21. Метод «Пирамиды»-1Pyramid Evaluation
•Разработанв 2005 году Колумбийским университетом.

•Эксперты выделяют из «эталонных» аннотаций «информационные единицы» -Summary Content Units (SCUs).

•Каждый SCU получает вес, равный количеству «эталонных» аннотаций, где она встречалась.

•Оценка –суммарный вес входящихSCU.

•Неоднократное вхождение SCU в автоматическую аннотацию не поощряется.
Метод «Пирамиды»-2
•Итоговый результат:

[Сумма весов SCUв авт.аннотациях]
[Сумма весов SCU в экспертных аннотациях]
•Пример **SCU:**

Мини-субмарина попала в ловушку под водой.
1.мини-субмарина... была затоплена... на дне моря...

2.маленькая... субмарина... затоплена... на глубине 625 футов.

3.мини-субмарина попала в ловушку... ниже уровня моря.

4.маленькая... субмарина... затоплена... на дне морском...

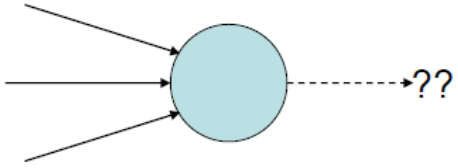
Метод «Пирамиды»:
+ Наиболее объективная оценка содержания аннотации
-Отсутствие оценки читабельности, большое участие человека

22. Анализ ссылок. PageRank

— это числовая величина, характеризующая «важность» веб-страницы. Чем больше ссылок на страницу, тем она «важнее». Кроме того, «вес» страницы A определяется весом ссылки, передаваемой страницей B. Таким образом, PageRank — это метод вычисления веса страницы путём подсчёта важности ссылок на неё.

Вес Pagerank
•Представим, что пользователь случайно бродит по страницам:
–Начинает на случайной странице

–На каждом шагу переходит на следующую по ссылку с равной вероятностью
•В пределе каждая страница получит рейтинг посещений –можно использовать как вес страницы



«Телепортация»

•В тупиковой странице –переход на случайную страницу

•Для любой не тупиковой страницы -с вероятностью10%, переходим на случайную страницу
–С оставшейся вероятностью(90%)–переход по одной из исходящих ссылок (с равной вероятностью)

–10% -параметр
•Теперь можно говорить о посещаемости страницы как о ее рейтинге

•Как можно посчитать такой рейтинг?
ФормулаPageRank

•**PR(A) = d + (1-d)(PR(T1)/C(T1) +…+PR(Tn)/C(Tn))**

•PR(Tn)–исходная значимость страницы
•C(Tn)–количество исходящих ссылок со страницы
•PR(Tn)/C(Tn) –значимость страницы равномерно распределяется по исходящим ссылкам и переносится в значимость страницы A по входящим в нее ссылкам
•*n* — **количество страниц**, ссылающихся на страницу-акцептор (на которые не наложен фильтр);
•d–например, 0.15 значимость страницы, без входящих ссылок (коэффициент телепортации)
•(1-**d)(…)**–0.85

23. Анализ ссылок: HITS

В ответ на запрос вместо упорядоченного множества страницнайдем два множества взаимосвязанных страниц:
–Hub pages–*хорошие списки ссылок по теме*.
•“Bob’s list of cancer-related links.”
–Authority pages–часто упоминаются на страницах хабов
•Хорошо работает на широких тематических запросах

Авторитетный документ (авторитетная страница, автор) — это документ, соответствующий запросу пользователя, имеющий больший удельный вес среди документов данной тематики, то есть большее число документов ссылаются на данный документ.

Хаб-документ (хаб-страница, посредник) — это документ, содержащий много ссылок на авторитетные документы.

Первым шагом в алгоритме HITS, является получение наиболее релевантных страниц в поисковом запросе. Это множество называется корневым набором и может быть получено путём принятия самых популярных страниц n, возвращаемых текстовым алгоритмом поиска. Базовый набор формируется путём увеличения корневого набора со всеми веб-страницами, которые с ним связаны и с некоторыми страницами, ссылающихся на него. Веб-страницы в базовом наборе и все гиперссылки между этих страниц, образуют сосредоточенный подграф. HITS вычисления выполняются только на этом подграфе.

Оценки авторитетного документа и посредника определены в терминах друг друга во взаимной рекурсии. Оценка авторитетности страницы вычисляется как сумма значений оценок посреднических страниц, которые указывают на эту страницу. Значение оценки посредника вычисляется как сумма оценок авторитетных страниц, на которые он указывает.

