

Анализ поисковых запросов

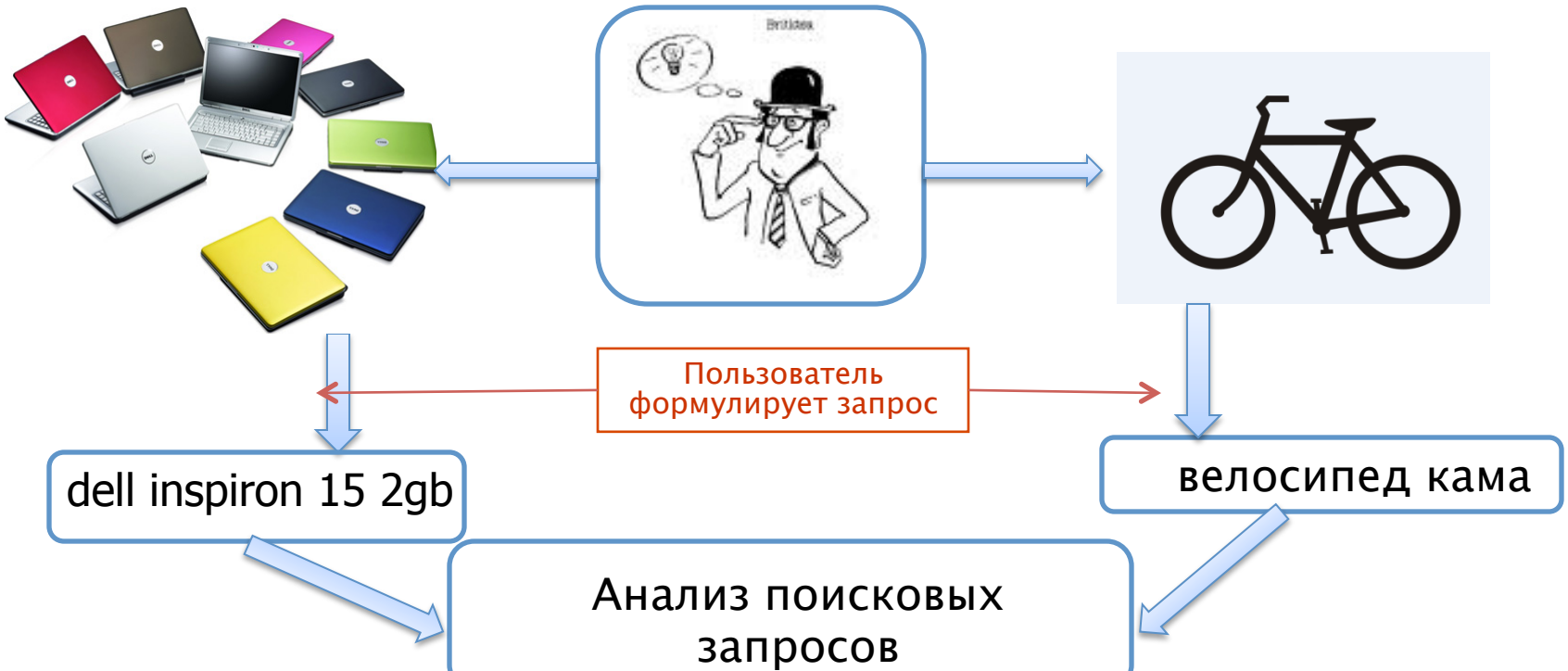
Юлия Киселёва

Санкт-Петербург
2010

План

- Введение
- Сегментация запросов
- Преобразование неструктурированных данных в структурированные
- Анализ поисковых запросов на предмет синонимов
- Анализ намерений пользователей

Поисковые запросы пользователей




Dell Inspiron 1545 T4200/15.6"WXGA/2Gb/250Gb/H...
красный

Код 53353
 Модель: Inspiron 154
 Процессор: Intel Dual
 Память: 2048MB.
 HDD: 250GB.
 Оптический привод:
 Видеоадаптер: ATI M
 LAN: 10/ 1000.
 Faxmodem: нет.
 Wireless LAN: Wi-Fi (Bluetooth 2.0.
 Порты: 3 x USB 2.0, f
 подключения микро
 Аудио: HD Audio.
 Card-reader: SD/ MM
 Дисплей: 15.6" (1366
 ОС: Windows Vista Hi
 Питание: батарейны
 Габариты: 34.5 x 3.8
 Вес: 2.64 кг.




