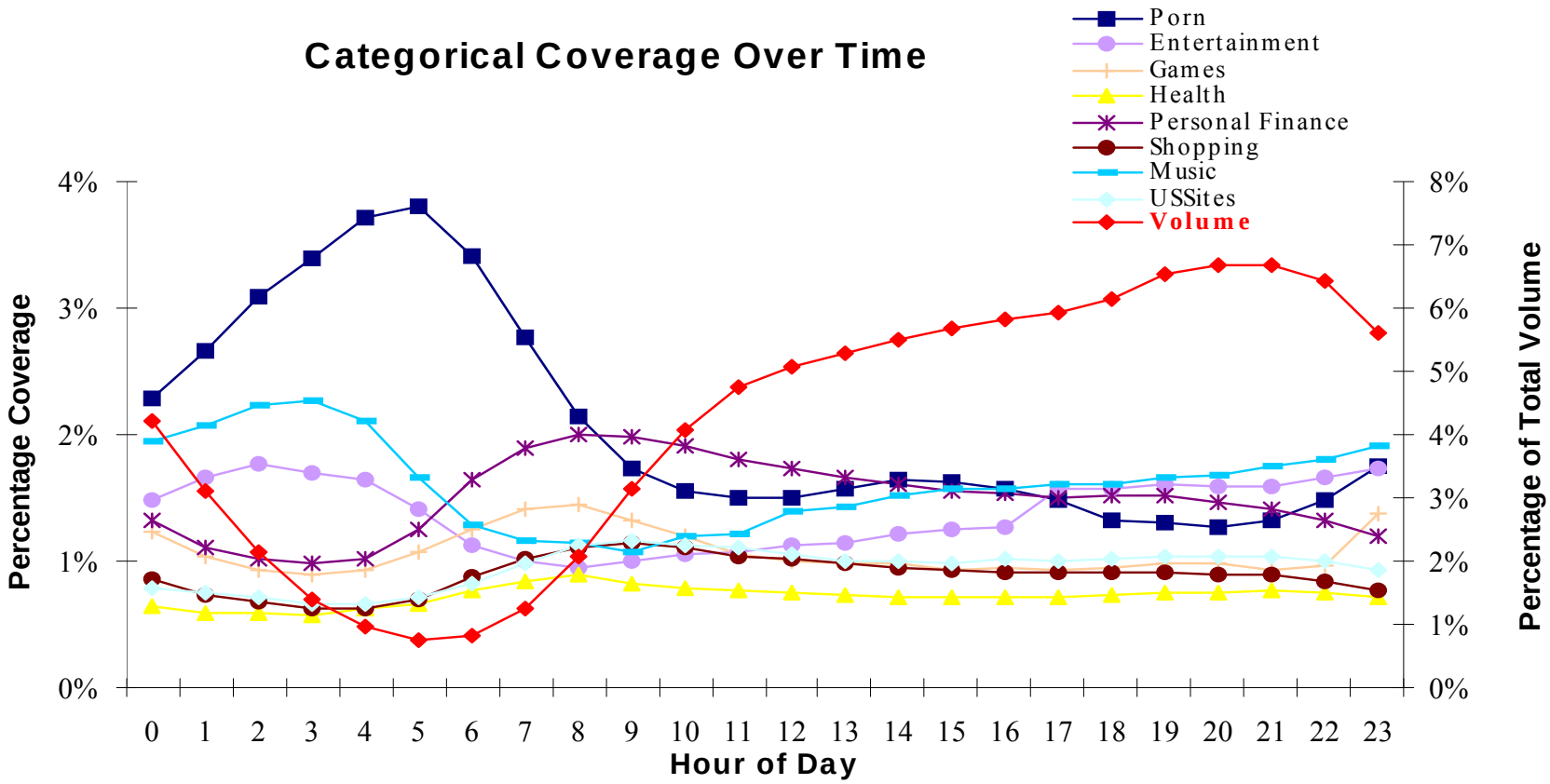
- извлечение нечеткой географической информации из текстов, моделирование различных сценариев поиска с учетом географии
- анализ запросов, содержащий географические аспекты
- поиск по Википедии с учетом географической информации
- поиск изображений

