ТЕХНОСФЕРА

Поведенческое ранжирование 2

Владимир Гулин

7 апреля 2018 г.

План лекции

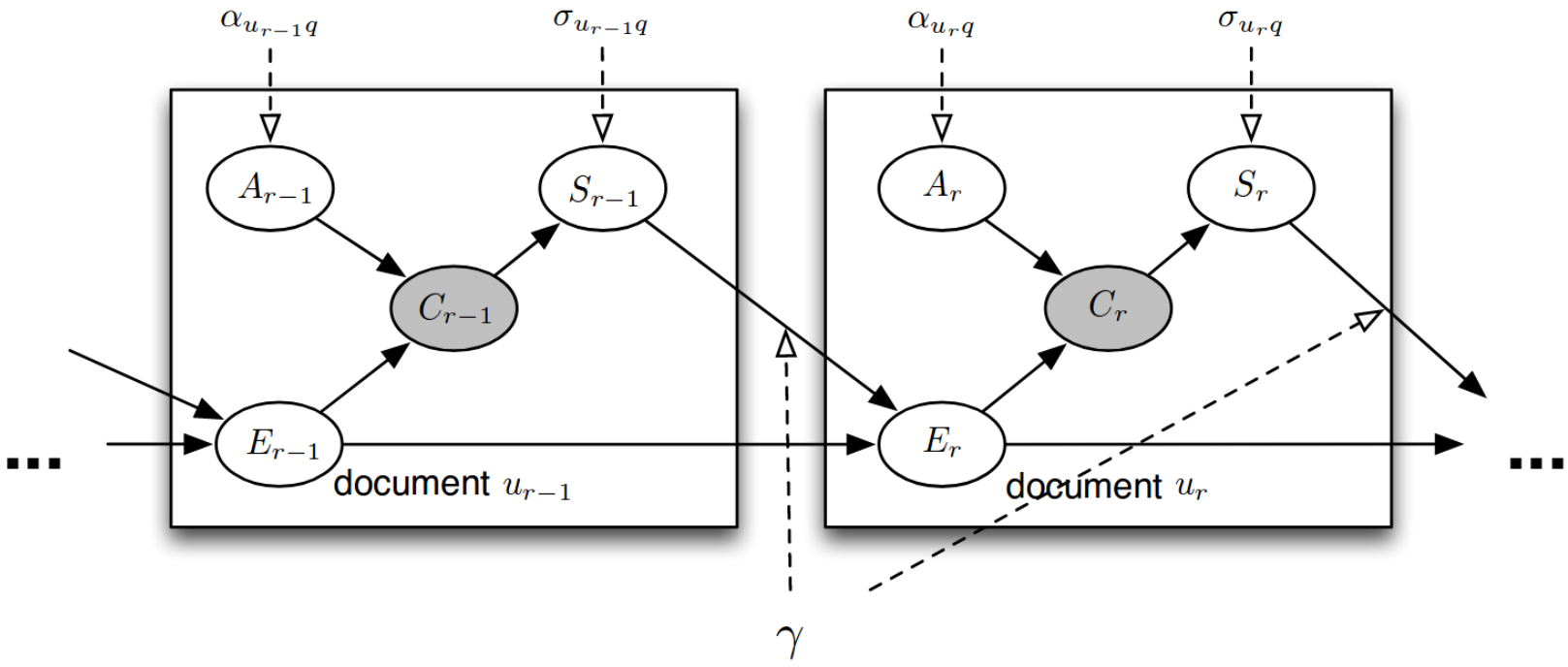
Базовые кликовые модели: выводы

- **CTR model:** подсчет кликов
- **Position-based model:** просмотры и привлекательность документа
- **Cascade model:** имеют значения предыдущие просмотры и клики
- **Dynamic Bayesian Network model:** удовлетворяемость документом
- **User browsing model:** ранки кликнутых документов

