

# Анализ логов запросов

По материалам курса

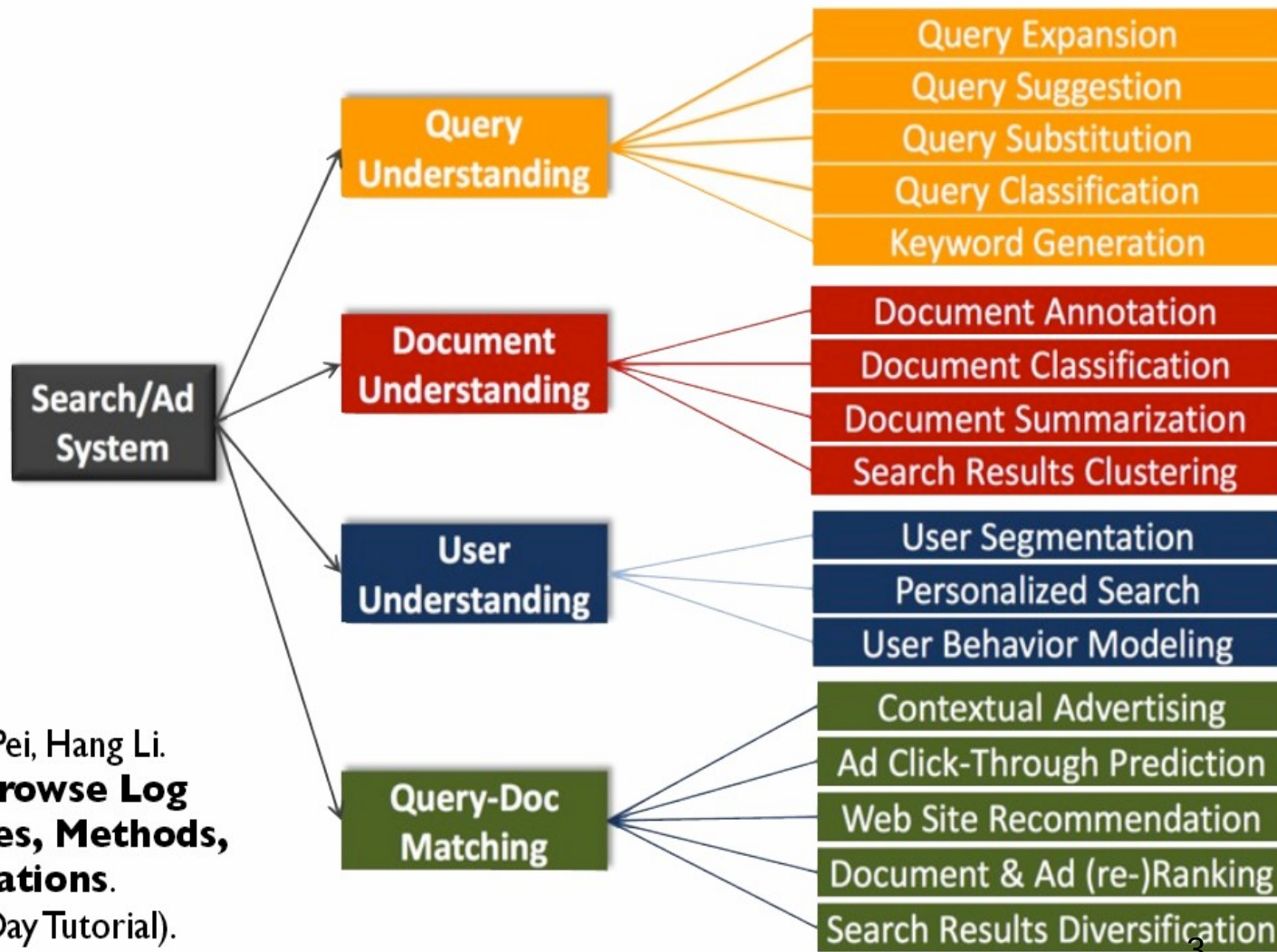
Mining query logs to improve web search  
engines' operations (QLM)

(Russir 2011)

# Приложения

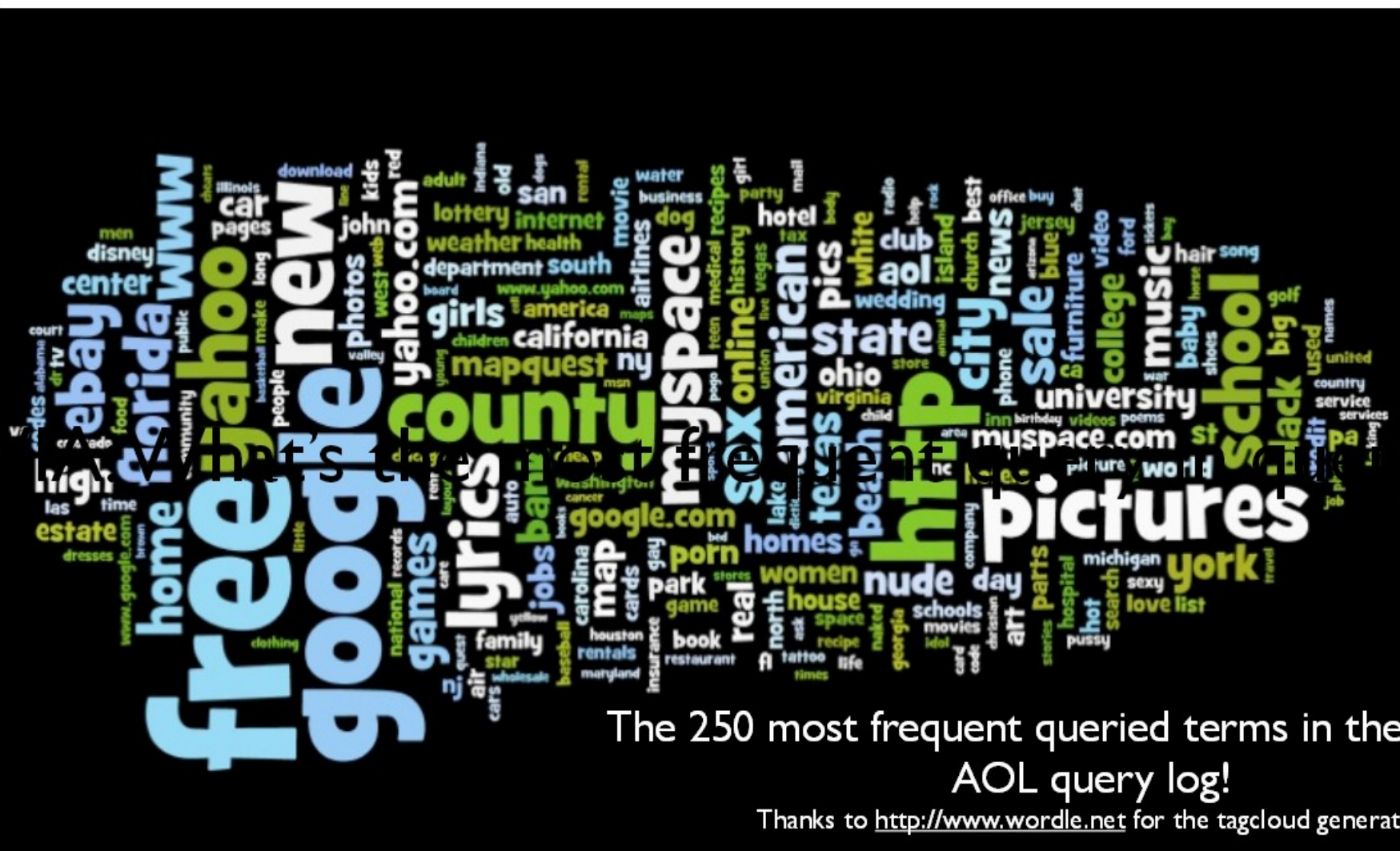
- Расширение запросов
- Предложение запросов
- Персонализация результатов
- Обучение ранжированию (learning to rank)
- Повышение скорости обработки запросов
  - Кеширование в поисковых машинах
  - Разделение и отбор коллекций

# Log (Usage) Mining Apps



From:  
Daxin Jiang, Jian Pei, Hang Li.  
**Web Search/Browse Log  
Mining: Challenges, Methods,  
and Applications.**  
WWW'10 (Full-Day Tutorial).

## y lc



## The 250 most frequent queried terms in the AOL query log!

Thanks to <http://www.wordle.net> for the tagcloud generator

# Some Popular Terms: Excite and Altavista

query	freq.
<i>*Empty Query*</i>	2,586
sex	229
chat	58
lucky number generator	56
p****	55
porno	55
b****y	55
nude beaches	52
playboy	46
bondage	46
porn	45
rain forest restaurant	40
f****ing	40
crossdressing	39
crystal methamphetamine	36
consumer reports	35
xxx	34
nude tanya harding	33
music	33
sneaker stories	32

(a) Excite.

query	freq.
christmas photos	31,554
lyrics	15,818
cracks	12,670
google	12,210
gay	10,945
harry potter	7,933
wallpapers	7,848
pornografia	6,893
"yahoo com"	6,753
juegos	6,559
lingerie	6,078
sybiosis logic 53c400a	5,701
letras de canciones	5,518
humor	5,400
pictures	5,293
preteen	5,137
hypnosis	4,556
cpc view registration key	4,553
sex stories	4,521
cd cover	4,267

(b) Altavista.



# Topic Distribution: Excite and AOL

Topic	Percentage
Entertainment or recreation	19.9%
Sex and pornography	16.8%
Commerce, travel, employment, or economy	13.3%
Computers or Internet	12.5%
Health or sciences	9.5%
People, places, or things	6.7%
Society, culture, ethnicity, or religion	5.7%
Education or humanities	5.6%
Performing or fine arts	5.4%
Non-English or unknown	4.1%
Government	3.4%

Excite

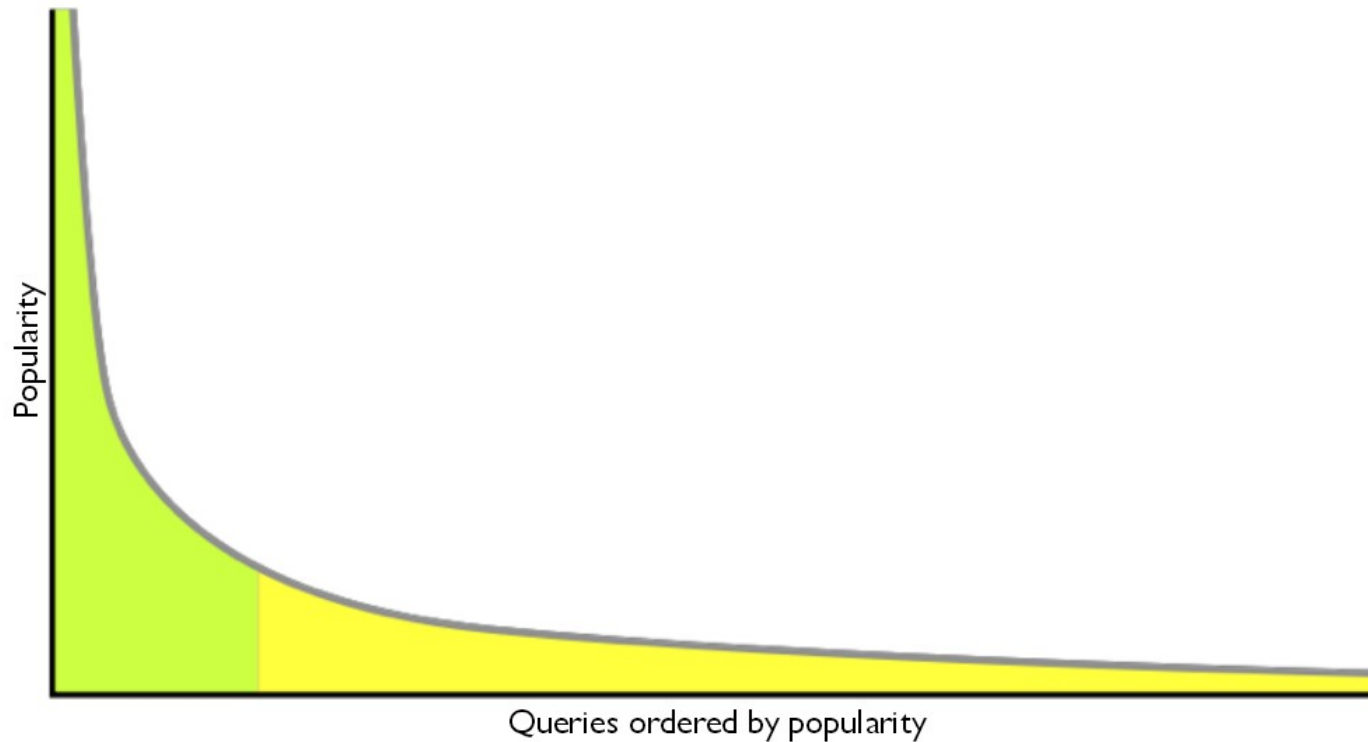
Topic	Percentage
Entertainment	13%
Shopping	13%
Porn	10%
Research & learn	9%
Computing	9%
Health	5%
Home	5%
Travel	5%
Games	5%
Personal & Finance	3%
Sports	3%
US Sites	3%
Holidays	1%
Other	16%