GeoCLEF: query parsing

- выделить запросы с географической составляющей
- исходя из структуры запроса *“what” + “geo-relation” + “where”*,
  - выделить *where*
  - *relation* (IN, ON, AT, NEAR,…)
  - определить тип *what* (map, yellow pages, information)

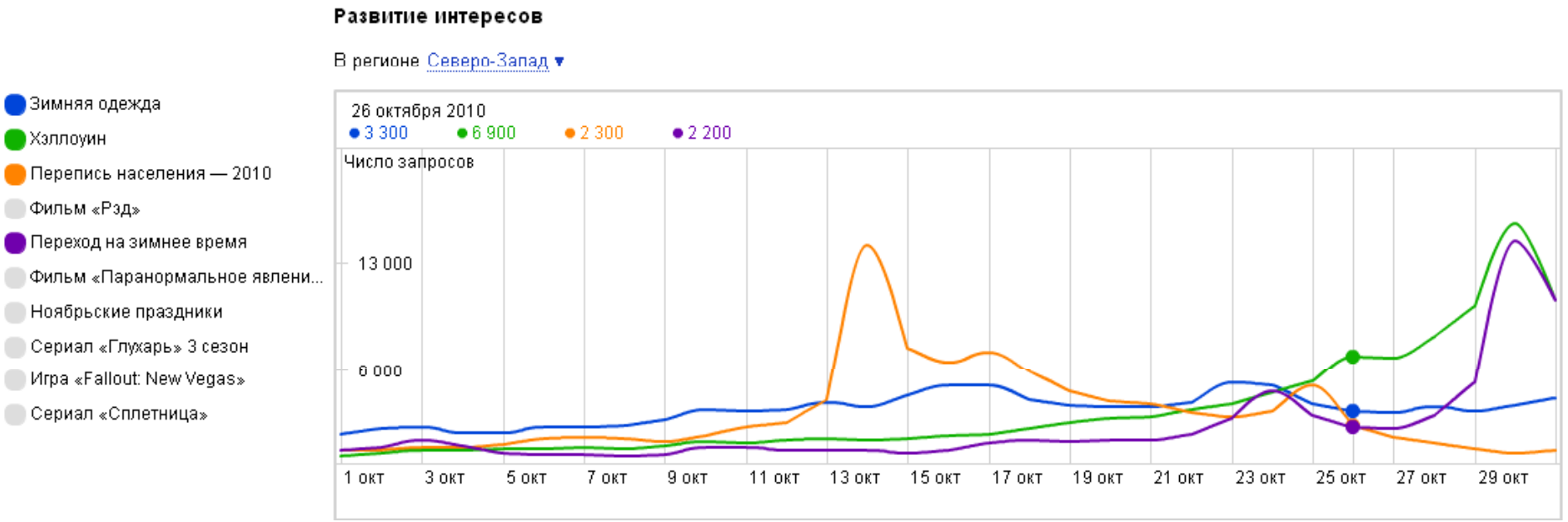
ВРЕМЕННЫЕ АСПЕКТЫ ЗАПРОСОВ

Category Popularity Over a Day

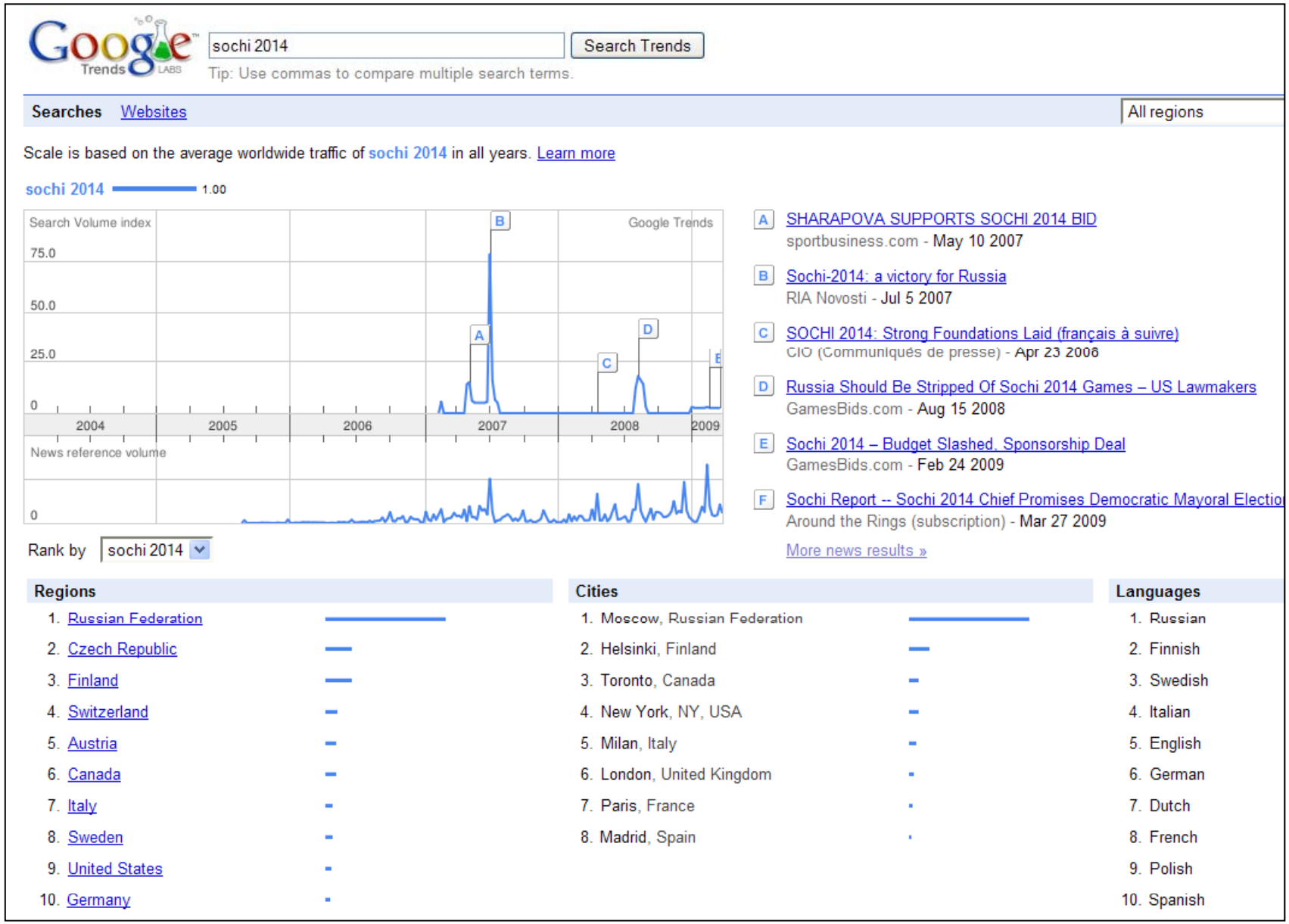


[Beitzel]

Динамика запросов



http://interes.yandex.ru/


Google Trends
sochi 2014
Search Trends
Tip: Use commas to compare multiple search terms.
Searches
Websites
All regions
Scale is based on the average worldwide traffic of sochi 2014 in all years. Learn more
sochi 2014 1.00
Search Volume index
75.0
50.0
25.0
0
2004
2005
2006
2007
2008
2009
News reference volume
0
Google Trends
A SHARAPOVA SUPPORTS SOCHI 2014 BID
sportbusiness.com - May 10 2007
B Sochi-2014: a victory for Russia
RIA Novosti - Jul 5 2007
C SOCHI 2014: Strong Foundations Laid (français à suivre)
CIO (Communiqués de presse) - Apr 23 2008
D Russia Should Be Stripped Of Sochi 2014 Games – US Lawmakers
GamesBids.com - Aug 15 2008
E Sochi 2014 – Budget Slashed, Sponsorship Deal
GamesBids.com - Feb 24 2009
F Sochi Report -- Sochi 2014 Chief Promises Democratic Mayoral Electio
Around the Rings (subscription) - Mar 27 2009
More news results »
Rank by
sochi 2014
Regions
1. Russian Federation
2. Czech Republic
3. Finland
4. Switzerland
5. Austria
6. Canada
7. Italy
8. Sweden
9. United States
10. Germany
Cities
1. Moscow, Russian Federation
2. Helsinki, Finland
3. Toronto, Canada
4. New York, NY, USA
5. Milan, Italy
6. London, United Kingdom
7. Paris, France
8. Madrid, Spain
Languages
1. Russian
2. Finnish
3. Swedish
4. Italian
5. English
6. German
7. Dutch
8. French
9. Polish
10. Spanish