Click models

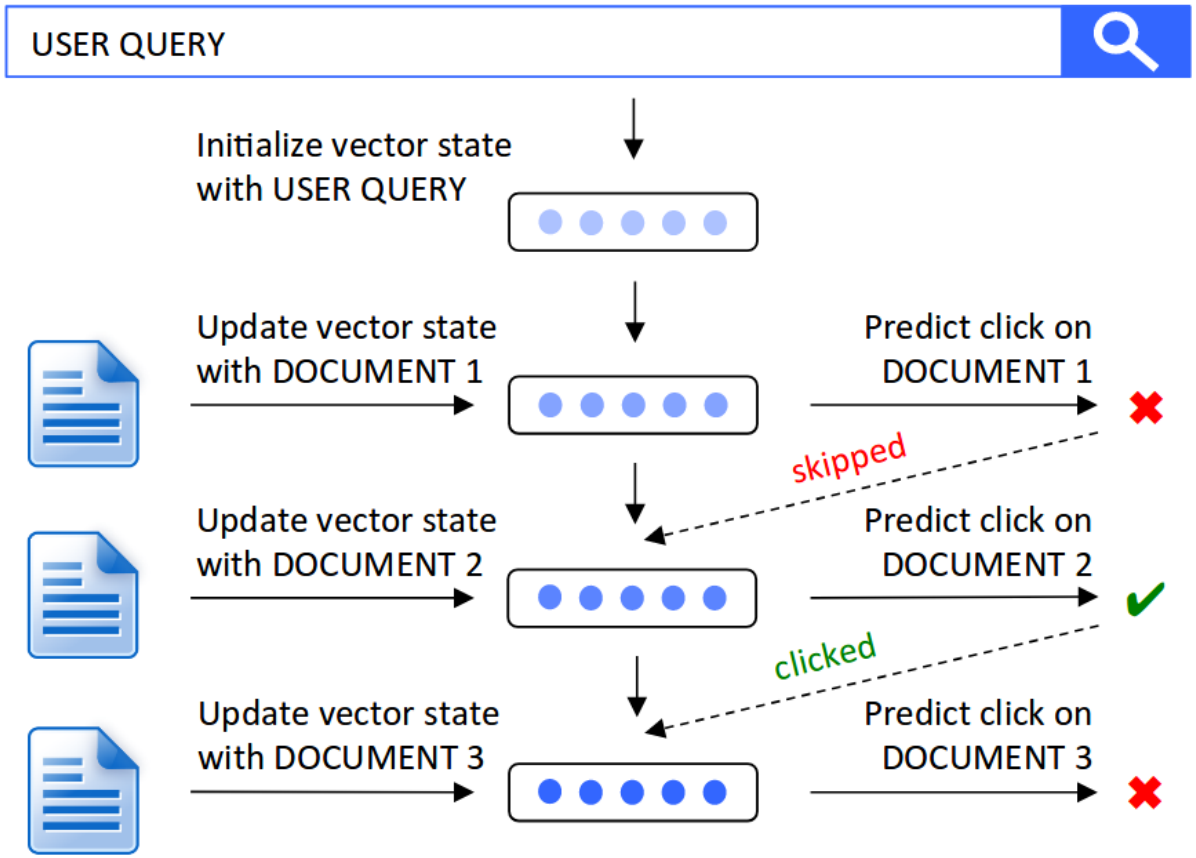
Click model	$P(C_u = 1)$	$P(C_u = 1 \mid \mathbf{C}_{<r_u})$
RCM	ρ	ρ
RCTR	ρ_{r_u}	ρ_{r_u}
DCTR	ρ_{uq}	ρ_{uq}
PBM	$\alpha_{uq}\gamma_{r_u}$	$\alpha_{uq}\gamma_{r_u}$
CM	$\alpha_{uq}\epsilon_{r_u}$, where $\epsilon_{r+1} = \epsilon_r(1 - \alpha_{u_r q})$	$\alpha_{uq}\epsilon_{r_u}$, where $\epsilon_r = \begin{cases} 1 & \text{if no clicks before } r \\ 0 & \text{otherwise} \end{cases}$
UBM	$\sum_{j=0}^{r_u-1} P(C_j = 1)\left(\prod_{k=j+1}^{r_u-1}(1 - \alpha_{u_k q}\gamma_{kj})\right)\alpha_{uq}\gamma_{r_u j}$, where $P(C_0 = 1) = 1$	$\alpha_{uq}\gamma_{rr'}$
DCM	$\alpha_{uq}\epsilon_{r_u}$, where $\epsilon_{r+1} = \epsilon_r\left(\alpha_{u_r q}\lambda_{r_u} + (1 - \alpha_{u_r q})\right)$	$\alpha_{uq}\epsilon_{r_u}$, where $\epsilon_{r+1} = c_r^{(s)}\lambda_r + \left(1 - c_r^{(s)}\right)\frac{(1-\alpha_{u_r q})\epsilon_r}{1-\alpha_{u_r q}\epsilon_r}$
CCM	$\alpha_{uq}\epsilon_{r_u}$, where $\epsilon_{r+1} = \epsilon_r\left(\alpha_{u_r q}((1 - \alpha_{u_r q})\tau_2 + \alpha_{u_r q}\tau_3) + (1 - \alpha_{u_r q})\tau_1\right)$	$\alpha_{uq}\epsilon_{r_u}$, where $\epsilon_{r+1} = c_r^{(s)}((1 - \alpha_{u_r q})\tau_2 + \alpha_{u_r q}\tau_3) + \left(1 - c_r^{(s)}\right)\frac{(1-\alpha_{u_r q})\epsilon_r\tau_1}{1-\alpha_{u_r q}\epsilon_r}$
DBN	$\alpha_{uq}\epsilon_{r_u}$, where $\epsilon_{r+1} = \epsilon_r\gamma(\alpha_{u_r q}(1 - \sigma_{u_r q}) + (1 - \alpha_{u_r q}))$	$\alpha_{uq}\epsilon_{r_u}$, where $\epsilon_{r+1} = c_r^{(s)}\gamma(1 - \sigma_{u_r q}) + \left(1 - c_r^{(s)}\right)\frac{(1-\alpha_{u_r q})\epsilon_r\gamma}{1-\alpha_{u_r q}\epsilon_r}$
SDBN	Same as DBN with $\gamma = 1$	Same as DBN with $\gamma = 1$

PGM-based click models



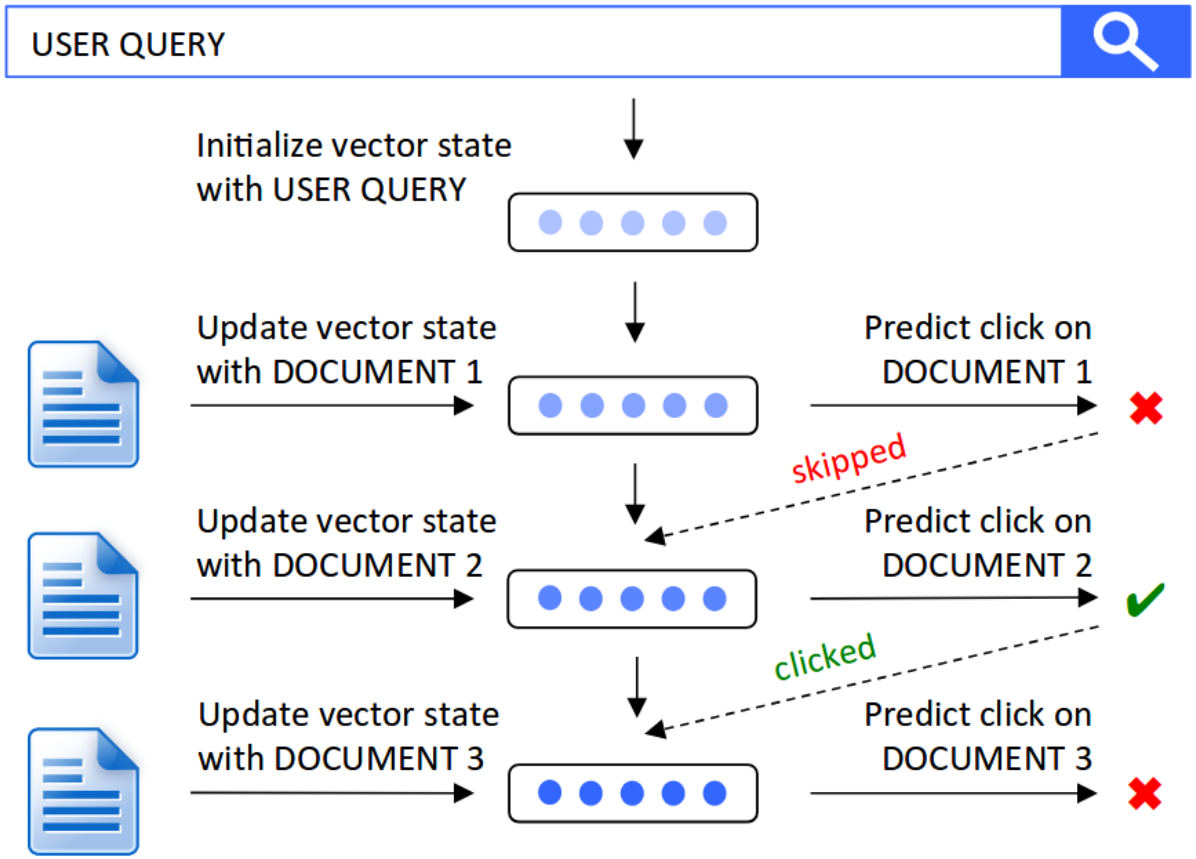
- Основаны на фреймворке probabilistic graphical model (PGM)
- Структура соответствующей PGM должна задаваться вручную

Альтернативный фреймворк



Выучиваем паттерны пользовательского поведения непосредственно из кликовых данных

Disributed representations $(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \ldots)$