AOL

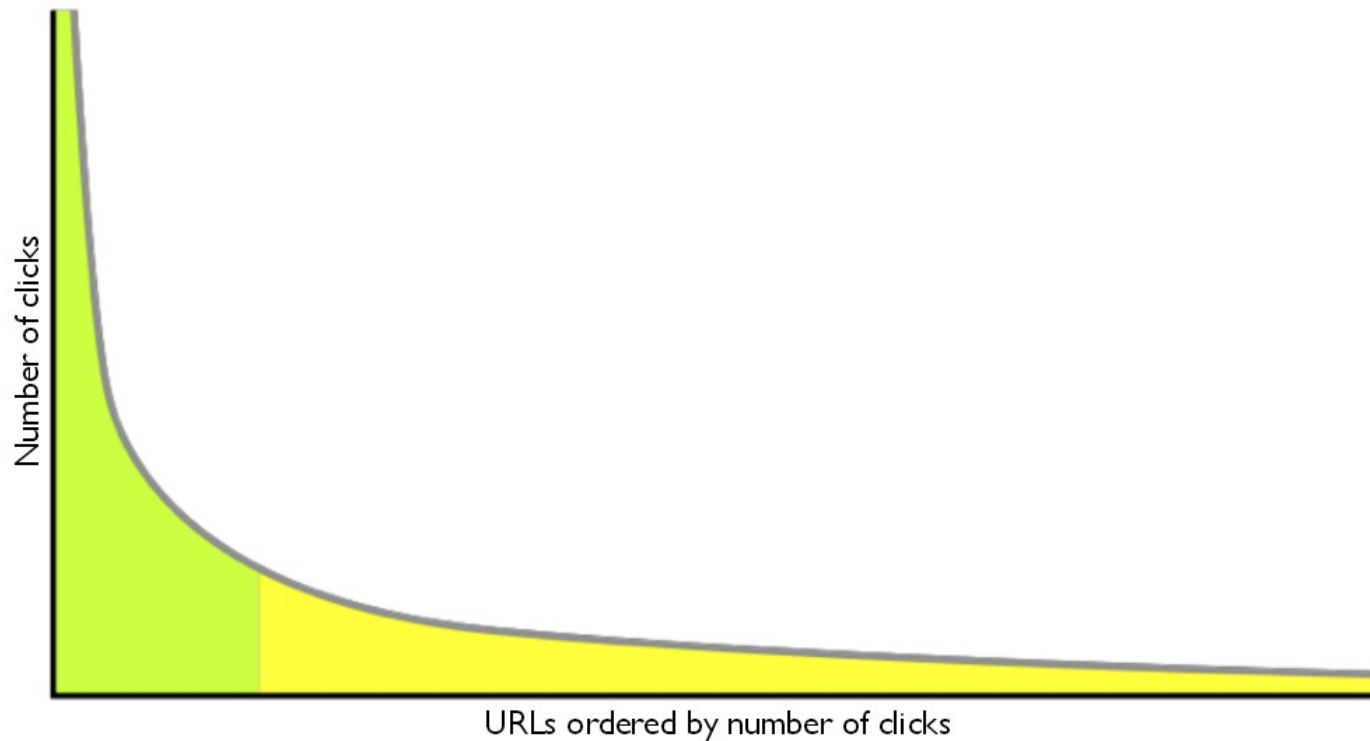
A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, "**From e-sex to e-commerce: Web search changes.**" Computer, vol. 35, no. 3, pp. 107–109, 2002.

S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, "**Temporal analysis of a very large topically categorized web query log.**" J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 2, pp. 166–178, 2007.

# Long Tail Distribution

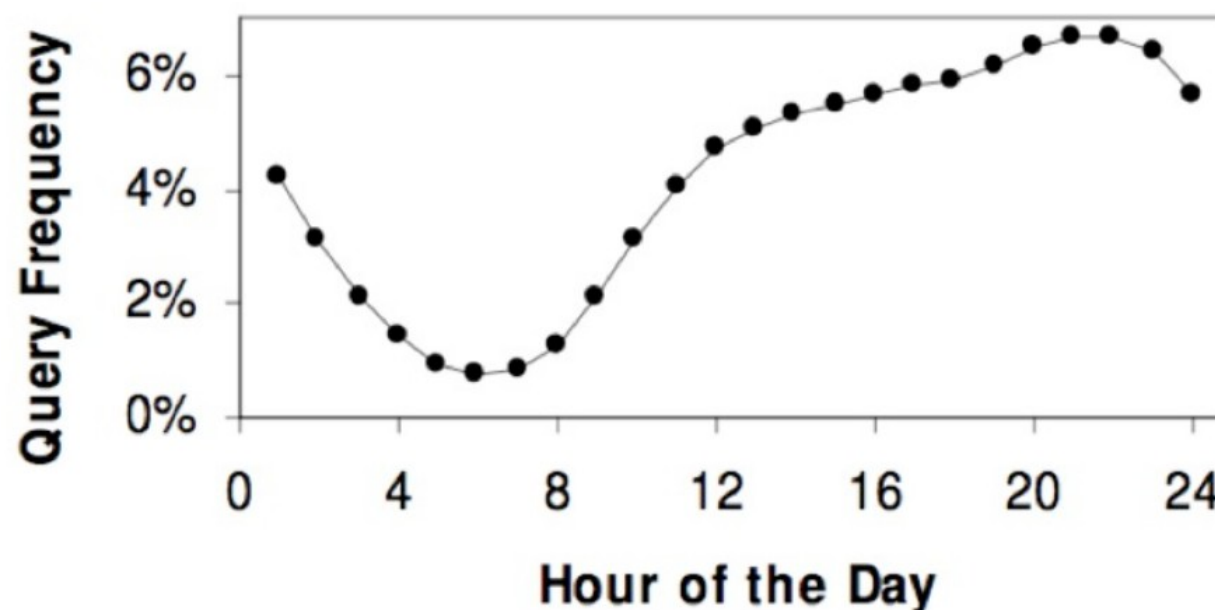


# Long Tail Distribution



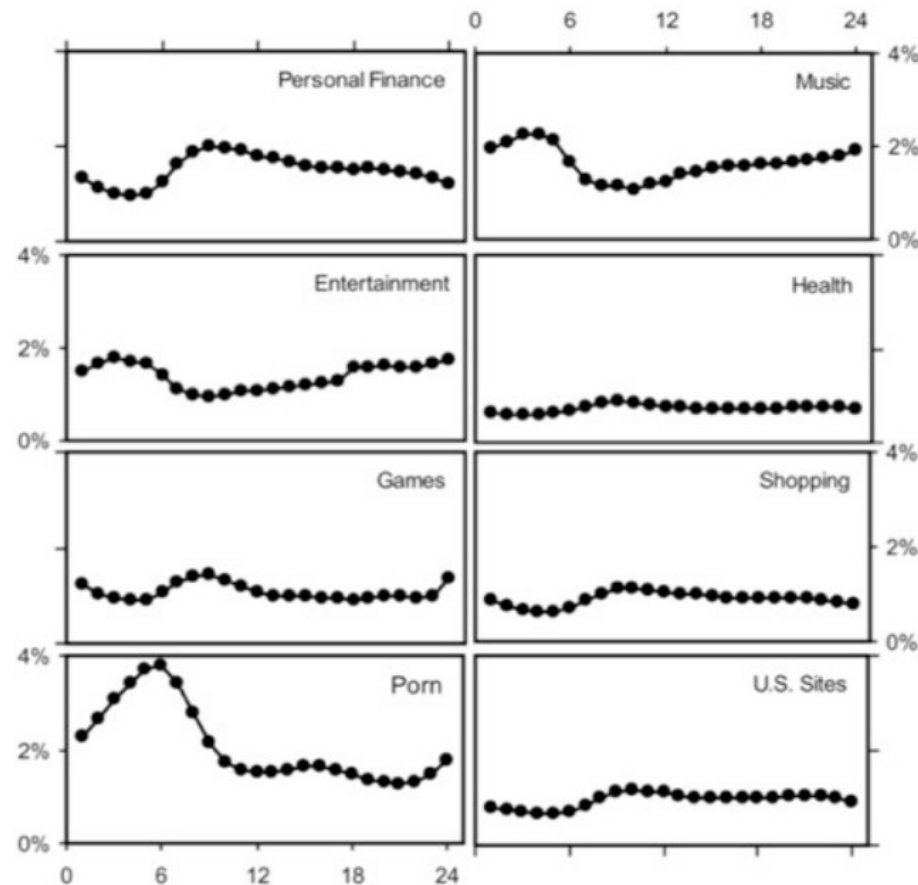


# Frequency of Query Submission



S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, "Temporal analysis of a very large topically categorized web query log," J. Am. Soc. Inf. Sci. Technol., vol. 58, no. 2, pp. 166–178, 2007.

# Hourly Topic Distribution



# Классификация поисковых запросов

- Навигационные запросы:
  - Нужно найти определенный сайт
- Информационные запросы
  - Найти информацию по запросу
- Транзакционные запросы (запросы на услугу, на ресурсы)
  - Скачать, купить, послушать, заказать и др.

# Navigational Queries

- American Airlines
- AA
- Google
- Yahoo
- CNN



They account for the 20 ~ 25% of the total queries.

# Informational Queries

- High Dynamic Resolution Photos
- Escher
- Transfinite Numbers



They account for the 40 ~ 45% of the total queries.





# Transactional Queries

- MP3
- Hotels Saint Petersburg
- Tickets for the Hermitage



They account for the 30 ~ 35% of the total queries.



# A Refined Taxonomy

Navigational

Informational

Resource

Looking for a particular web site.

Willing to satisfy an information need

Looking for obtaining resource (not information) available on the Web

# A Refined Taxonomy

Navigational

Informational

Resource


Download

Entertainment





Interact

Obtain



# Запрос ВМК МГУ: первый документ по навигационному запросу, второй – по информационному







✕Найти


плюс



[Поиск](#) [Картинки](#) [Видео](#) [Карты](#) [Маркет](#) [Новости](#) [ТВ онлайн](#) [Музыка](#) [Ещё](#)


**ВМК МГУ**  
[Дополнительное образование](#) [Контакты](#) [Новости](#) [Наши выпускники](#)  
[en.cs.msu.ru](#)   
Информация о факультетах. Справочник по кафедрам. Учебный процесс: расписание занятий и экзаменов, учебные и рабочие планы. Направления научной деятельности. Условия поступления. Контакты.