Кама

[Летний ассортимент/](#) [Велосипеды/](#) Кама



Кама 2020 Размер колеса 20", 6 скор., дв
пластиковые крылья, Shimano

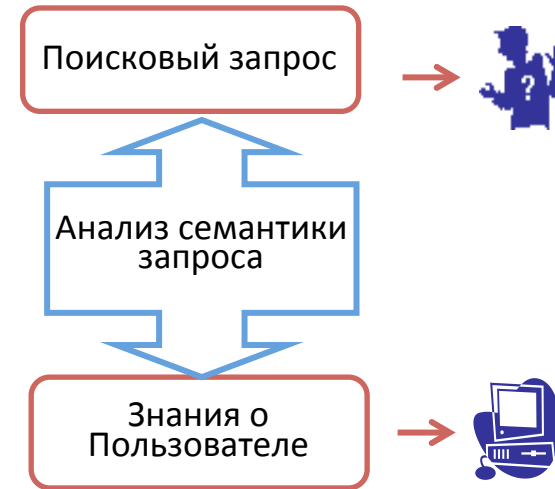


Кама F - 200 Luxe
 1 скор. низкая складная рама,
 руля с быстрым съемом, горна
 длин. хром. крылья, втулки КТ
 влагозащ. кареточный узел, зе
 Цвета: синий, красный, зелен
 фиолетовый, бордовый.



Цели

- Улучшение качества поиска.
- Улучшение ранжирования результатов поиска.
- Персонализация веб-пространства.
- Структуризация для неструктурированных запросов с «ключевыми словами».
 - Большая часть информации для Интернета изначально храниться в структурированных каталогах
- Показ контекстной рекламы.
- Показ запросов подсказок.