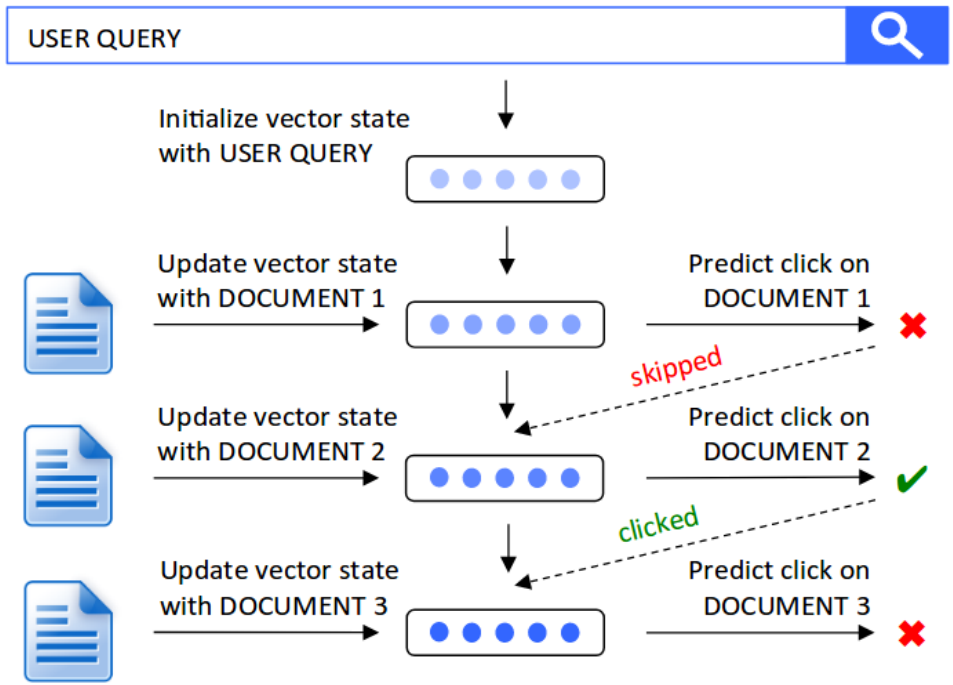
Моделируем пользовательское поведение в виде последовательности состояний, представленных векторами $(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \ldots)$, которые описывают информационную потребность пользователя и получаемую информацию в процессе поиска.

Описание модели

q – user query
d_r – document at rank r

i_r – user interaction
with document at rank r

$$\mathbf{s}_0 = \mathcal{I}(q)$$
$$\mathbf{s}_{r+1} = \mathcal{U}(\mathbf{s}_r, i_r, d_{r+1})$$



$$P(C_{r+1} = 1 \mid q, i_1, \ldots, i_r, d_1, \ldots, d_{r+1}) = \mathcal{F}(\mathbf{s}_{r+1})$$

Neural click modeling framework

Representations of q, d_r and i_r

- Используем 3 набора представлений: QD, QD+Q, QD+Q+D

Parametrization of $\mathcal{I}, \mathcal{U}$ and $\mathcal{F}$

- $\mathcal{I}$ Feed-forward neural network
- $\mathcal{U}$ Recurrent neural network (RNN, LSTM)
- $\mathcal{F}$ Feed-forward neural network
  (with one output and sigmoid activation)

Training

- Stochastic gradient descent (AdaDelta + gradient clipping)

Experimantal setup

Dataset

- Yandex relevance prediction dataset (WSCD 2012) (146,278,823 query sessions)

Evaluation

- Click prediction (log-likelihood, perplexity)
- Ranking (NDCG)

Baselines

- DBN, DCM, CCM, UBM

Evaluation

Likelihood

$$\mathcal{LL}(M) = \frac{1}{|S|} \sum_{s \in S} log\, P_M(C_1 = c_1^{(s)}, \ldots, C_n = c_n^{(s)})$$

- C_r - случайная бинарная величина, означающая клик по позиции r
- $c_r^{(s)}$ - наблюдаемый клик по позиции r в поисковой сессии s
- $P(C_r = c_r^{(s)})$ - вероятность пронаблюдать $c_r^{(s)}$ в сессии s
- $P_M(C_1 = c_1^{(s)}, \ldots, C_n = c_n^{(s)})$ - вероятность пронаблюдать серию $c_1^{(s)}, \ldots, c_n^{(s)}$ в сессии s.

Evaluation

Perplexity

Perplexity измеряет на сколько хорошо модель оценивает вероятность клика по кокнретной позиции

$$p_r(M) = 2^{-\frac{1}{S}\sum_{s \in S}(log_2 P_M(C_r^{(s)}=c_r^{(s)}))}$$

$$p_r(M) \in [1 \dots 2]$$

Ранжирование

$$DCG = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

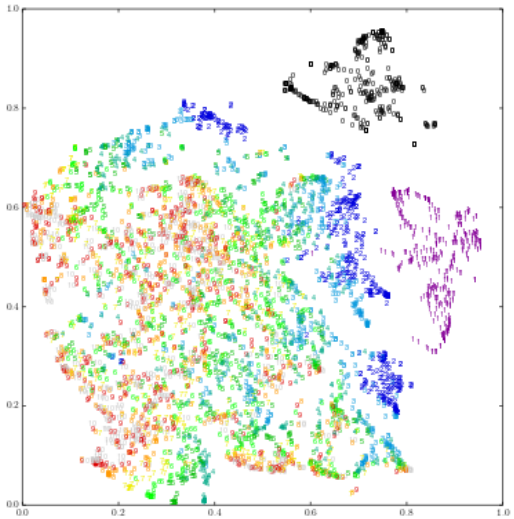
Click prediction

Click model	Perplexity	Log-likelihood
DBN	1.3510	−0.2824
DCM	1.3627	−0.3613
CCM	1.3692	−0.3560
UBM	1.3431	−0.2646
$\text{NCM}_{\text{QD}}^{\text{RNN}}$	1.3379	−0.2564
$\text{NCM}_{\text{QD}}^{\text{LSTM}}$	1.3362	−0.2547
$\text{NCM}_{\text{QD+Q}}^{\text{LSTM}}$	1.3355	−0.2545
$\text{NCM}_{\text{QD+Q+D}}^{\text{LSTM}}$	1.3318	−0.2526