Новостные запросы (Maslov et al., 2006)

- Идея: выделение запросов, относящихся недавним, текущим или близким событиям реального мира, находящим отражение в новостях: *новостные запросы*
- поиск: context transfer, query routing
- использование при обработке новостного потока: кластеризация, реферирование, ранжирование
- ср.: Henzinger, M. et al. Query-Free News Search, WWW2003.

Новостные запросы - 2

Query significance:

$$S(q, \Delta_1, \Delta_2) = \frac{F(q, \Delta_1)}{F(q, \Delta_2)}$$

Momentary query novelty:

$$MQN(q) = S(q,\ \Delta last\_int,\ \Delta prec\_day)$$

Hourly query novelty:

$$HQN(q) = S(q, \Delta last\_int,\ \Delta prec\_week)$$

Query novelty:

$$QN(q) = min\{MQN(q), HQN(q)\}$$

News-related queries:

- queries with more than 0.1% of relevant documents web database are removed
- there are relevant news within the three-hour time window around the query timestamp

Новостные запросы - 3

Typical daily breakdown:
25366048 Web queries → 196579 novel queries (0,77%) → 1039 news-related queries (0,53%).

Примеры:
пресс-конференция путина
пресс-конференция в кремле
компьютерный вирус 3 февраля
вирус пухет
горбатая гора энга ли
номинанты на оскар
бедствие в таиланде
число жертв цунами
землетрясение в юго-восточной азии
цунами 26 декабря
высота волн цунами
код да винчи в канне
ден браун код да винчи

Новостные запросы - 4

Test sample:
• four one-hour intervals between 10 am and 7 pm in two consequent workdays in December 2005;
• all queries automatically detected as news-related plus randomly selected 2% of the remaining queries within the respective intervals.

831 queries, 244 (30%) of which were automatically detected as news-related.

The test sample was presented to an assessor who evaluated queries in sequence. The assessor answered the question: “Is it safe to presume that the vast majority of the users making the query at the given time were interested in current news?” There are two values for each hour: plain (calculated over unique queries) and frequency-weighted (query frequency is accounted).

	12-7-2005: 1 pm		12-7-2005: 6 pm		12-8-2005: 10 am		12-8-2005: 3 pm	
	plain	wghtd	plain	wghtd	plain	wghtd	plain	wghtd
Misses	9*50	38*50	7*50	60*50	7*50	20*50	5*50	11*50
TruePos	122	923	130	808	101	755	145	710
FalsePos	30	75	25	64	12	32	27	71
Precision	0.80	0.92	0.84	0.93	0.89	0.96	0.84	0.91
Recall	0.21	0.33	0.27	0.21	0.22	0.43	0.37	0.56
F1	0.34	0.48	0.41	0.35	0.35	0.59	0.51	0.70

Burst Query Identification Using Language Model [Dong10]

- Calculating probabilities for generating query at current time slot $P(q \mid M_{C,t})$ $P(q \mid M_{Q,t})$
- Calculating probabilities for generating query from previous time slot to current time slot

  $P(q \mid M_{C,t-r_i})$ $P(q \mid M_{Q,t-r_i})$

- Calculating buzziness of query from two language models and linearly combining them

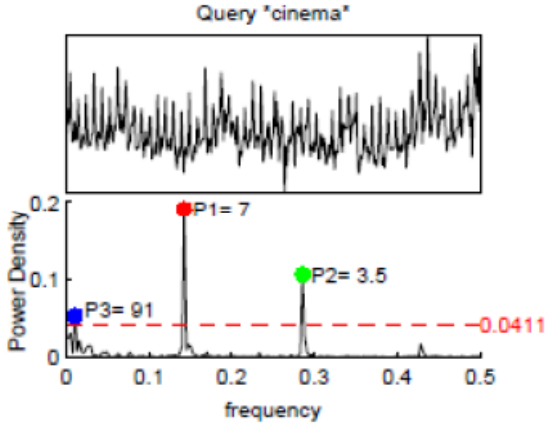
$$\mathrm{buzz}(q,t,C) = \max_i \log P(q \mid M_{C,t}) - \log P(q \mid M_{C,t-r_i})$$

$$\mathrm{buzz}(q,t,Q) = \max_i \log P(q \mid M_{Q,t}) - \log P(q \mid M_{Q,t-r_i})$$