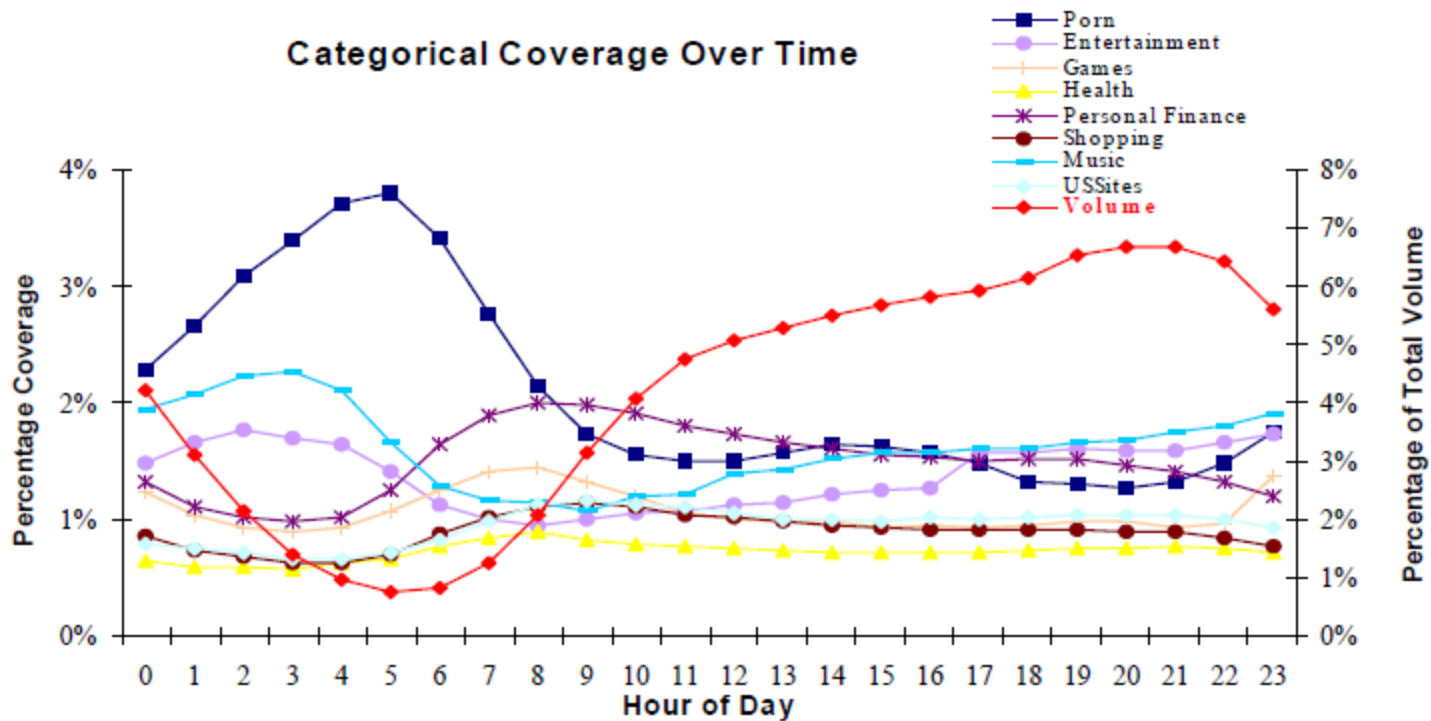
Несколько цифр

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

Классификация запросов (Bernard, 2007)

- Информационные (80,6%) – “maximization –expectation algorithm”
- Навигационные (10,2%) - “google.com”, “Warsaw Airport”
- Транзакционные (9,2 %) - “Mars surface image”, “Christmas present ideas”

Популярность запросов по тематикам [Beitzel]



Источники и типы данных [Jiang et al. 2010]



Как оценивать?

- Вручную?
- Друзья?

The screenshot shows the top navigation bar of the Amazon Mechanical Turk website. On the left is the logo for Amazon Mechanical Turk, with the tagline 'Artificial Intelligence'. The navigation menu includes 'Your Account', 'HITS', and 'Qualifications'. On the right, there are links for 'Already have an account?' and 'Sign in as a Worker | Requester'. Below the navigation bar is a blue banner with the text: 'Introduction | Dashboard | Status | Account Settings'. The main banner is yellow and contains the text: 'Mechanical Turk is a marketplace for work. We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient. 60,421 HITS available. View them now.'

Make Money by working on HITS

HITS - *Human Intelligence Tasks* - are individual tasks that you work on. [Find HITS now.](#)

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



[or learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITS - *Human Intelligence Tasks* - and get results using Mechanical Turk. [Register Now.](#)

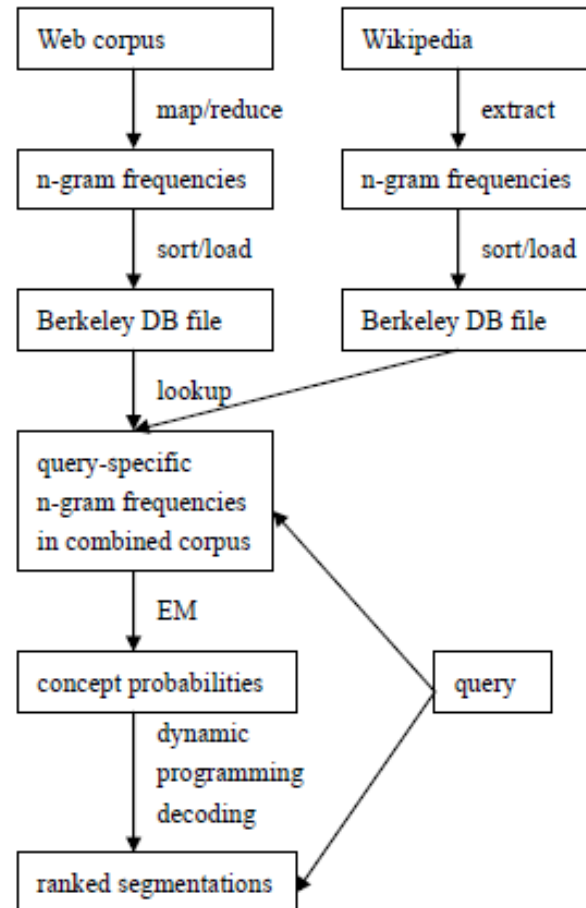
As a Mechanical Turk Requester you:

- I have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITS completed in minutes
- Pay only when you're satisfied with the results



Сегментация запросов пользователей «без учителя» [Tan et al. 2008]

- [New] [York] [times] [subscription]
- [New York] [times] [subscription]
- [New York] [times subscription]



[<http://www.oracle.com/database/berkeley-db>]
AOL search

Conditional Random Fields

$X = (X_1, X_2, \dots, X_n)$ - запрос, состоящий из n слов

$y = (y_1, y_2, \dots, y_n)$ - последовательность атрибутов для этих n слов

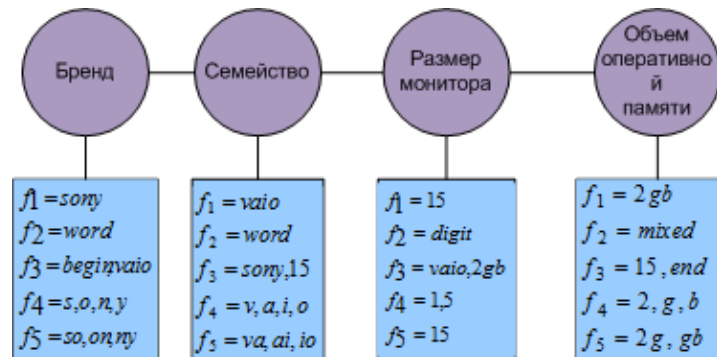
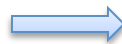
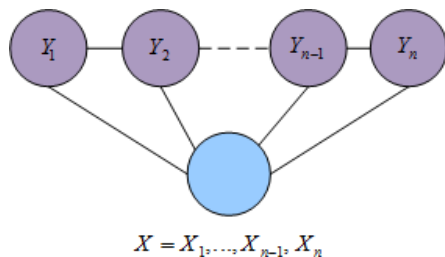
$$f_j(y_{i-1}, y_i, x, i) = \exp(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i))$$