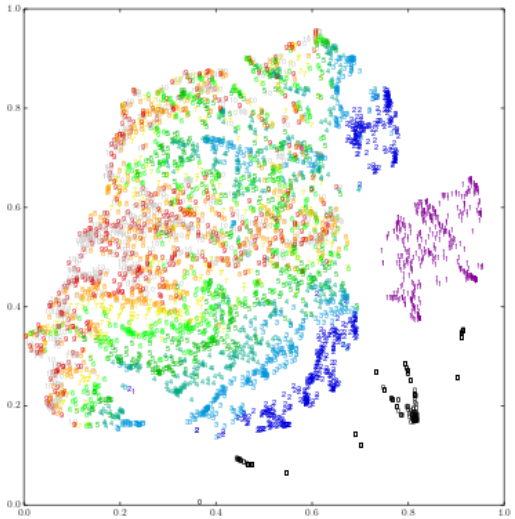
Ranking

Click model	NDCG @1	NDCG @3	NDCG @5	NDCG @10
DBN	0.717	0.725	0.764	0.833
DCM	0.736	0.746	0.780	0.844
CCM	0.741	0.752	0.785	0.846
UBM	0.724	0.737	0.773	0.838
$\text{NCM}^{\text{RNN}}_{\text{QD}}$	0.762	0.759	0.791	0.851
$\text{NCM}^{\text{LSTM}}_{\text{QD}}$	0.756	0.759	0.789	0.850
$\text{NCM}^{\text{LSTM}}_{\text{QD+Q}}$	0.775	0.773	0.799	0.857
$\text{NCM}^{\text{LSTM}}_{\text{QD+Q+D}}$	0.755	0.755	0.787	0.847

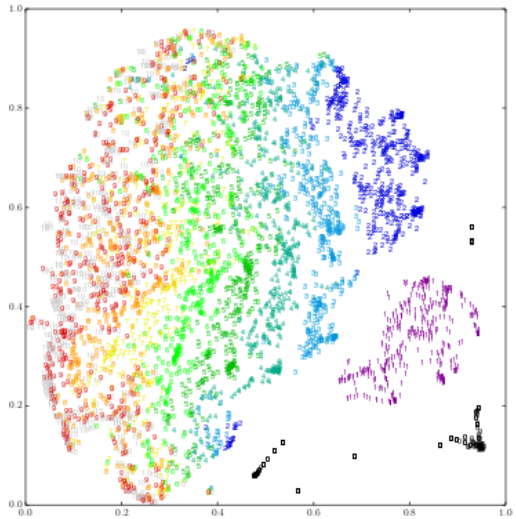
Анализ



(a) All query sessions.



(b) Query sessions generated by queries that occur one time in the training set.

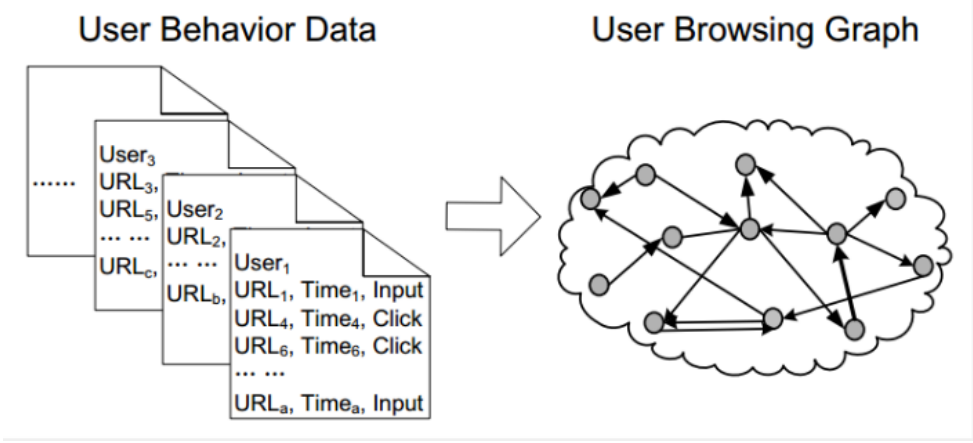


(c) Query sessions with no clicks generated by queries that occur one time in the training set.

Two-dimensional t-SNE projections of vector states s_r for different ranks r. Colors correspond to ranks: black 0; purple 1; dark blue 2; light blue 3; light blue-green 4; green 5; light green 6; yellow 7; orange 8; red 9; grey 10.

Поведенческие факторы

BrowseRank


User Behavior Data
User Browsing Graph
.....
User3
URL3,
URL5,
... ...
URLc,
User2
URL2,
... ...
URLb,
User1
URL1, Time1, Input
URL4, Time4, Click
URL6, Time6, Click
... ...
URLa, Timea, Input

Идентификация сессии

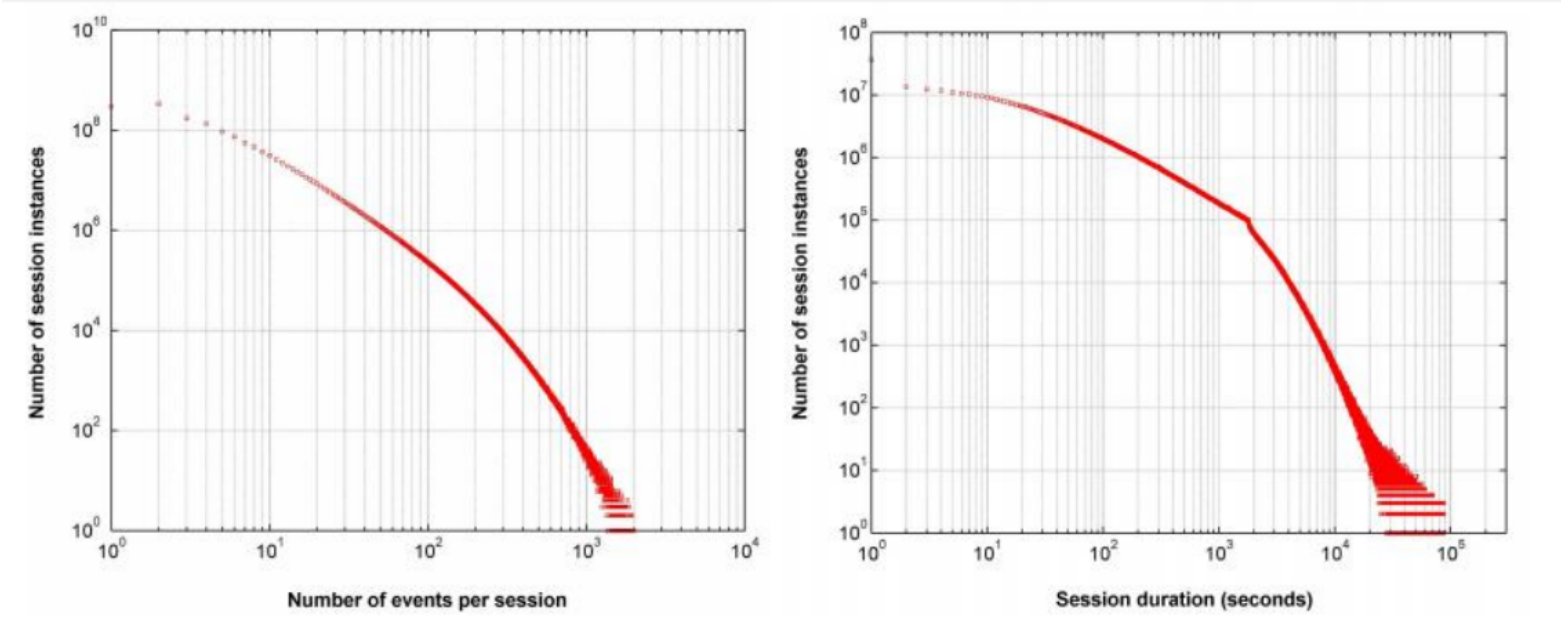
Событие:

- Cookie
- timestamp
- URL
- referral URL
- attribute

События объединяются в сессию:

1. По пользователю
2. referral URL(i) = URL(i - 1)
3. Время неактивности не должно превышать порог (30 мин.)