**Факультет вычислительной математики и кибернетики...**  
[ru.wikipedia.org](#) > [Факультет вычислительной математики и кибернетики ...](#)   
Факультет вычислительной математики и кибернетики (ВМК, ранее ВМиК) **Московского государственного университета** имени М. В. Ломоносова — учебный центр по подготовке кадров в области фундаментальных исследований по прикладной математике, вычислительной ... [Читать ещё >](#)

**ВМК МГУ — ВКонтакте**  
[vk.com](#) > [smcmsu](#)   
Перейдите на страницу пользователя, чтобы посмотреть публикации или отправить сообщение.  
**О себе:** Москва, Россия, Неофициальная группа студентов и в...  
[Читать ещё >](#)



**Факультет вычислительной математики и кибернетики МГУ**  
[msu.ru](#) > [info/struct/dep/vmc.html](#)   
Факультет вычислительной математики и кибернетики (ВМК) **Московского государственного университета** имени М.В.Ломоносова является ведущим...



6 фото

# Поисковая сессия

- Совокупность запросов, которая задается пользователем в течение некоторого интервала времени.
- Граница интервала – пауза в запросах
- Типичная сессия
  - Два запроса
  - Из двух слов
  - Две страницы выдачи
  - Два клика на страницу

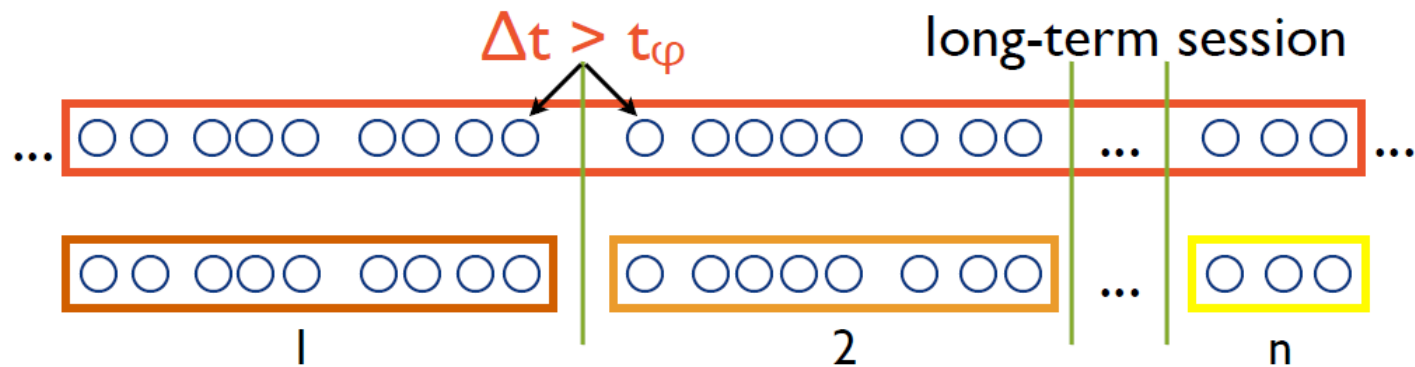


# The Big Picture

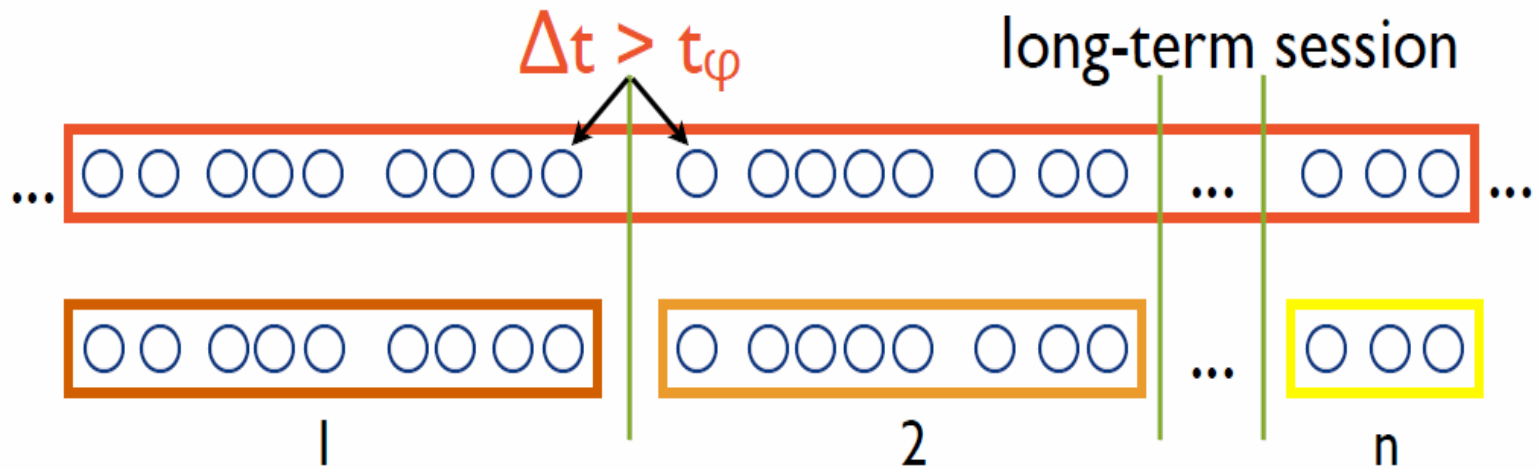
long-term session



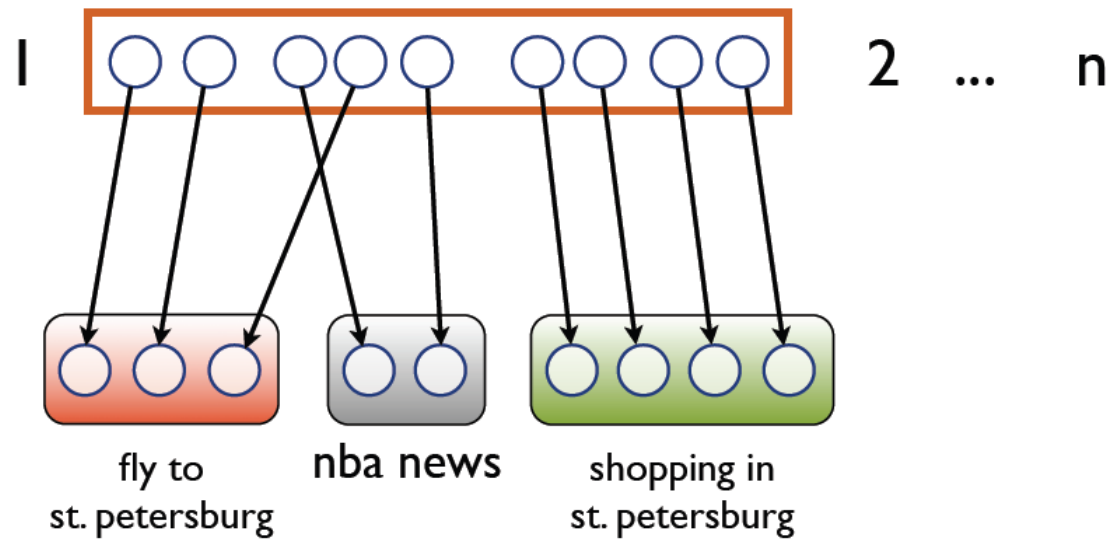
# The Big Picture



# The Big Picture



# The Big Picture

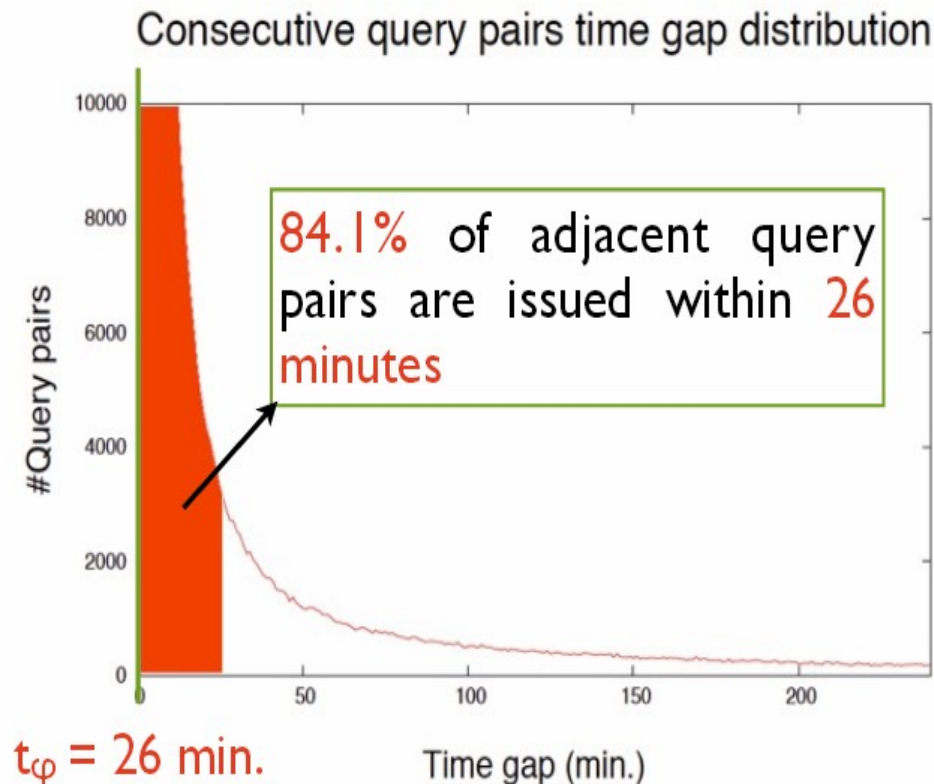


# Методы разделения на сессии

- Разделение по времени
  - Проблема: многозадачная сессия
- Разделение по сходству (content-based)
  - Близкие по смыслу запросы могут быть не похожи по словам (vocabulary mismatch)
- Комплексный подход
  - Включает дополнительные ресурсы (например, Википедию) (Semantic-based)



# Data Analysis: query time gap



# Query Features

## Content-based ( $\mu_{\text{content}}$ )

- ✓ two queries ( $q_i, q_j$ ) sharing common terms are likely related
- ✓  $\mu_{\text{jaccard}}$ : Jaccard index on query character **3-grams**

$$\mu_{\text{jaccard}}(q_1, q_2) = 1 - \frac{|T(q_1) \cap T(q_2)|}{|T(q_1) \cup T(q_2)|}$$

- ✓  $\mu_{\text{levenshtein}}$ : normalized Levenshtein distance

$$\mu_{\text{content}}(q_1, q_2) = \frac{(\mu_{\text{jaccard}} + \mu_{\text{levenshtein}})}{2}$$

## Semantic-based ( $\mu_{\text{semantic}}$ )

- ✓ using **Wikipedia** and **Wiktionary** for “expanding” a query  $q$
- ✓ “**wikification**” of  $q$  using **vector-space model**

$$\vec{C}(t) = (c_1, c_2, \dots, c_W) \quad \vec{C}(q) = \sum_{t \in q} \vec{C}(t)$$