Алгоритм выполняет ряд итераций, каждая из которых состоит из двух основных этапов:

- **Обновление авторитетности**. Обновление авторитетной оценки каждой вершины подграфа, эквивалентное сумме посреднических оценок каждой из вершин, указывающих на них.
- **Хаб-обновление**. Обновление посреднической оценки каждой вершины подграфа, путём суммирования авторитетных оценок каждой из вершин, на которые они указывают.

Оценка авторитетности и посредническая оценка для вершины рассчитывается по следующему алгоритму:

- Начните с вершин, оценка авторитетности и посредническая оценка которых равна 1.
- Выполнение правила обновления авторитетности.
- Выполнение правила хаб-обновления.
- Нормализация значений путём деления каждой посреднической оценки на корень квадратный из суммы квадратов всех посреднических оценок, и деления каждой оценки авторитетности на корень квадратный из суммы квадратов всех оценок авторитетности.
- Повторение со второго шага по мере необходимости.

Основная схема
•Извлечь исходное множество(базовый набор) потенциально хороших хабов или авторитетов
•Из них формируем небольшой топ-лист хабов или авторитетов
–Итеративный алгоритм
Базовый набор страниц
•Дан текстовый запрос (например, ***браузер***), получаем страницы, содержащие слово ***браузер.***
–Это **корневой набор** страниц.
•Добавляем любую страницу, которая
–указывает на страницу из корневого набора или

–На которую есть ссылка со страницы корневого множества.
•Это **базовый** набор

Разделение хабов и авторитетов
•Вычисляем для каждой страницы в базовом наборehub score*h(x)*иauthority score*a(x).*

•Инициализация: для всех *x, h(x)*□*1; a(x)* □*1*;

•Итеративно пересчитываем*h(x), a(x)*;

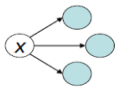
•В результате итераций
–Выдать страницы с наивысшими*h()*как топ-хабы

–С наивысшими*a()*scores как топ-авторитеты

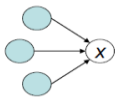
Итеративный пересчет

- Повторяем следующий пересчет для всех *x*:

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$



$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$



Итеративный пересчет
•Таким образом,
–оценка авторитетности страницы вычисляется как сумма значений оценок посреднических страниц, которые указывают на эту страницу.

–посредническая оценка страницы вычисляется как сумма значений оценок авторитетности страниц, на которых она ссылается.
•Рост значений авторитетности и посредника –необходима нормализация.

•Значения, полученные в результате этого процесса, в конечном итоге сходятся.

•Обычно требуется около 5 итераций

Недостатки HITS
•Сдвиг темы (Topic drift)
–Нерелевантные документы могут вызвать сдвиг темы
•Нерелеватные страницы на первых позициях выдачи
–приводят к ошибочным результатам
•Взаимное усиление страниц, ссылающихся друг на друга

•Поисковая оптимизация SEO: создание искусственного множества ссылок

HITS vs. PageRank

•Алгоритм HITS вычисляет не только ранг каждого узла, но также дает посредническую оценку.

•Алгоритм PageRank содержит свободный параметр α, который обычно не включен в алгоритм HITS.

•Приоритетом, в результате работы алгоритма PageRank, пользуются, как правило, более старые ресурсы, в то время как HITS алгоритм имеет меньший уклон в этом отношении.

•Алгоритм PageRank может находить единственное уникальное решение

24. Особенности использования кликов пользователя в качестве фидбека от пользователя. Каскадная модель при обработке кликов

•Клики –это хорошо…

–Одинаково ли хороши?

•Отсутствие кликов может объясняться:

–Не релевантно

–Не видел

Неравноценность позиций относительно кликов

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

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

- Документ должен быть прочитан перед кликом.
- Условная вероятность клика после прочтения зависит от релевантности документа
- 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

•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+o8]

- Формальная спецификация модели:
  - $P(C_i=1|E_i=0) = 0$, $P(C_i=1|E_i=1) = r_{u_i}$
  - $P(E_1=1) = 1$, $P(E_{i+1}=1|E_i=0) = 0$
  - $P(E_{i+1}=1|E_i=1, C_i=0)=1$