Характеристики сессий



- Среднее число событий в сессии: 9.1
- Средняя продолжительность сессии: 420.3 с
- Среднее число сессий на пользователя в день: 15.5
- Процент поисковых сессий: 4.85%

Local ClickRank

Опредилим local ClickRank как функцию

$$ClickRank(p_i, s_j) = \sum_{p_i \in s_j} w_r(i, s_j) w_t(p, s_j) I(p = p_i)$$

- The weight function $w_r(i, s_j)$ is computed from the rank of the page visit event p_i in session s_j
- The weight function $w_t(p, s_j)$ is computed from temporal information associated with browsing of the page p_i
- I - индикатор функция

ClickRank учитывает порядок кликов

Определим функцию взвешивания $w_r(i, s_j)$ для события i на позиции $r(i)$ в сессии s_j с общам числом событий n_j как

$$w_r(i, s_j) = \frac{2(n_j + 1 - r(i))}{n_j(n_j + 1)}$$

где $r(i) \in \{1, \ldots, n_j\}$

- $w_r(i, s_j)$ монотонно обывающая функция по $r(i)$
- $w_r(i-1, s_j) - w_r(i, s_j) = w_r(i, s_j) - w_r(i+1, s_j)$
- $\sum_{i=1}^{n_j} w_r(i, s_j) = 1$

ClickRank учитывает время

Определим функцию взвешивания $w_t(p, s)$ с учетом временной информации

$$w_t(p, s) = (1 - e^{-\lambda_1 t_d})e^{-\lambda_2 t_l} I(t(p) \in \mathcal{T})$$

где t_d и t_l это нормализованный dwelltime и время загрузки страницы
Индикатор функция $I(t(p) \in \mathcal{T})$ определяющая интересующий временной интервал

Global ClickRank

Имея набор сессий $S = (s_1, \ldots, s_k)$, global ClickRank может быть вычислен из local ClickRank функций путем агрегации

$$ClickRank(p, S) = AGGR_{s \in S}[ClickRank(p, s)]$$

Замечание:

При этом в качестве функции агрегации можно использовать помимо суммы, среднего еще и фильтрацию (например по времени или соц. дему)

Применение в поиске

Данные

3.3 миллиарда веб сессий извлеченных из Yahoo! тулбара за 6 месяцев 2008

Site ranking

$$ClickRank(w, S) = \sum_{p \in w} ClickRank(p, S)$$

Page ranking

Использование в качестве дополнительного фактора в ранжировании

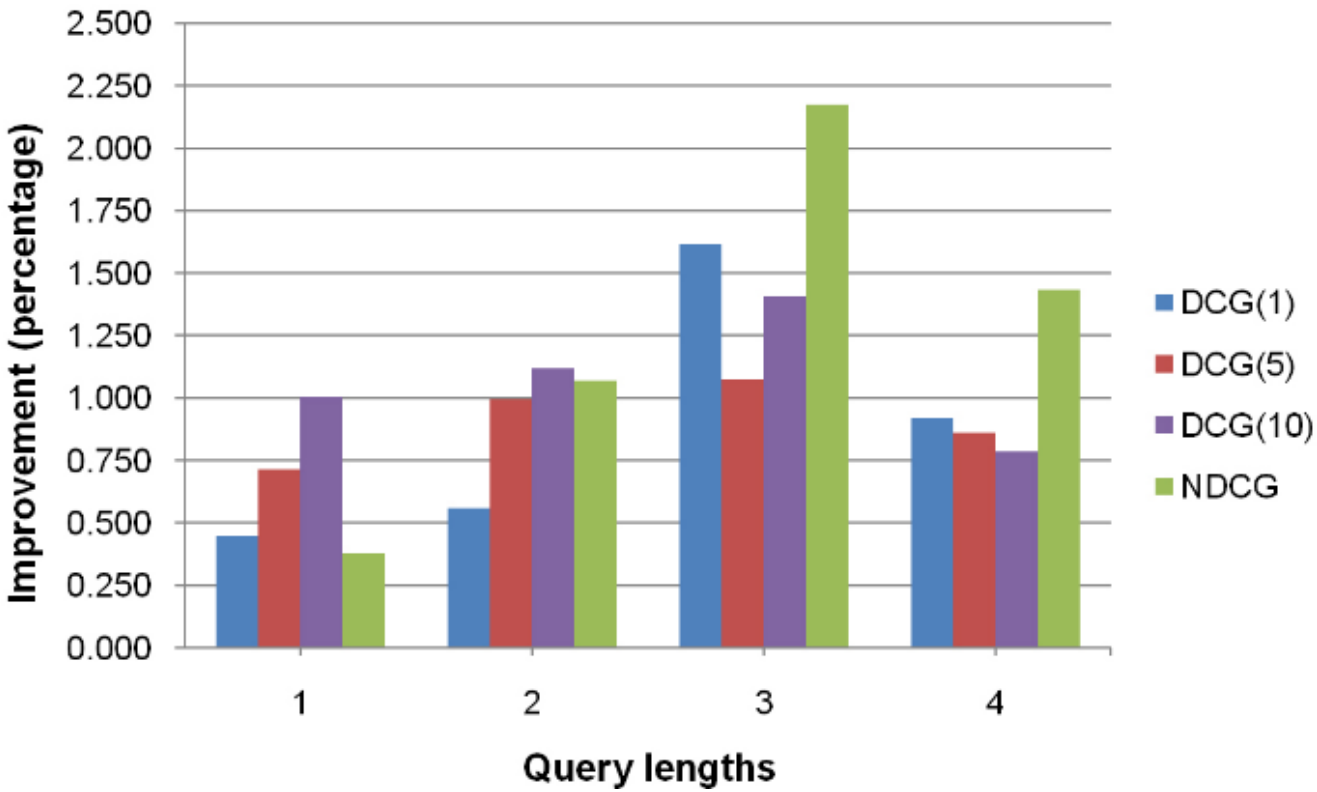
Site Ranking

Rank	PageRank	BrowseRank	ClickRank
1	adobe.com	myspace.com	yahoo.com
2	wordpress.com	msn.com	google.com
3	w3.org	yahoo.com	myspace.com
4	miibeian.gov.cn	youtube.com	live.com
5	statcounter.com	live.com	youtube.com
6	phpbb.com	facebook.com	facebook.com
7	baidu.com	google.com	msn.com
8	php.net	ebay.com	friendster.com
9	microsoft.com	hi5.com	pogo.com
10	mysql.com	bebo.com	aol.com
11	mapquest.com	orkut.com	microsoft.com
12	cnn.com	aol.com	wikipedia.org
13	google.com	friendster.com	ebay.com
14	blogger.com	craigslist.org	craigslist.org
15	paypal.com	google.co.th	hi5.com
16	macromedia.com	microsoft.com	go.com
17	jalbum.net	comcast.net	ask.com
18	nytimes.com	wikipedia.org	google.co.th
19	simplemachines.org	pogo.com	comcast.net
20	yahoo.com	photobucket.com	orkut.com