- ✓ relatedness between ( $q_i, q_j$ ) computed using **cosine-similarity**

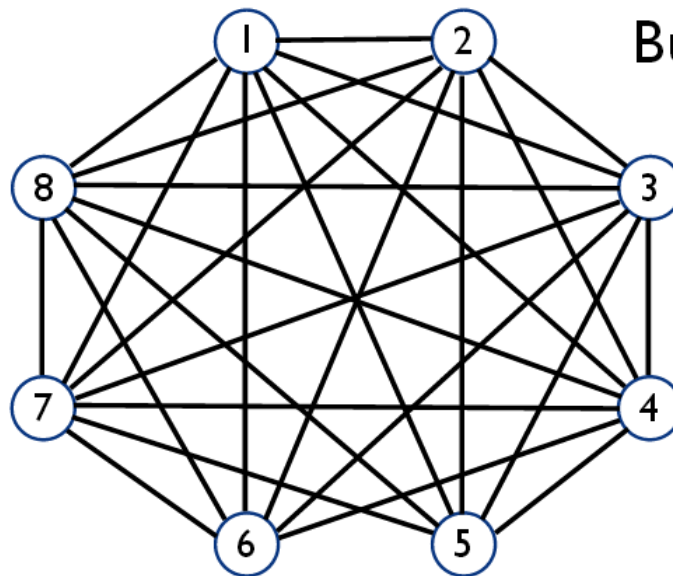
$$\text{rel}(q_1, q_2) = \frac{\vec{C}(q_1) \cdot \vec{C}(q_2)}{|\vec{C}(q_1)| |\vec{C}(q_2)|}$$

$$\mu_{\text{wikification}}(q_1, q_2) = 1 - \text{rel}(q_1, q_2)$$

$$\mu_{\text{semantic}}(q_1, q_2) = \min(\mu_{\text{wiktionary}}, \mu_{\text{wikipedia}})$$

Сходство между запросами  
на основе расширенного пред-  
ставления Википедии

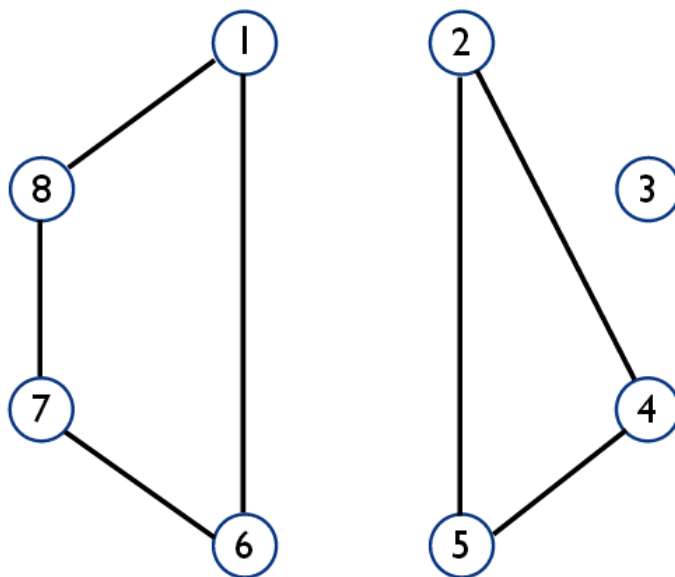
# QC-WCC



Build similarity graph  $G_\varphi$

Показано значительное  
улучшение выделения  
поисковых сессий на  
основе сопоставления  
с Википедией

# QC-WCC



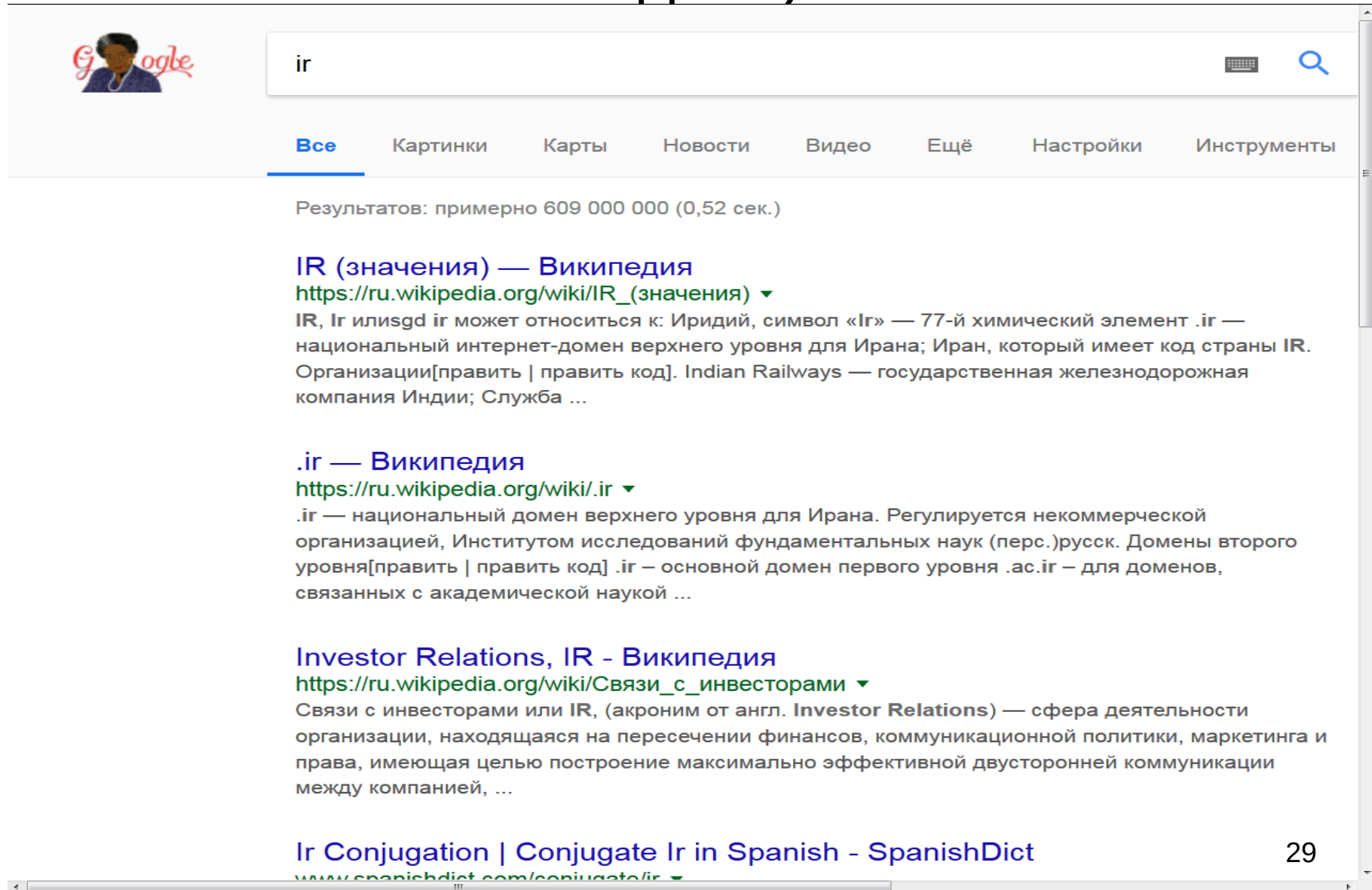
Drop “weak edges”

# Персонализация обработки запросов

- Персонализация состоит в представлении разных результатов поиска
  - Пользователям с различными интересами
  - Пользователям в разных контекстах (место, время)
- Пример: экономист и математик задают вопрос «теория игр»
  - Математик: теоретические вопросы
  - Экономист: приложение теории игр к экономики
- Поиск по сокращениям:
  - IR, CHI, ФБК и тп.



# Запрос в Google: IR (диверсификация выдачи)



The screenshot shows a Google search interface with the query 'ir' entered in the search bar. The search results are displayed below the navigation tabs. The first result is 'IR (значения) — Википедия' with a URL link. The second result is '.ir — Википедия' with a URL link. The third result is 'Investor Relations, IR - Википедия' with a URL link. The fourth result is 'Ir Conjugation | Conjugate Ir in Spanish - SpanishDict' with a URL link. The search results are sorted by 'Все' (All).

Google

ir

Все Картинки Карты Новости Видео Ещё Настройки Инструменты

Результатов: примерно 609 000 000 (0,52 сек.)

**IR (значения) — Википедия**  
[https://ru.wikipedia.org/wiki/IR\\_\(значения\)](https://ru.wikipedia.org/wiki/IR_(значения)) ▼  
IR, Ir илиsgd ir может относиться к: Иридий, символ «Ir» — 77-й химический элемент .ir — национальный интернет-домен верхнего уровня для Ирана; Иран, который имеет код страны IR. Организации[править | править код]. Indian Railways — государственная железнодорожная компания Индии; Служба ...

**.ir — Википедия**  
<https://ru.wikipedia.org/wiki/.ir> ▼  
.ir — национальный домен верхнего уровня для Ирана. Регулируется некоммерческой организацией, Институтом исследований фундаментальных наук (перс.)русск. Домены второго уровня[править | править код] .ir — основной домен первого уровня .ac.ir — для доменов, связанных с академической наукой ...

**Investor Relations, IR - Википедия**  
[https://ru.wikipedia.org/wiki/Связи\\_с\\_инвесторами](https://ru.wikipedia.org/wiki/Связи_с_инвесторами) ▼  
Связи с инвесторами или IR, (акроним от англ. **Investor Relations**) — сфера деятельности организации, находящаяся на пересечении финансов, коммуникационной политики, маркетинга и права, имеющая целью построение максимально эффективной двусторонней коммуникации между компанией, ...

**Ir Conjugation | Conjugate Ir in Spanish - SpanishDict**  
[www.spanishdict.com/conjugate/ir](http://www.spanishdict.com/conjugate/ir) ▼

# Методы персонализации

- 1. Расширение запроса
  - Например, автоматический добавить в запрос IR - information retrieval, если известно, что пользователя интересует эта тема
- 2. Переранжирование

# Как выяснить интересы пользователя

- Пользователь описывает сам свой профиль – обычно получается плохо
- Автоматическое выявление интересов:
  - Ранее заданные поисковые запросы
  - Посещенные страницы
  - Документы пользователя
  - Электронная почта
- Приватность? Охрана персональных данных

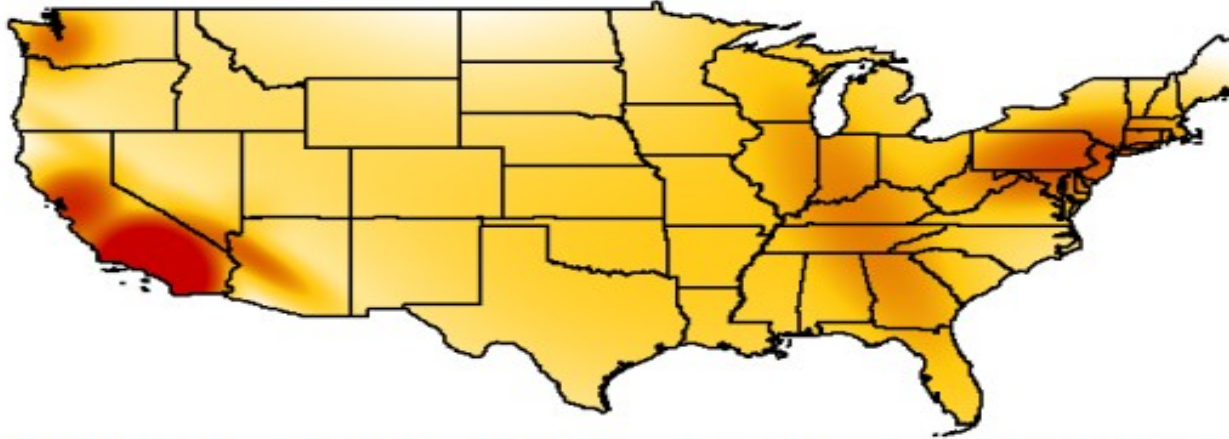
# Геотаргетинг

- Учет местоположения пользователя – один из часто встречающихся видов персонализации
- Геозависимые запросы
- Изучение поведения пользователей:
  - Пользователи в некотором месте предпочитают выбирать страницы, важные для этого места