$t_j(y_{i-1}, y_i, x, i)$ - вероятность перехода

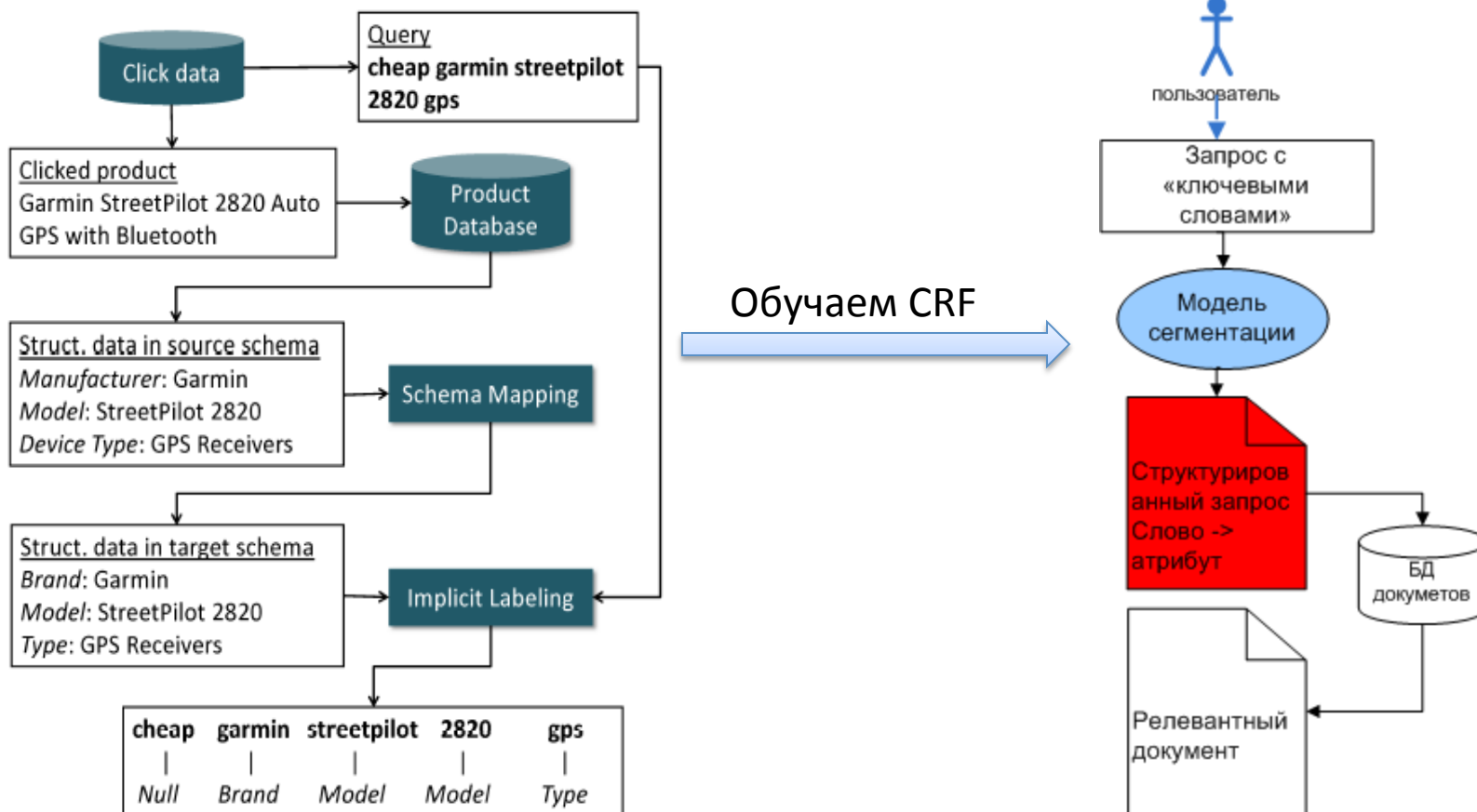
$s_k(y_i, x, i)$ - признак

$$p(y | x, \lambda) = \frac{1}{Z(x, \lambda)} \exp(\sum_j \lambda_j f_j(y, x))$$

$\{(X^{(i)}, y^{(i)})\}_{i=1}^m$ - обучающее множество



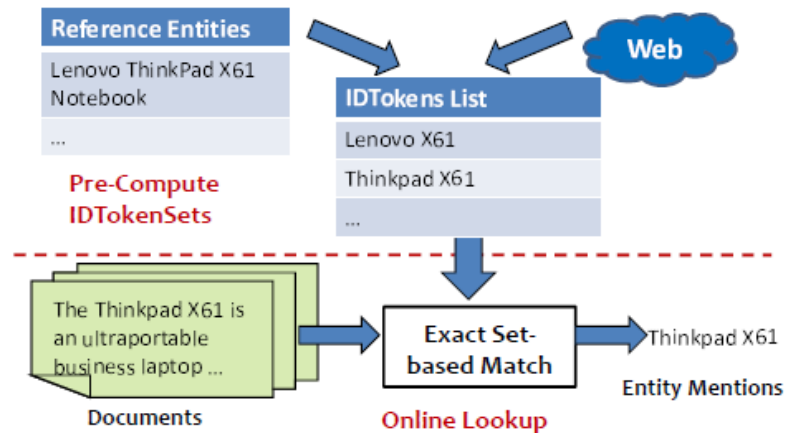
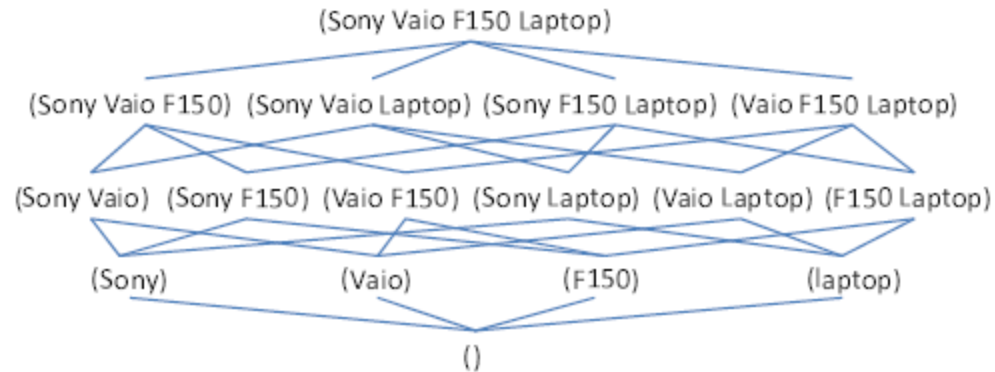
Сегментация запросов о продуктах