Top Ranked sites with different algorithms

Вклад в ранжирование

Query length	Number of queries	Affected queries	Improvements in				Significance test p-value
			DCG(1)	DCG(5)	DCG(10)	NDCG	
1	1484	1232	0.45%	0.71%	1.00%	0.38%	5.33×10^{-2}
2	2992	2450	0.56%	0.99%	1.12%	1.07%	4.65×10^{-4}
3	2153	1722	1.62%	1.08%	1.41%	2.18%	1.10×10^{-4}
4+	2412	1937	0.92%	0.86%	0.78%	1.43%	1.61×10^{-5}
All	9041	7341	1.02%	0.97%	1.11%	1.33%	9.98×10^{-5}



Сглаживание поведенческих данных

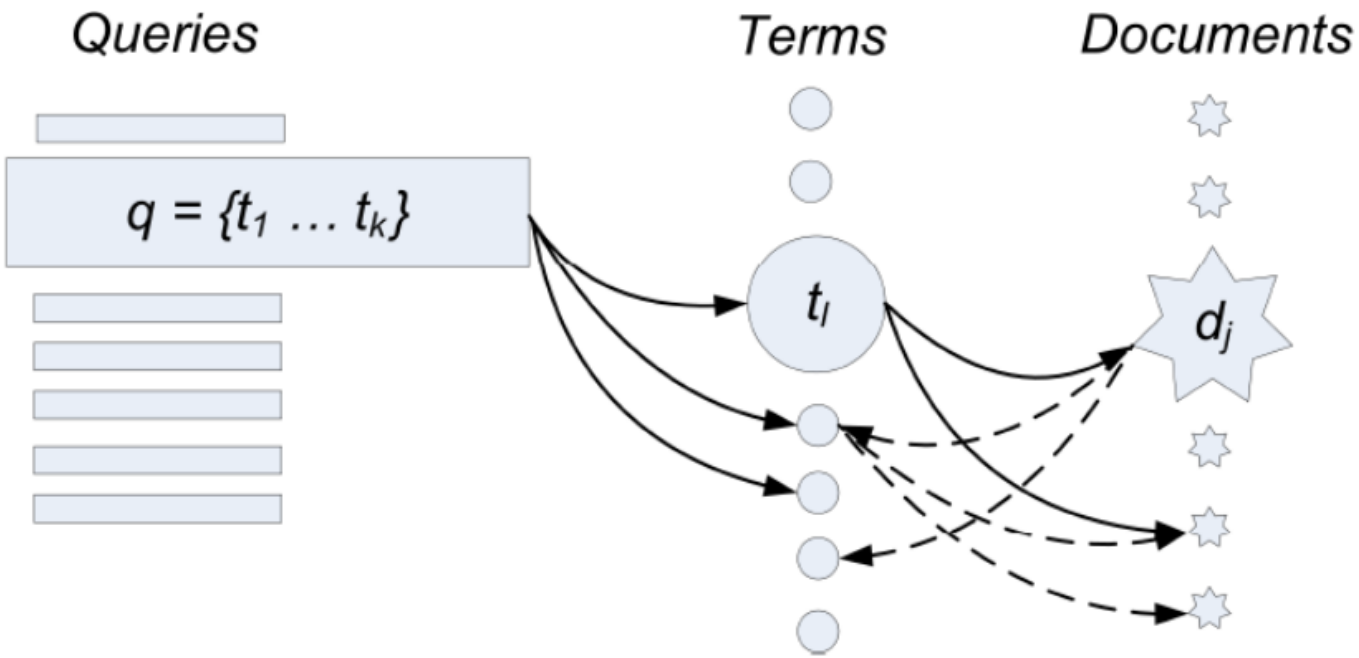
Query:[**гдз по математике 6 класс забурева мордкович учебник**]

гдз по математике	11158
гдз по математике 6 класс	11974
учебник зубарева мордкович 6 класс	169
учебник по математике зубарева	17
гдз зубарева мордкович 6 класс	2089
гдз 6 класс математика	6036

Сглаживание поведенческих данных

Идея:

Будем использовать двудольный граф: терм-документ



Heuristic Retrieval Model

$$w_{d_i,t_j} = QTF_{i,j} \cdot IQF_j =$$

$$= \frac{(\lambda+1)n(d_i,d_j)}{\lambda((1-\beta)+\beta\frac{n(d_i)}{\bar{n}(d_i)}+n(d_i,t_j))} \cdot log\,\frac{N_d - n(t_j)+0.5}{n(t_j)+0.5}$$

$$n(d_i,t_j) = \sum_{q \rightsquigarrow d_i, t_j \in q} f(q \rightsquigarrow d_i)$$

- $n(d_i)$ - число термов, связанных с d_i
- $\bar{n}(d_i)$ - среднее число термов, связанных с d_i
- $n(t_j)$ - число документов, связанных с t_j
- $\lambda = 0.5$, $\beta = 0.75$ - коэффициенты сглаживания
- N_d - число документов

Heuristic Retrieval Model

Для нового запроса $\hat{q} = \{\hat{t}_1, \ldots, \hat{t}_k\}$:

$$Rel_H(d_i, \hat{q}) = \sum_{\hat{t}_j \in \hat{q}} w_{d_i, \hat{t}_j} \cdot w_{\hat{t}_j}$$

$$w_{\hat{t}_j} = log \frac{N_q - n(\hat{t}_j) + 0.5}{n(\hat{t}_j) + 0.5}$$

- $n(t_j)$ - число документов, связанных с t_j
- N_q - общее число запросов

Probabilistic Retrieval Model

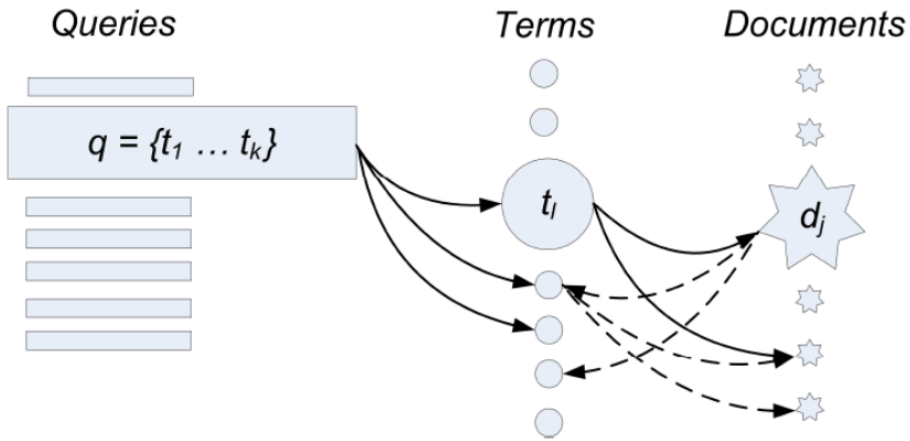
$$Rel_P(d_i, \hat{q}) = p(d_i|\hat{q}) = \sum_{\hat{t}_j \in \hat{q}} p(\hat{t}_j|\hat{q}) p(d_i|\hat{t}_j)$$