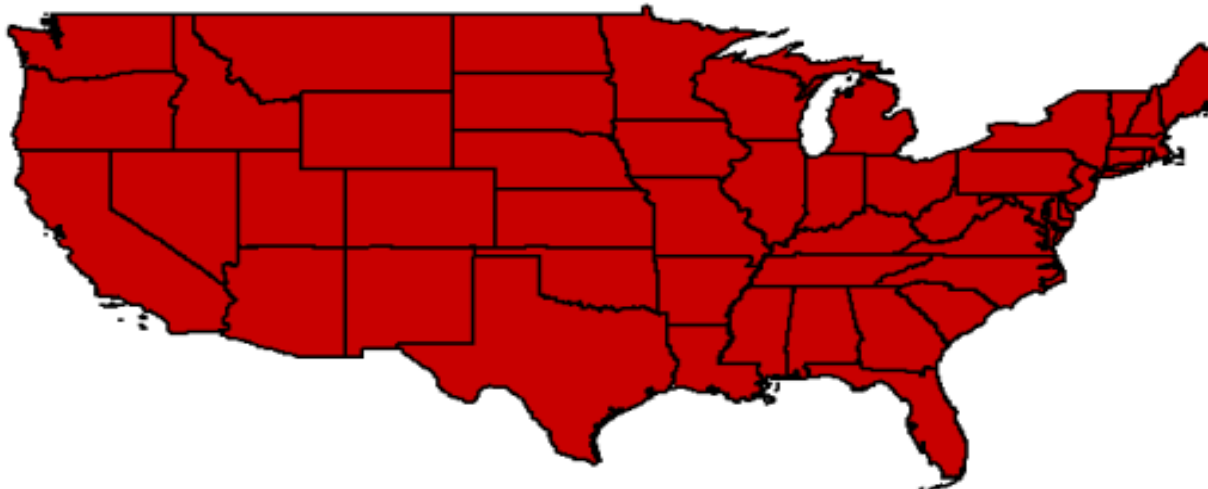
$$P(location = x | URL)$$

- Учет IP адресов

# Интерес к сайту в зависимости от местоположения пользователя



(c) Los Angeles Times: Reviews and Recommendations  
<http://findlocal.latimes.com/>



(d) Los Angeles Times: Crossword Puzzles and Games  
<http://games.latimes.com/>

# Были рассмотрены

- Логи поисковых запросов
- Зависимость задаваемых запросов от времени суток
- Тематическая классификация и классификация по цели
- Разбиение запросов на сессии
- Персонализация обработки запросов

# Анализ кликов для улучшения качества поиска

# Запрос CIKM – это известная конференция по knowledge management.

## Дело происходит в 2009 году

ALL RESULTS 1-10 of 131,000 results · [Advanced](#)

**CIKM 2008 | Home**  
Napa Valley Marriott Hotel & Spa: Napa Valley, California October 26-30, 2008  
[cikm2008.org](#) · [Cached page](#)

Papers Program Committee  
Themes News  
Important Dates Napa Valley  
Banquet Posters  
[Show more results from cikm2008.org](#)

**Conference on Information and Knowledge Management (CIKM)**  
Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...  
[www.cikm.org](#) · [Cached page](#)

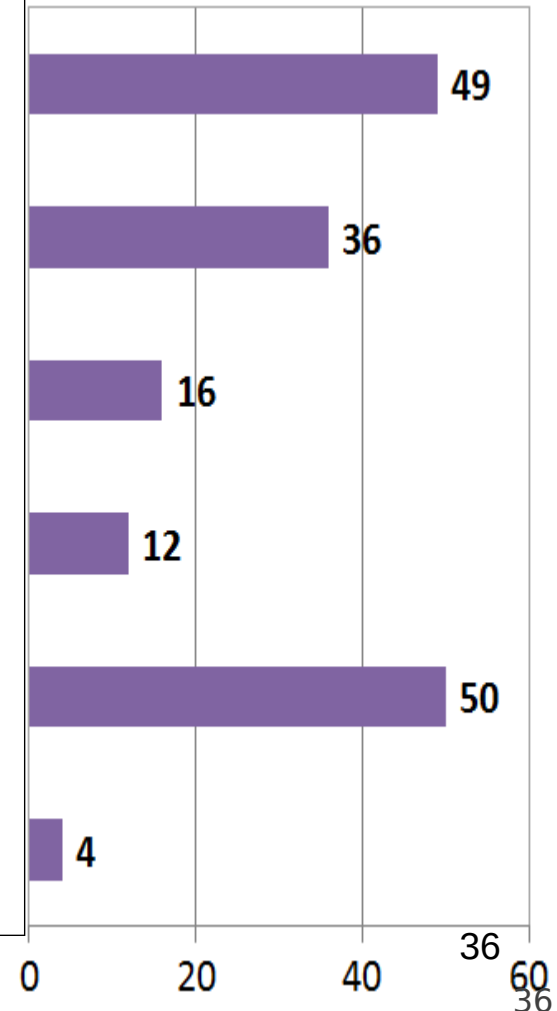
**Conference on Information and Knowledge Management (CIKM'02)**  
SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.  
[www.cikm.org/2002](#) · [Cached page](#)

**ACM CIKM 2007 - Lisbon, Portugal**  
News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.  
[www.fc.ul.pt/cikm2007](#) · [Cached page](#)

**CIKM 2009 | Home**  
CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...  
[www.comp.polyu.edu.hk/conference/cikm2009](#) · [Cached page](#)

**Conference on Information and Knowledge Management (CIKM)**  
CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...  
[cikmconference.org](#) · [Cached page](#)

# of clicks  
received





# Как улучшить выдачу используя клики?

ALL RESULTS

1-10 of 131,000 results · [Advanced](#)

**CIKM 2008 | Home**  
Napa Valley Marriott Hotel & Spa: Napa Valley, California October 26-30, 2008  
[cikm2008.org](#) · [Cached page](#)

[Papers](#) [Program Committee](#)  
[Themes](#) [News](#)  
[Important Dates](#) [Napa Valley](#)  
[Banquet](#) [Posters](#)

Show more results from cikm2008.org

**Conference on Information and Knowledge Management (CIKM)**  
Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...  
[www.cikm.org](#) · [Cached page](#)

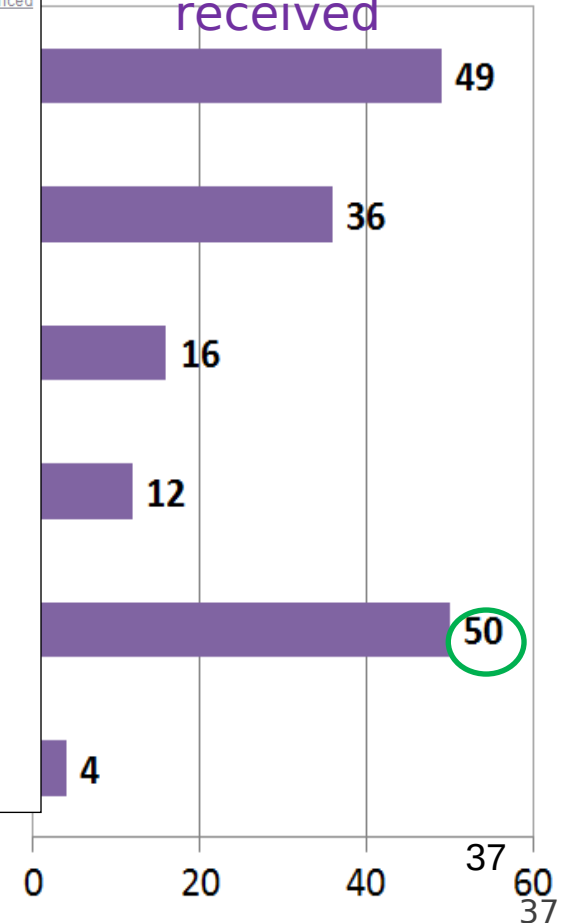
**Conference on Information and Knowledge Management (CIKM'02)**  
SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.  
[www.cikm.org/2002](#) · [Cached page](#)

**ACM CIKM 2007 - Lisbon, Portugal**  
News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.  
[www.fc.ul.pt/cikm2007](#) · [Cached page](#)

**CIKM 2009 | Home**  
CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...  
[www.comp.polyu.edu.hk/conference/cikm2009](#) · [Cached page](#)

**Conference on Information and Knowledge Management (CIKM)**  
CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...  
[cikmconference.org](#) · [Cached page](#)

# of clicks  
received



# Лог веб кликов

Query	cikm	Session ID	f851c5af178384d12f3d
Position	URL	Click	
1	<a href="http://cikm2008.org">cikm2008.org</a>	1	
2	<a href="http://www.cikm.org">www.cikm.org</a>	0	
3	<a href="http://www.cikm.org/2002">www.cikm.org/2002</a>	0	
4	<a href="http://www.fc.ul.pt/cikm2007">www.fc.ul.pt/cikm2007</a>	0	
5	<a href="http://www.comp.polyu.edu.hk/conference/cikm2009">www.comp.polyu.edu.hk/conference/cikm2009</a>	1	
6	<a href="http://cikmconference.org">cikmconference.org</a>	0	
7	<a href="http://lr.iit.edu/cikm2004">lr.iit.edu/cikm2004</a>	0	
8	<a href="http://www.informatik.uni-trier.de/~ley/db/conf/cikm/index.html">www.informatik.uni-trier.de/~ley/db/conf/cikm/index.html</a>	0	
9	<a href="http://www.tzi.de/CIKM2005">www.tzi.de/CIKM2005</a>	0	
10	<a href="http://www.cikm.com">www.cikm.com</a>	0	

# Интерпретация кликов как relevance feedback

[CIKM 2008 | Home](#)  
Napa Valley Marriott Hotel Napa: Napa Valley, California October 26-30, 2008  
[cikm2008.org](#) - [Cached page](#)

<a href="#">Papers</a>	<a href="#">Program Committee</a>
<a href="#">Themes</a>	<a href="#">News</a>
<a href="#">Important Dates</a>	<a href="#">Napa Valley</a>
<a href="#">Banquet</a>	<a href="#">Posters</a>

[Show more results from cikm2008.org](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)  
Provides an international forum for presentation and discussion of research on information and knowledge management, as well as recent advances on data and knowledge bases ...  
[www.cikm.org](#) - [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM'02\)](#)  
SAIC Headquarters, McLean, Virginia, USA, 4-9 November 2002.  
[www.cikm.org/2002](#) - [Cached page](#)

[ACM CIKM 2007 - Lisbon, Portugal](#)  
News and announcements: 12/02 - Best interdisciplinary paper award at CIKM 2007 went to Fei Wu and Daniel Weld for Autonomously Semantifying Wikipedia.  
[www.fc.ul.pt/cikm2007](#) - [Cached page](#)

[CIKM 2009 | Home](#)  
CIKM 2009 (The 18th ACM Conference on Information and Knowledge Management) will be held on November 2-6, 2009, Hong Kong. Since 1992, CIKM has successfully brought together ...  
[www.comp.polyu.edu.hk/conference/cikm2009](#) - [Cached page](#)

[Conference on Information and Knowledge Management \(CIKM\)](#)  
CIKM Conference on Information and Knowledge Management The Conference on Information and Knowledge Management (CIKM) provides an international forum for presentation and ...  
[cikmconference.org](#) - [Cached page](#)