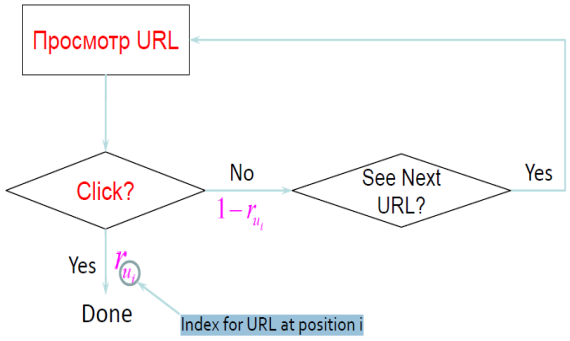
Гипотеза просмотра

Каскадная гипотеза

Моделирование клика

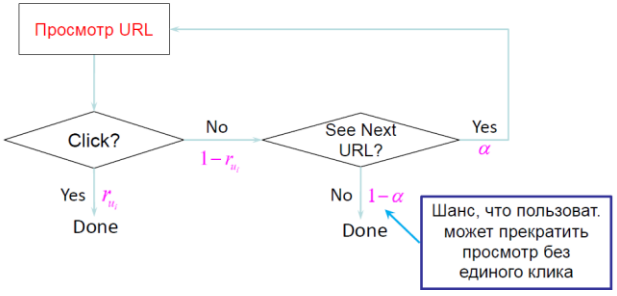
Каскадная модель

- Блок схема поведения пользователя:



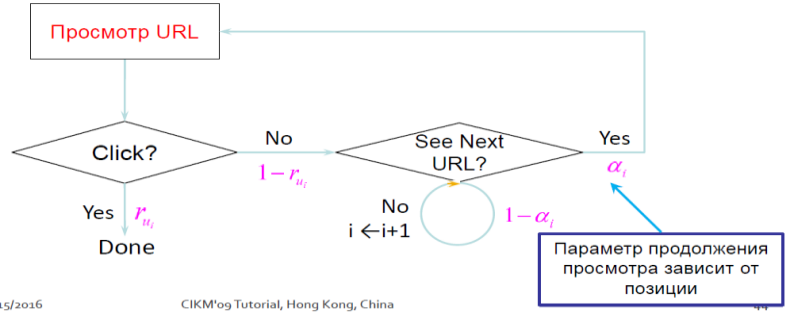
Альтернативы

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



Альтернативы

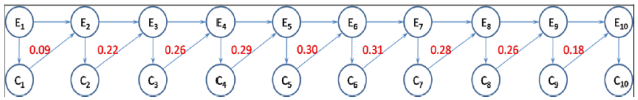
- Первый клик в User Browsing Model [Dupret+08]



Моделирование нескольких кликов (Guo et al., 2009)

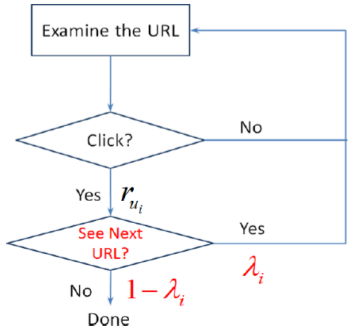
- Обобщение каскадной модели для 1+ кликов:
  - $P(C_i=1|E_i=0) = 0$, $P(C_i=1|E_i=1) = r_{u_i}$
  - $P(E_1=1) = 1$, $P(E_{i+1}=1|E_i=0) = 0$
  - $P(E_{i+1}=1|E_i=1, C_i=0)=1$
  - $P(E_{i+1}=1|E_i=1, C_i=1)= \lambda_i$

λ:глобальные параметры, характеризующие поведение пользователя



Моделирование нескольких кликов (Guo et al., 2009)

- Обобщение каскадной модели для 1+ кликов:



25. Классификация запросов по цели. Зачем нужна. Особенности обработки разных типов запросов

По целям пользователей:

Классификация запросов по цели во многом определяет характер страницы результатов поиска. По общим запросам будет получена неоднородная выдача, по транзакционным — более точная информация, например, карта вашего города с отметкой ближайшего ресторана.

- **Информационные** – запросы, по которым ищут различного рода информацию, например, «как приготовить плов» или «что такое франшиза». Соответственно на первый запрос мы получаем кучу рецептов приготовления плова, а на второй – определение слова «франшиза».

Такие запросы, как правило, используют контентные сайты, которые заинтересованы в получении большего трафика.

- **Навигационные** – запросы, по которым люди ищут конкретный сайт, сервис или компанию. Например: «официальный сайт компании Тойота».

Если ваш ресурс будут часто находить, используя навигационные запросы, то это говорит о том, что сайт хорошо раскручен и в него есть своя целевая аудитория.

- **Транзакционные** – запросы, задаваемые с целью совершить какое-либо действие (скачать фильм, музыку, фото, программу; купить или заказать какой-нибудь товар или услугу). Примеры транзакционных запросов: «купить ноутбук», «заказать кухню».

На данный момент транзакционные запросы характеризуются очень большой конкуренцией, это связано с тем, что их, как правило, используют коммерческие сайты, которые занимаются продажей каких-либо товаров или услуг. Все дело в том, что данные запросы имею достаточно хорошую конверсию, то есть пользователи, которые пришли по таким запросам, чаще что-то покупают или заказывают.