$$p(\hat{t}_j|\hat{q}) = \frac{exp(-p(\hat{t}_j))}{\sum_{\hat{t}_l \in \hat{q}} exp(-p(\hat{t}_l))} = \frac{exp(-\frac{n(\hat{t}_j)+\mu}{\sum_{\hat{t}_s \in D} n(\hat{t}_s)+\mu})}{\sum_{\hat{t}_l \in \hat{q}} exp(-\frac{n(\hat{t}_l)+\mu}{\sum_{\hat{t}_s \in D} n(\hat{t}_s)+\mu})}$$

$$p(d_i|\hat{t}_j) = \frac{n(d_i, \hat{t}_j)}{\sum_{d_l \in D} n(d_l, \hat{t}_j)}$$

$$n(d_i, \hat{t}_j) = \sum_{q \rightsquigarrow d_i, s.t. \hat{t}_j \in q} f(q \rightsquigarrow d_i)$$

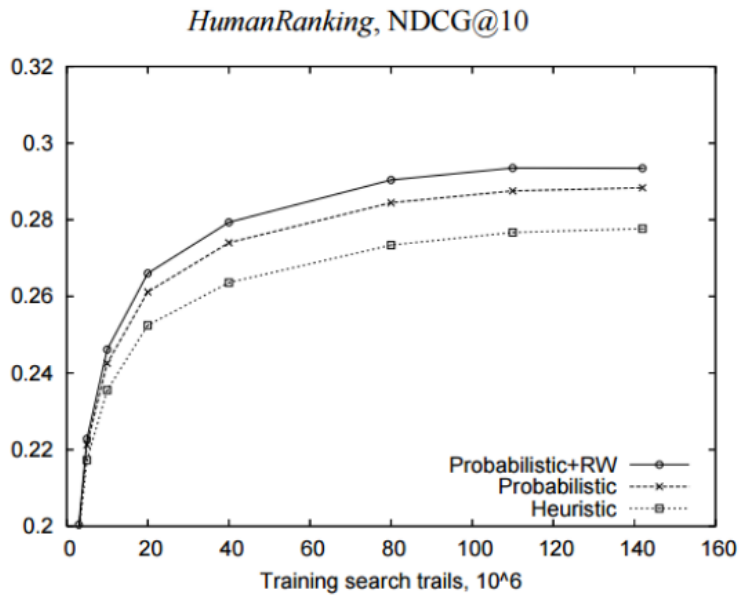
Random-Walk Extension

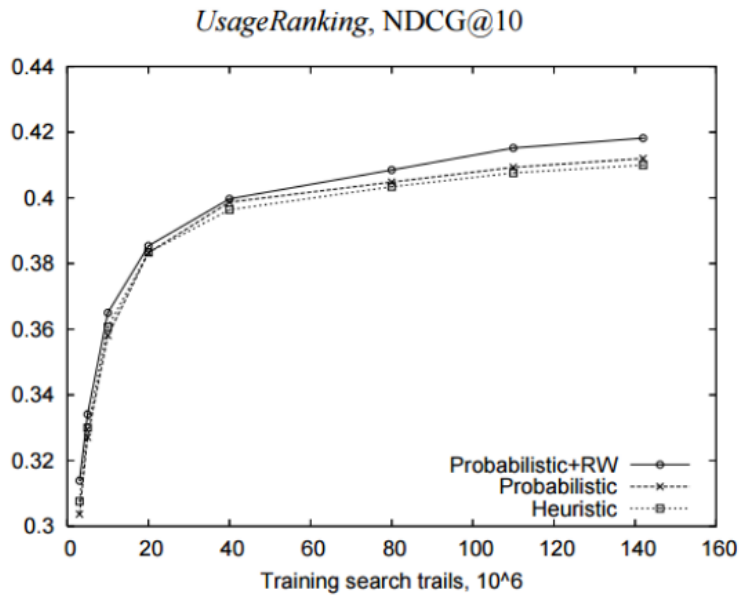


$$Rel_{P+RW}(d_i, \hat{q}) =$$

$$= \sum_{\hat{t}_j \in \hat{q}} p(\hat{t}_j|\hat{q}) \left(\alpha p(d_i, \hat{t}_j) + (1-\alpha) \sum_{\hat{t}_l \in \hat{q}, d_j} p(d_j, \hat{t}_j) p(\hat{t}_l, d_j) p(d_i, \hat{t}_l) \right)$$

Сравнение алгоритмов сглаживания





Collaborative ranking

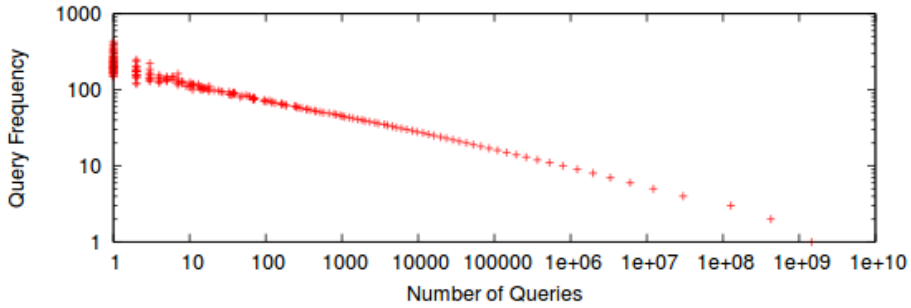


Figure 1: Query frequency with respect to the number of distinct queries

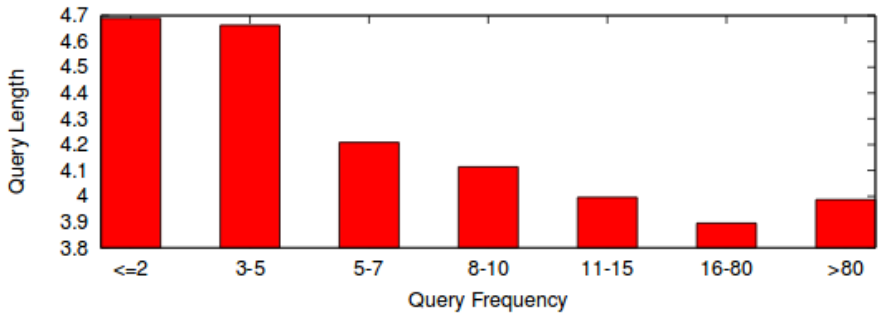


Figure 2: Query length with respect to different levels of query frequency

Collaborative ranking

$$f(q,d) = h(q,d) + \frac{1}{|S_q|}\sum_{q' \in S_q} s(q,q')h(q',d)$$

$$f(q,d) = h(x_{qd}) + \frac{1}{|S_q|}\sum_{q' \in S_q} s(x_{qq'})h(x_{q'd})$$

- $s(q,q')$ - функция похожести между запросами
- $h(q,d)$ - скоринг функция по одному запросу
- S_q - множесто “похожих” на запрос q запросов
- x_{qd} - множество факторов, описывающих пару (q,d)
- $x_{qq'}$ - множество факторов, описывающих пару (q,q')