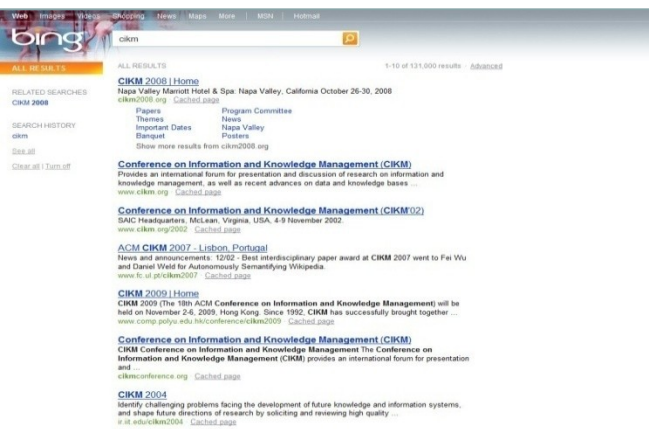
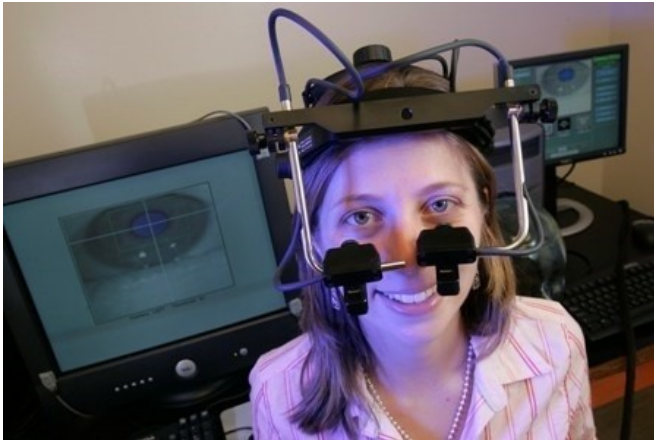
[CIKM 2004](#)  
Identify challenging problems facing the development of future knowledge and information systems, and shape future directions of research by soliciting and reviewing high quality ...  
[ir.iit.edu/cikm2004](#) - [Cached page](#)

[CIKM](#)  
International Conference on Information and Knowledge Management (CIKM) CIKM Home Page  
ACM DL: CIKM 17. CIKM 2008: Napa Valley, California, USA. James G. Shanahan, Sihem Amer-Yahia ...  
[www.informatik.uni-trier.de/~ley/db/conf/cikm/index.html](#) - [Cached page](#)

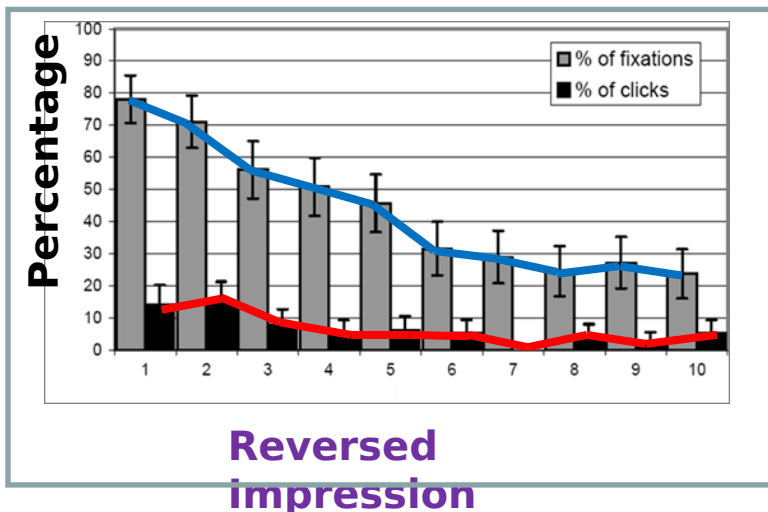
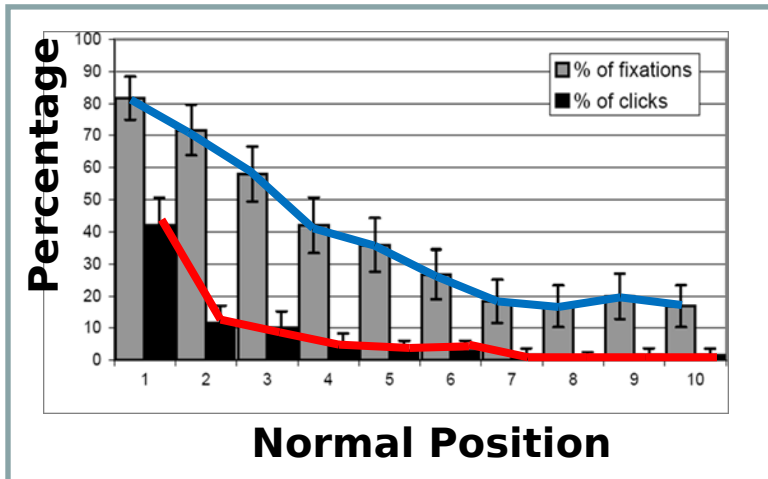
- Клики – это хорошо...
  - Одинаково ли хороши?
- Отсутствие кликов может объясняться:
  - Не релевант
  - Не видел



# Слежение за глазами (eye tracking)



# Неравноценность позиций относительно КЛИКОВ

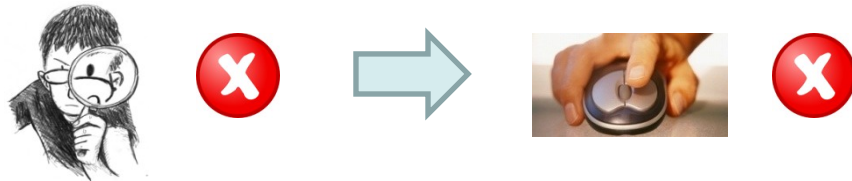


- Более высокие позиции получают больше кликов пользователя, чем более низкие позиции (**eye fixation**).
- Это справедливо, даже если выдачу переставить наоборот
- “Клики информативны, но смещены (biased)”.

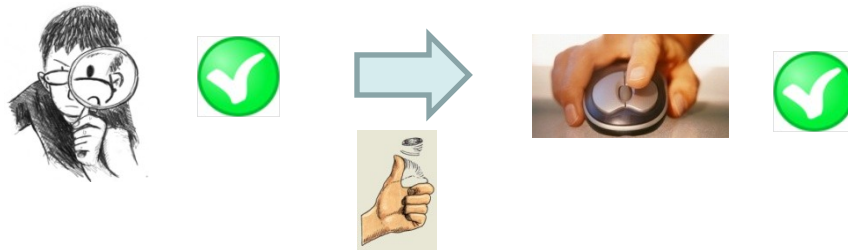
[Joachims+07]

# Гипотеза о «наблюдении» (Richardson и др. 2007)

- Документ должен быть прочитан перед кликом.

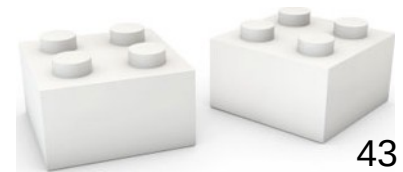


- Условная вероятность клика после прочтения зависит от релевантности документа



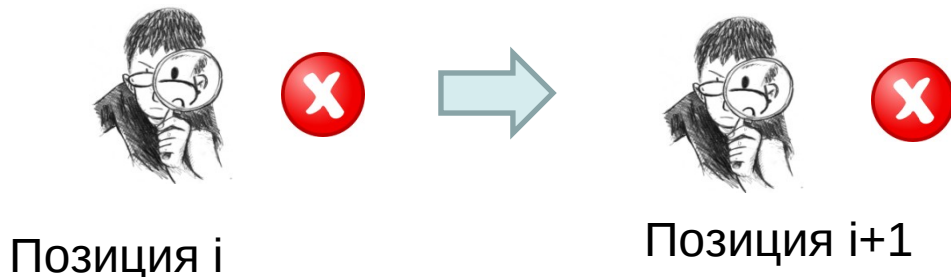
# Гипотеза о «наблюдении» (Richardson и др. 2007)

- Вероятность клика делится на две части
  - Глобальный компонент: вероятность увидеть – зависит от позиции документа
  - Локальный компонент: зависит от пары (запрос, документ)
- Это основа любой современной модели



# Каскадная модель (Craswell et al., 2008)

- Первый документ всегда просматривается
- Дальше модель Маркова
  - Просмотр на позиции  $i+1$  зависит от просмотра и клика на позиции  $i$
- Просмотр идет линейно





# Каскадная модель

- Объединяем две гипотезы:

**Cascade  
Model** =

[Craswell+08]



- Формальная спецификация модели:

$$- P(C_i=1|E_i=0) = 0, P(C_i=1|E_i=1) = r_{u_i}$$

$$- P(E_1=1) = 1, P(E_{i+1}=1|E_i=0) = 0$$

$$- P(E_{i+1}=1|E_i=1, C_i=0)=1$$

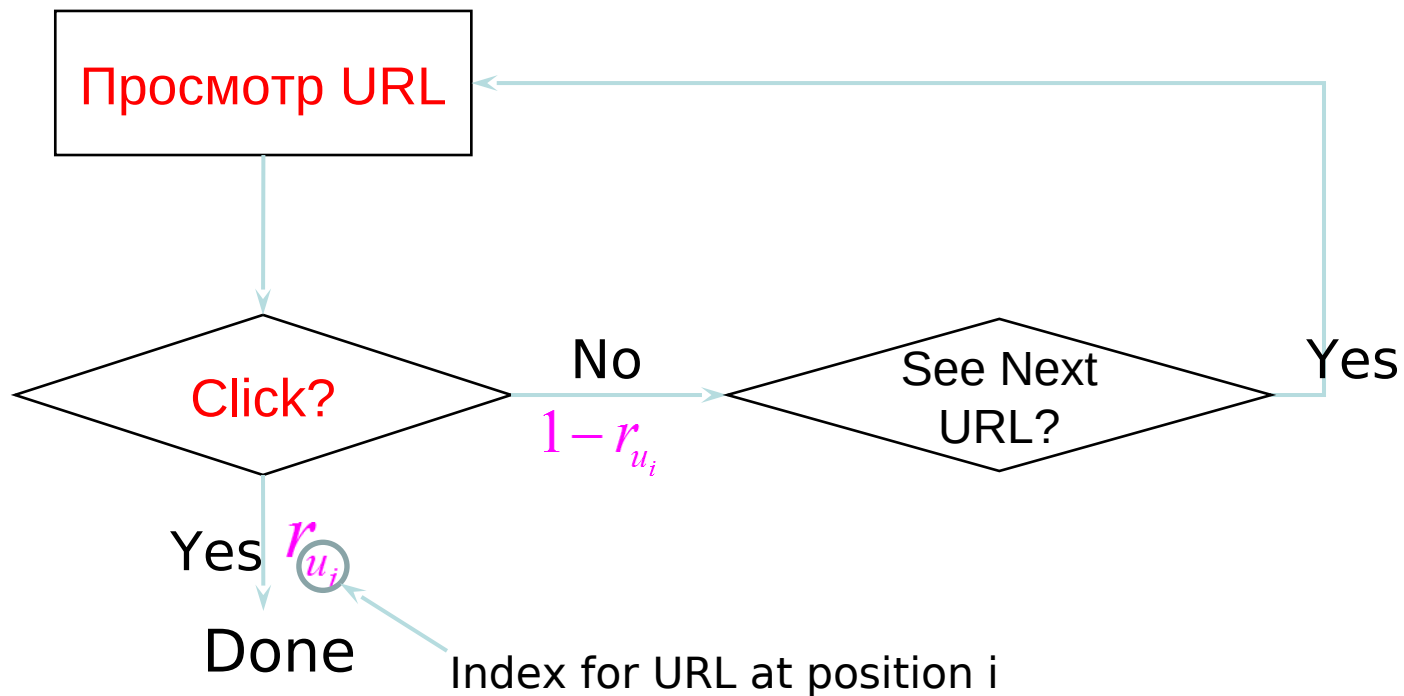
*Гипотеза просмотра*

*Каскадная гипотеза*

*Моделирование клика*

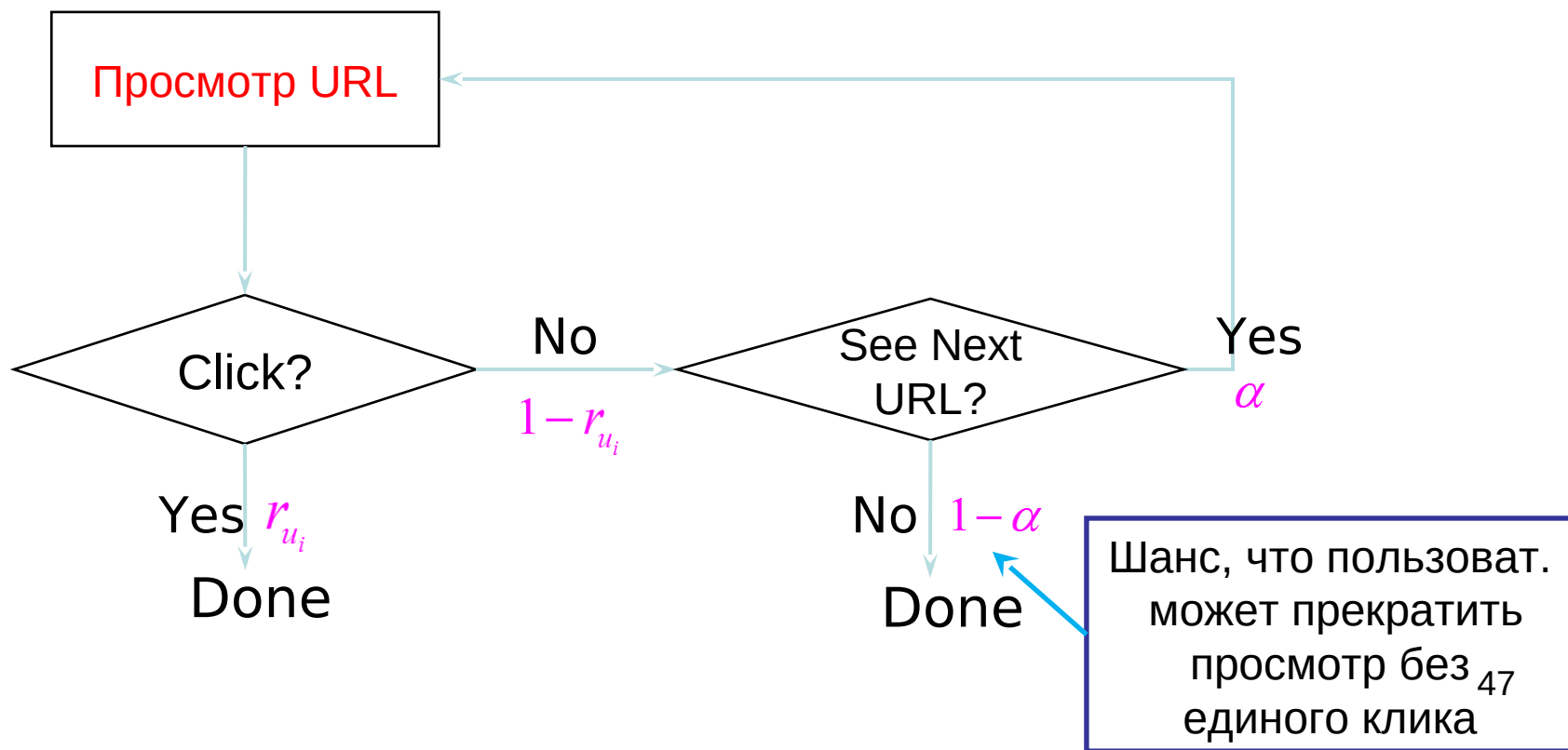
# Каскадная модель

- Блок схема поведения пользователя:



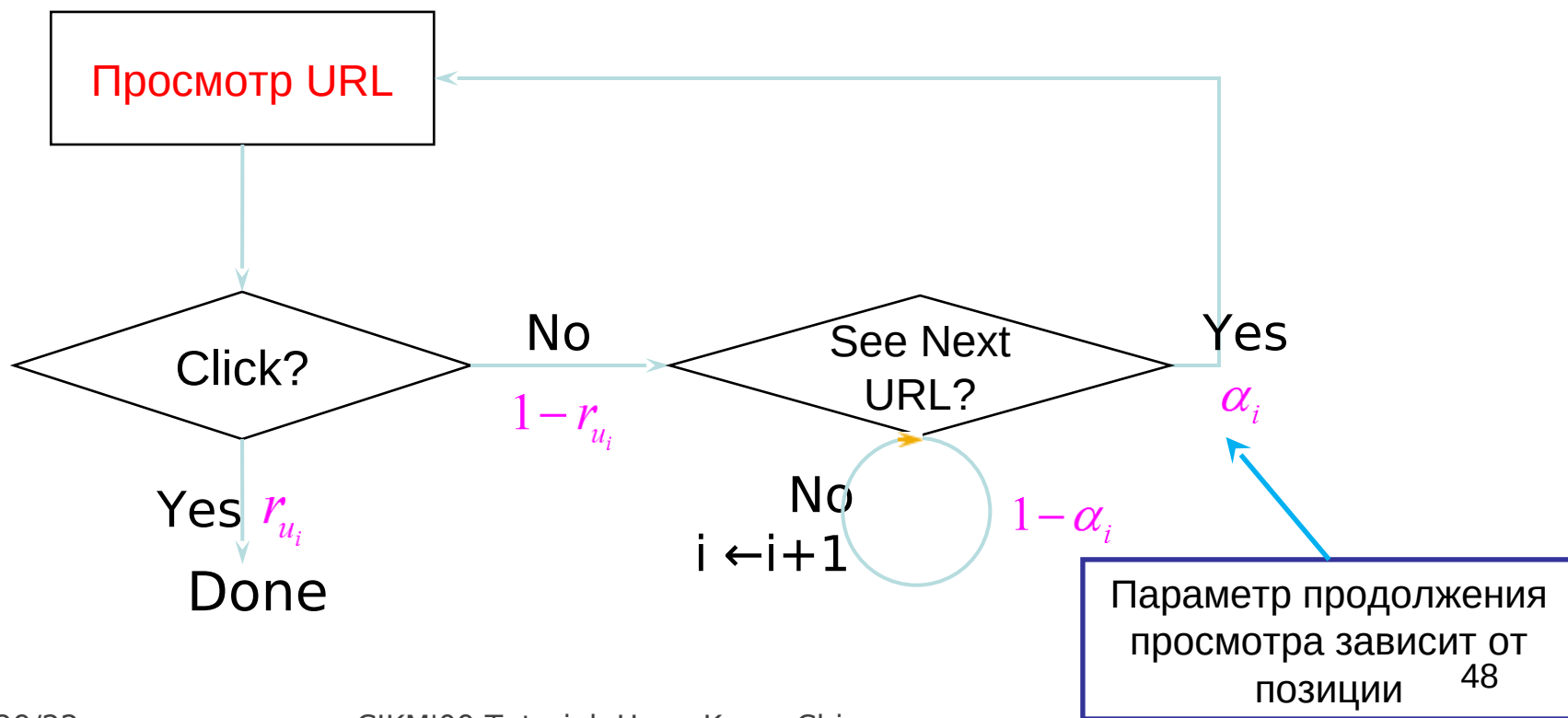
# Альтернативы

- Первый клик в **Click Chain Model** [Guo+09b] и **Dynamic Bayesian Network** model [Chapelle+09]



# Альтернативы

- Первый клик в **User Browsing Model** [Dupret+08]



# Моделирование нескольких кликов (Guo et al., 2009)

- Обобщение каскадной модели для 1+ КЛИКОВ:

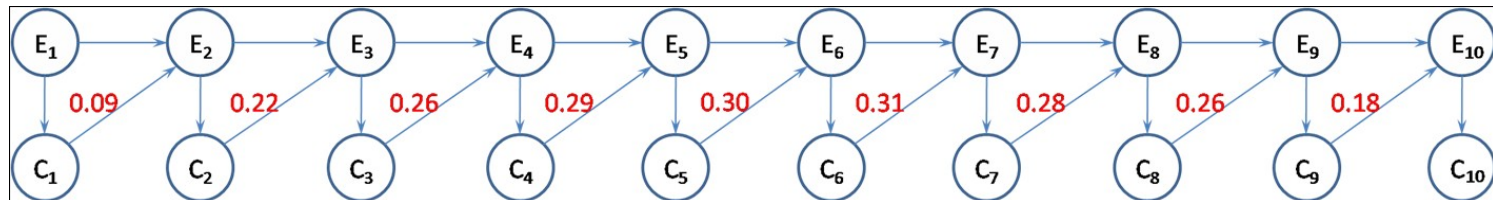
–  $P(C_i=1|E_i=0) = 0$ ,  $P(C_i=1|E_i=1) = r_{u_i}$

–  $P(E_1=1) = 1$ ,  $P(E_{i+1}=1|E_i=0) = 0$

–  $P(E_{i+1}=1|E_i=1, C_i=0)=1$

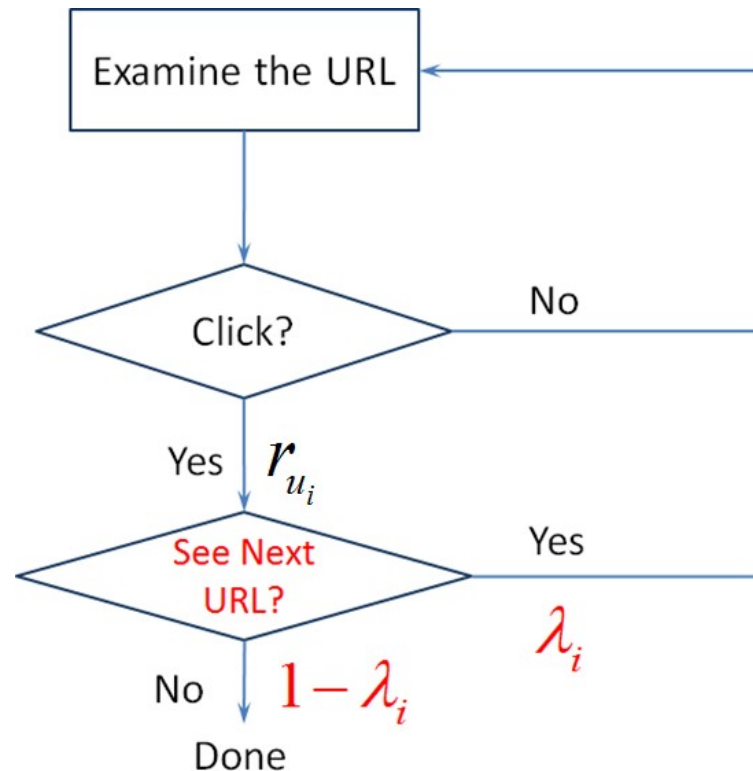
–  $P(E_{i+1}=1|E_i=1, C_i=1)=\lambda_i$

$\lambda$ : глобальные параметры,  
характеризующие поведение  
пользователя



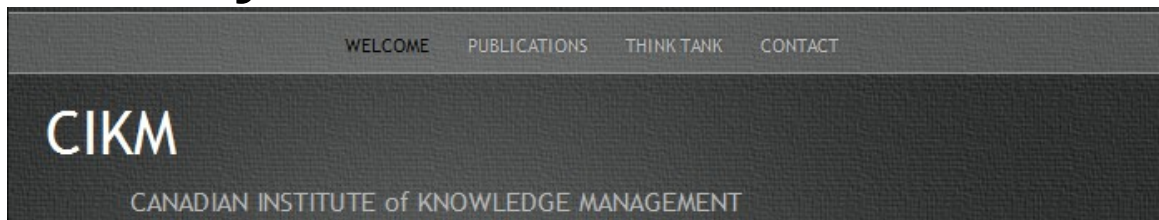
# Моделирование нескольких кликов (Guo et al., 2009)