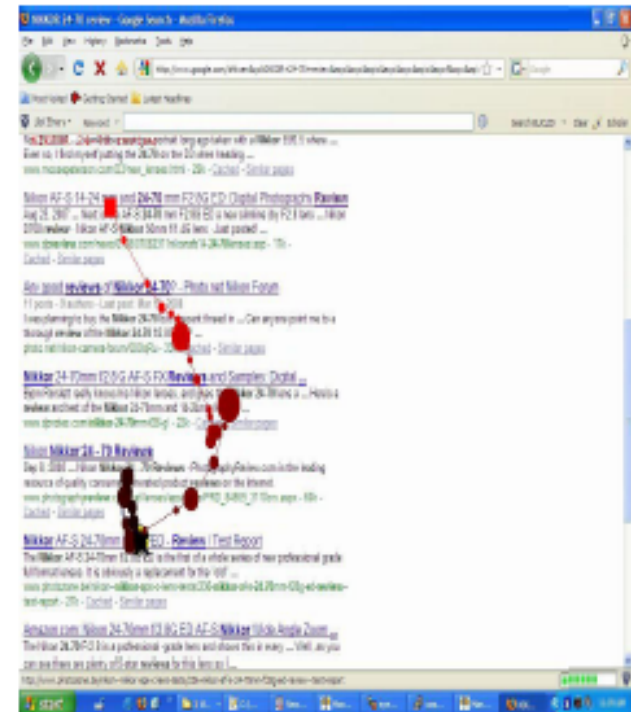
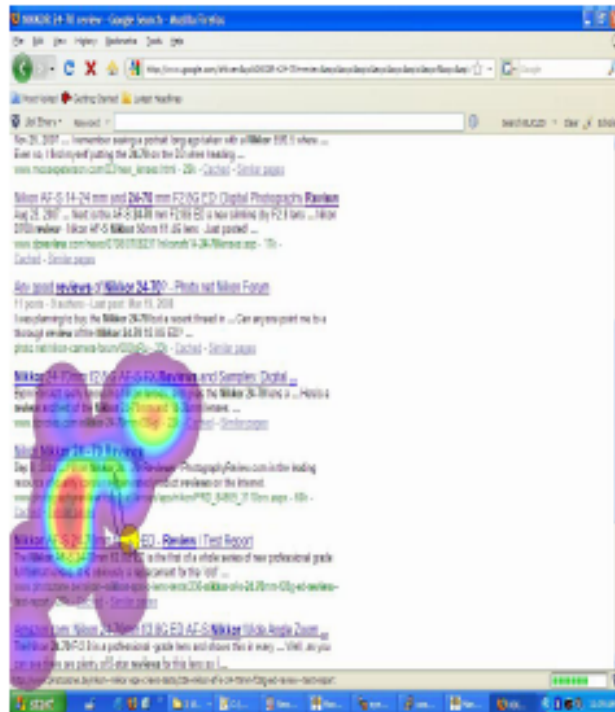
Сегментация с несколькими атрибутами [Drezde et al. 2009]

<i>John</i>	<i>studies</i>	<i>at</i>	<i>the</i>	<i>University</i>	<i>of</i>	<i>California</i>	<i>.</i>
<i>PER</i>	<i>O</i>	<i>O</i>	<i>O</i>	<i>ORG</i>	<i>ORG</i>	<i>ORG (0.33)</i>	<i>O</i>
						<i>LOC (0.67)</i>	

Анализ поисковых запросов на предмет синонимов [Chaudhuri et al. 2009]



Определение поисковых целей пользователя целей [Guo and Agichetein 2010]



Вопросы?