Collaborative ranking

Pointwise

$$\min_{h,s} \mathcal{R}_{reg}(f) = \sum_{q,d} (y_{qd} - f(q,d))^2$$

$$\mathcal{R}_{reg}(h,s) = \sum_{q,d} \left(y_{qd} - h(x_{qd}) - \frac{1}{|S_q|} \sum_{q' \in S_q} s(x_{qq'}) h(x_{q'd}) \right)^2$$

Pairwise

$$L(f(x_{qd}), f(x_{qd'})) = \max(0, f(x_{qd'}) - f(x_{qd}) + \tau)^2$$

Alternating Minimization

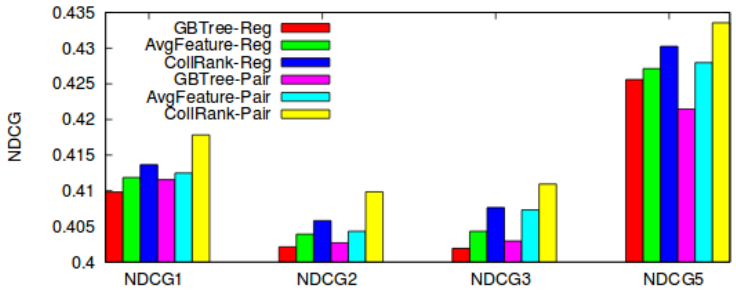
- Имея $s(x_{qq'})$, можем оптимизировать $\mathcal{R}_{reg}(h, s)$ по h

$$\min_{h \in \mathcal{H}_h} \sum_{q,d} \left(y_{qd} - h(x_{qd}) - \frac{1}{|S_q|} \sum_{q' \in S_q} s(x_{qq'}) h(x_{q'd}) \right)^2$$

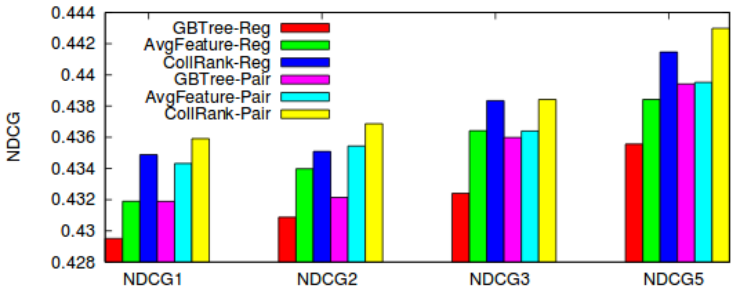
- Имея $h(x_{qd})$, можем оптимизировать $\mathcal{R}_{reg}(h, s)$ по s

$$\min_{h \in \mathcal{H}_s} \sum_{q,d} \left(y_{qd} - h(x_{qd}) - \frac{1}{|S_q|} \sum_{q' \in S_q} s(x_{qq'}) h(x_{q'd}) \right)^2$$

Сравнение алгоритмов

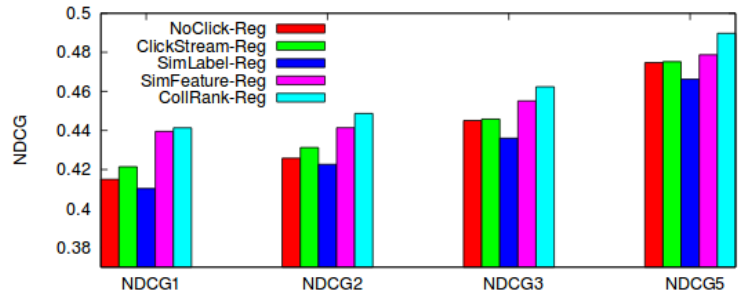


(a) Performance on Web10K

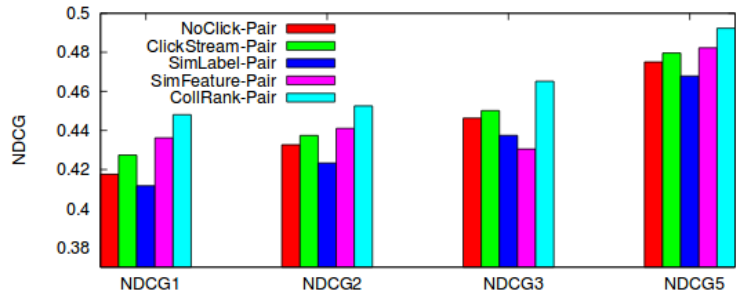


(b) Performance on Web30K

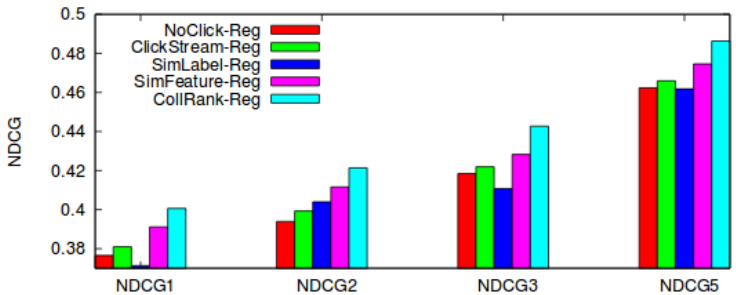
Figure 3: The performance comparisons of CollRank and baselines on MSLR data set



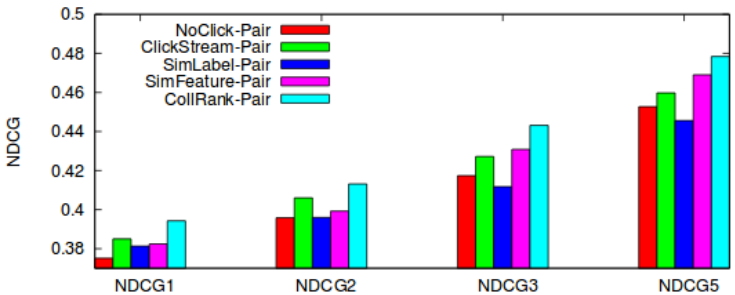
(a) Performance on Test1 with square loss



(b) Performance on Test1 with pair-wise loss



(c) Performance on Test2 with square loss



(d) Performance on Test2 with pair-wise loss

Figure 4: The performance comparisons of CollRank and four baseline methods

Вопросы