- Обобщение каскадной модели для 1+ КЛИКОВ:



# Click chain model (Guo et al. 2009)

- Вероятность просмотра следующего документа зависит от релевантности кликнутого URL:



Нет, не то.  
Нужно  
проверить  
следующий

Да, ОК.  
Конец  
просмотра

# Click chain model (Guo et al. 2009)

- Вероятность просмотра следующего документа зависит от релевантности кликнутого URL:

- :

$$- P(E_{i+1}=1|E_i=1, C_i=1) = \alpha_2(1-r_{u_i}) + \alpha_3 r_{u_i}$$

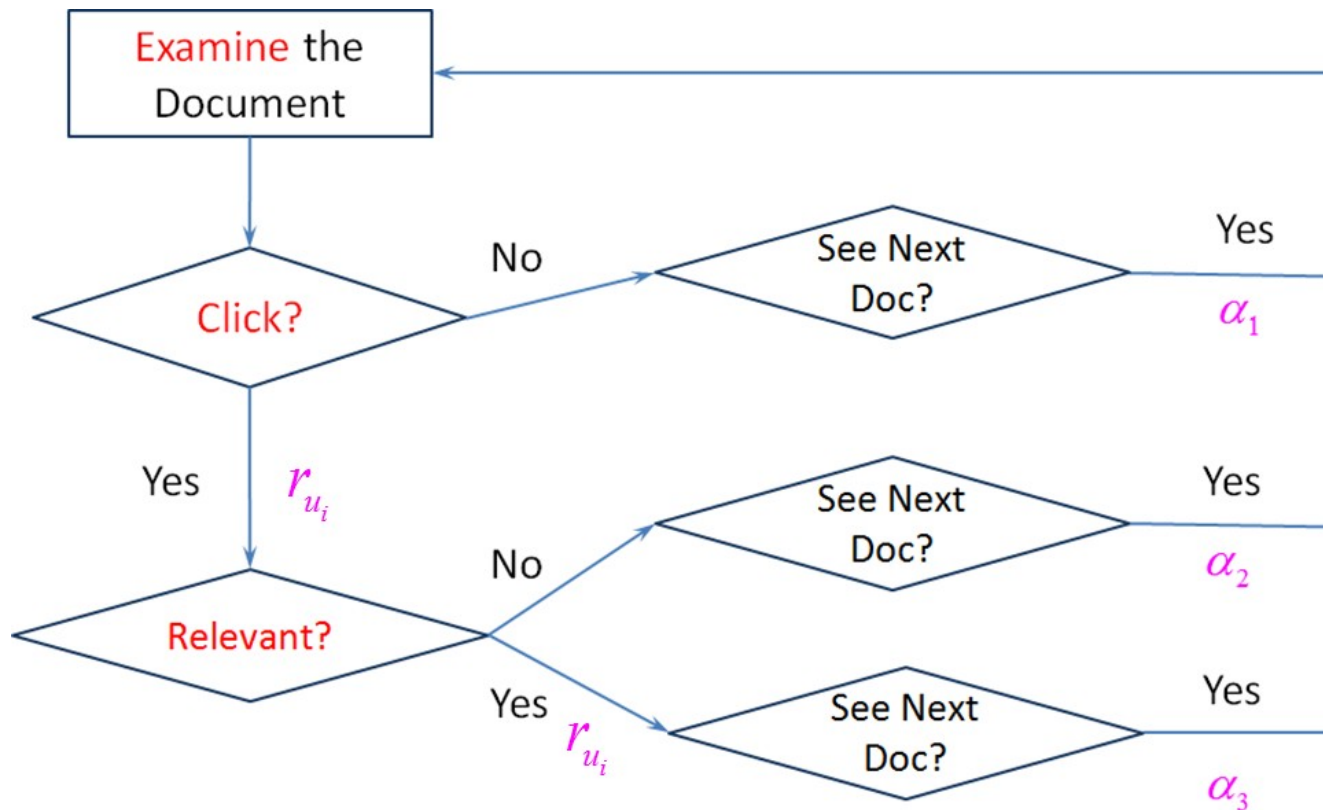
$$- P(E_{i+1}=1|E_i=1, C_i=0) = \alpha_1$$

where  $0 < \alpha_1 \leq 1$ ,  $0 \leq \alpha_3 < \alpha_2 \leq 1$



# Click chain model

- Полная картина:



# User browsing model (Dupret et al., 2008)

- Вероятность просмотра следующего документа зависит от: ранее кликнутой позиции  $r$ , and расстояния до позиции  $d$ .

$r =$   
0  
 $d =$   
1



Position n	URL	Click
1	cikm2008.org	0
2	www.cikm.org	1
3	www.cikm.org/2002	0
4	www.fc.ul.pt/cikm2007	0
5	cikmconference.org	0
6	www.comp.polyu.edu.hk/...	1
...	...	...

# User browsing model

Вероятность просмотра следующего документа зависит от: ранее кликнутой позиции  $r$ , and расстояния до позиции  $d$ .

$r =$   
0  
 $d =$   
2

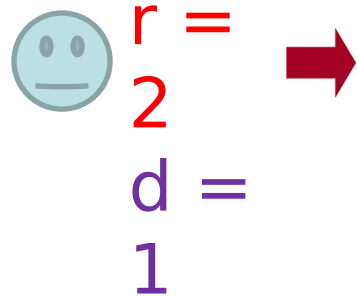


Position $n$	URL	Click
1	cikm2008.org	0
2	www.cikm.org	1
3	www.cikm.org/2002	0
4	www.fc.ul.pt/cikm2007	0
5	cikmconference.org	0
6	www.comp.polyu.edu.hk/...	1
...	...	...

55

# User browsing model


Вероятность просмотра следующего документа зависит от: ранее кликнутой позиции  $r$ , and расстояния до позиции  $d$ .



Position $n$	URL	Click
1	cikm2008.org	0
2	www.cikm.org	1
3	www.cikm.org/2002	0
4	www.fc.ul.pt/cikm2007	0
5	cikmconference.org	0
6	www.comp.polyu.edu.hk/...	1
...	...	...

# User browsing model

Вероятность просмотра следующего документа зависит от: ранее кликнутой позиции  $r$ , and расстояния до позиции  $d$ .

  $r = 2$   
 $d = 2$




Position $n$	URL	Click
1	cikm2008.org	0
2	www.cikm.org	1
3	www.cikm.org/2002	0
4	www.fc.ul.pt/cikm2007	0
5	cikmconference.org	0
6	www.comp.polyu.edu.hk/...	1
...	...	...

57

# User browsing model

Вероятность просмотра следующего документа зависит от: ранее кликнутой позиции  $r$ , and расстояния до позиции  $d$ .

  $r = 2$   
 $d = 3$  →

Position $n$	URL	Click
1	cikm2008.org	0
2	www.cikm.org	1
3	www.cikm.org/2002	0
4	www.fc.ul.pt/cikm2007	0
5	cikmconference.org	0
6	www.comp.polyu.edu.hk/...	1
...	...	...

# User browsing model

Вероятность просмотра следующего документа зависит от: ранее кликнутой позиции  $r$ , and расстояния до позиции  $d$ .

- пользователи могут потерять терпение, пока они ищут релевантный документ без кликов.
- Вероятность клика монотонно убывает с увеличением  $d$  при условии сохраняющего  $r$ .

# User browsing model

Вероятность просмотра следующего документа зависит от: ранее кликнутой позиции  $r$ , and расстояния до позиции  $d$ .

$$P(E_i=1|C_{1:i-1}) = \beta_{r_i, d_i}$$

Где  $r_i = \max\{j | j < i, C_j=1\}$ ,  $d_i$

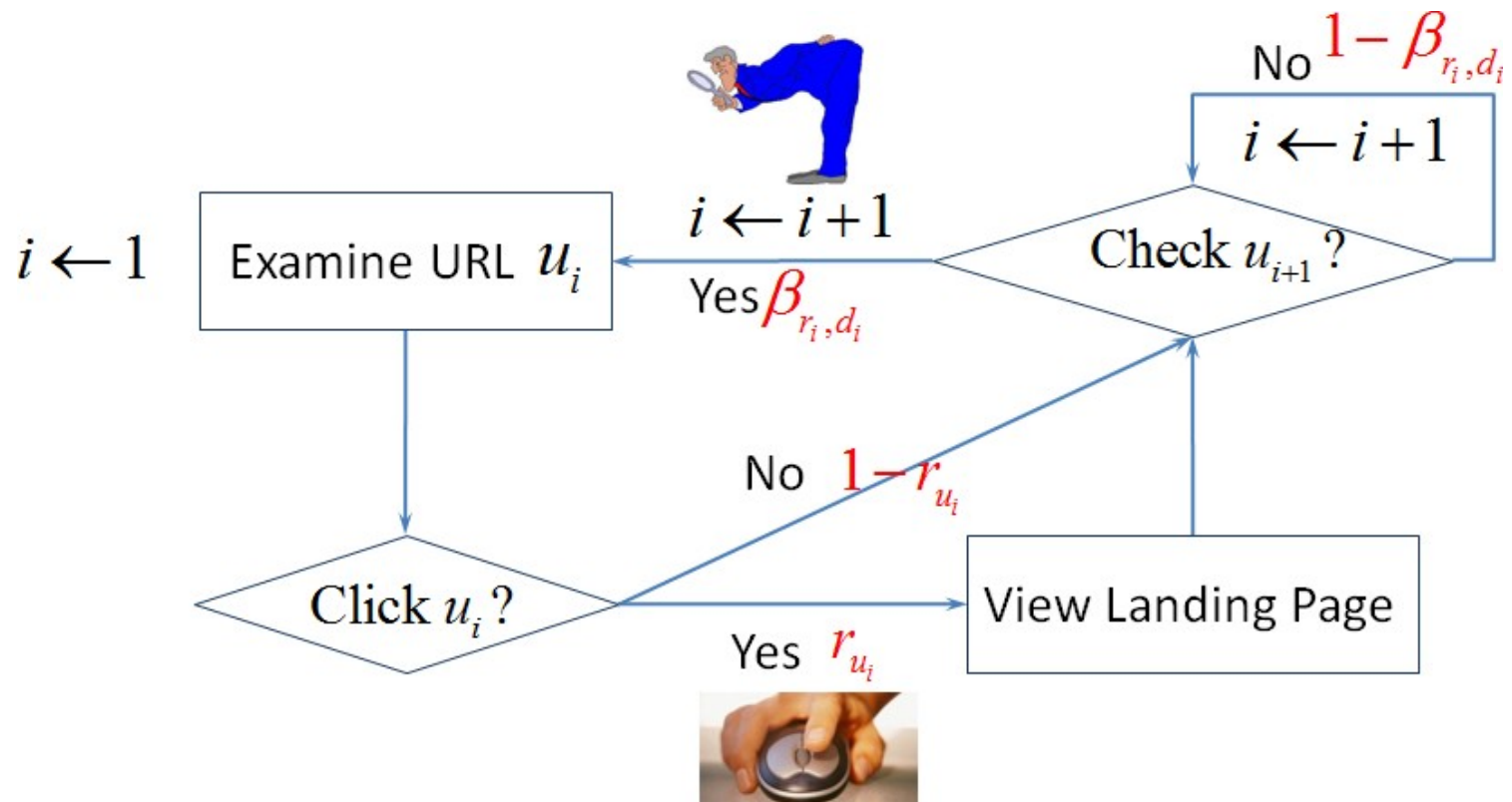
– 55 параметров для top-10 позиций  
( $0 \leq r < r+d \leq 10$ ).

– Каскадная модель не предполагается



# User browsing model

- Полная картина:



# Заключение: Кликовые модели

- Статистические модели учесть клики пользователя как feedback
- Различные модели делают различные предположения о закономерностях поведения пользователей
- Модели должны «бороться» с неравноценностью позиций