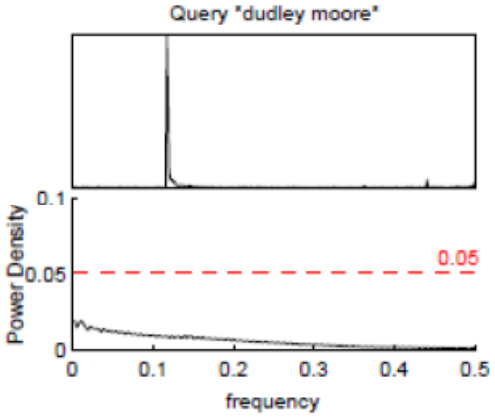
[Jing et al., 2010]

Periodic Queries vs Non-Periodic Queries

- Time periods can be found from power spectrum (period = 1 / frequency)
- Power spectrums of random (non-periodic) queries follow exponential distribution
- Hypothesis testing: find significant periods of query using power spectrum of query



Most significant period for query “cinema” is 7



No significant period found for query “dudley moore”

[Vlachos 04]
Слайд [Jing et al., 2010]

ПЕРЕВОД ЗАПРОСОВ

Подходы к CLIR

- Язык запроса (QL) != язык документов (DL)
  - поиск нетекстового содержания (!)
- Подходы:
  1. перевод запросов
  2. перевод документов
  3. 1 + 2

Ресурсы для перевода

- словари
- статистические словари на основе параллельных корпусов
- тезаурусы, в т.ч. [illegible]нные
- Wikipedia (!)
- логи запросов

примеры параллельных текстов?

построить тезаурус автоматически?

почему лучше словарей?

Ссылки – География

- Amitay E. et al. Web-a-Where: Geotagging Web Content, SIGIR'2004.
- Ding J., Gravano L., Shivakumar N.: Computing Geographical Scopes of Web Resources, VLDB2000.
- Chuang Wang et al. Web Resource Geographic Location Classification and Detection, WWW2005.
- Агеев М. и др. Некоторые способы определения географической привязки IP адресов, Интернет-математика, 2005.
- Pyalling A., Maslov M., Braslavski P. Automatic geotagging of Russian web sites, WWW2006.
- Lars Backstrom, Jon Kleinberg, Ravi Kumar, Jasmine Novak. Spatial Variation in Search Engine Queries. WWW 2008.
- Michael J. Welch, Junghoo Cho. Automatically Identifying Localizable Queries. SIGIR’08.
- Srinivas Vadrevu, Ya Zhang, Belle Tseng, Gordon Sun, Xin Li. Identifying Regional Sensitive Queries in Web Search. WWW 2008.
- Buyukkokten O., Cho J., Garcia-Molina J. Exploiting Geographical Location Information of Web Pages, SIGMOD'99.

Ссылки – Время

- Maslov M., Golovko A., Segalovich I., Braslavski P. Extracting news-related queries from web query log. WWW '06, 931-932.
- Henzinger, M. et al. Query-Free News Search, WWW2003, 1-10.
- Dong, A. et al. Towards Recency Ranking in Web Search. WSDM’10.
- Vlachos, M., et al. Identifying similarities, periodicities and bursts for online search queries. SIGMOD'04

Ссылки - Перевод

- Wang J., Oard D.W. Combining Bidirectional Translation and Synonymy for Cross-Language Information Retrieval, SIGIR'06.
- Nguyen, D. et al. WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia. Evaluating Systems for Multilingual and Multimodal Information Access, 2009.
- Hu, R. et al. Web Query Translation via Web Log Mining, SIGIR’08.

Павел Браславский
pb@yandex-team.ru