Towards Goal-driven Information Retrieval:

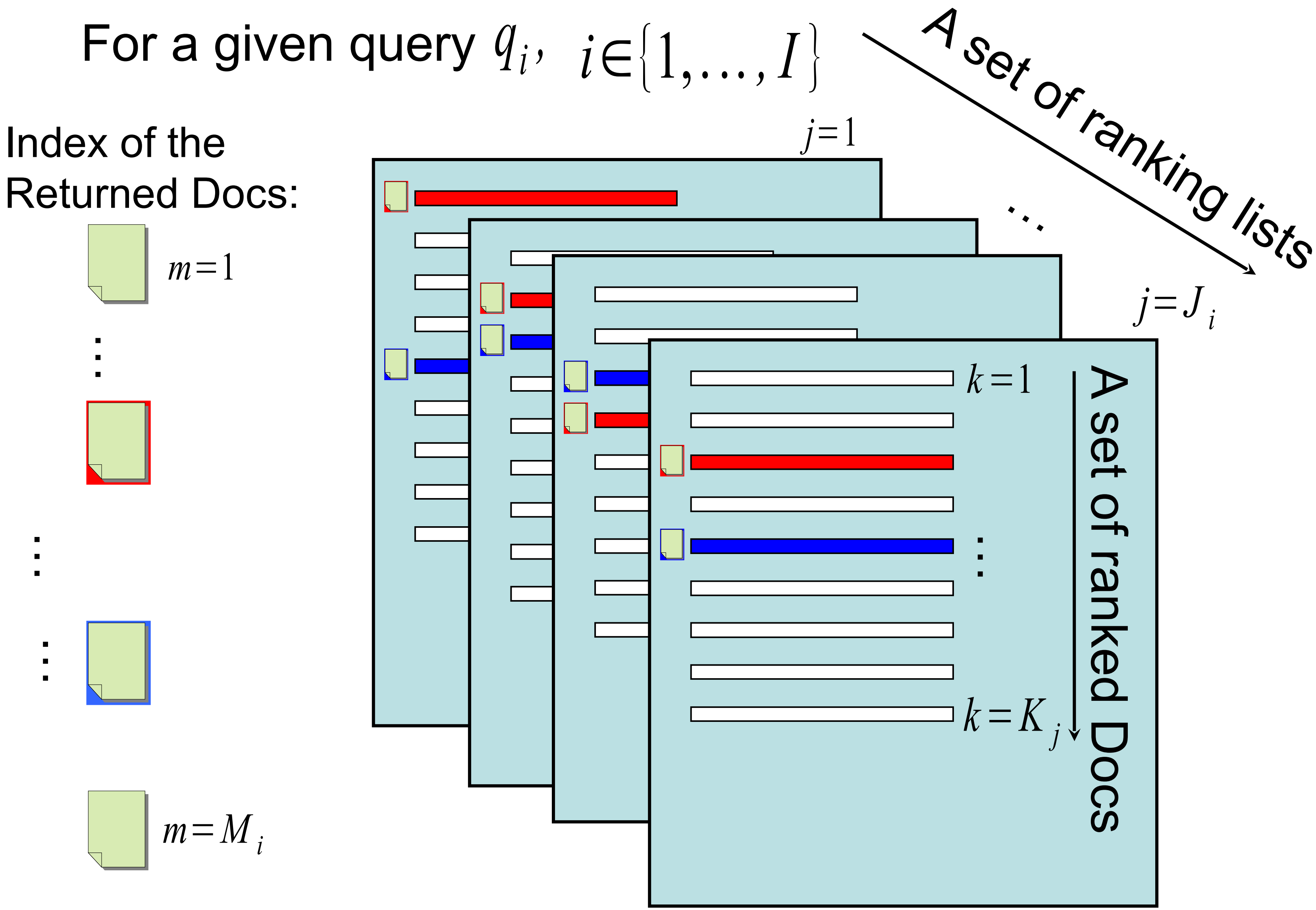
Graphical models for Click-through data

Jun Wang, **Tamas Jambor**, University College London

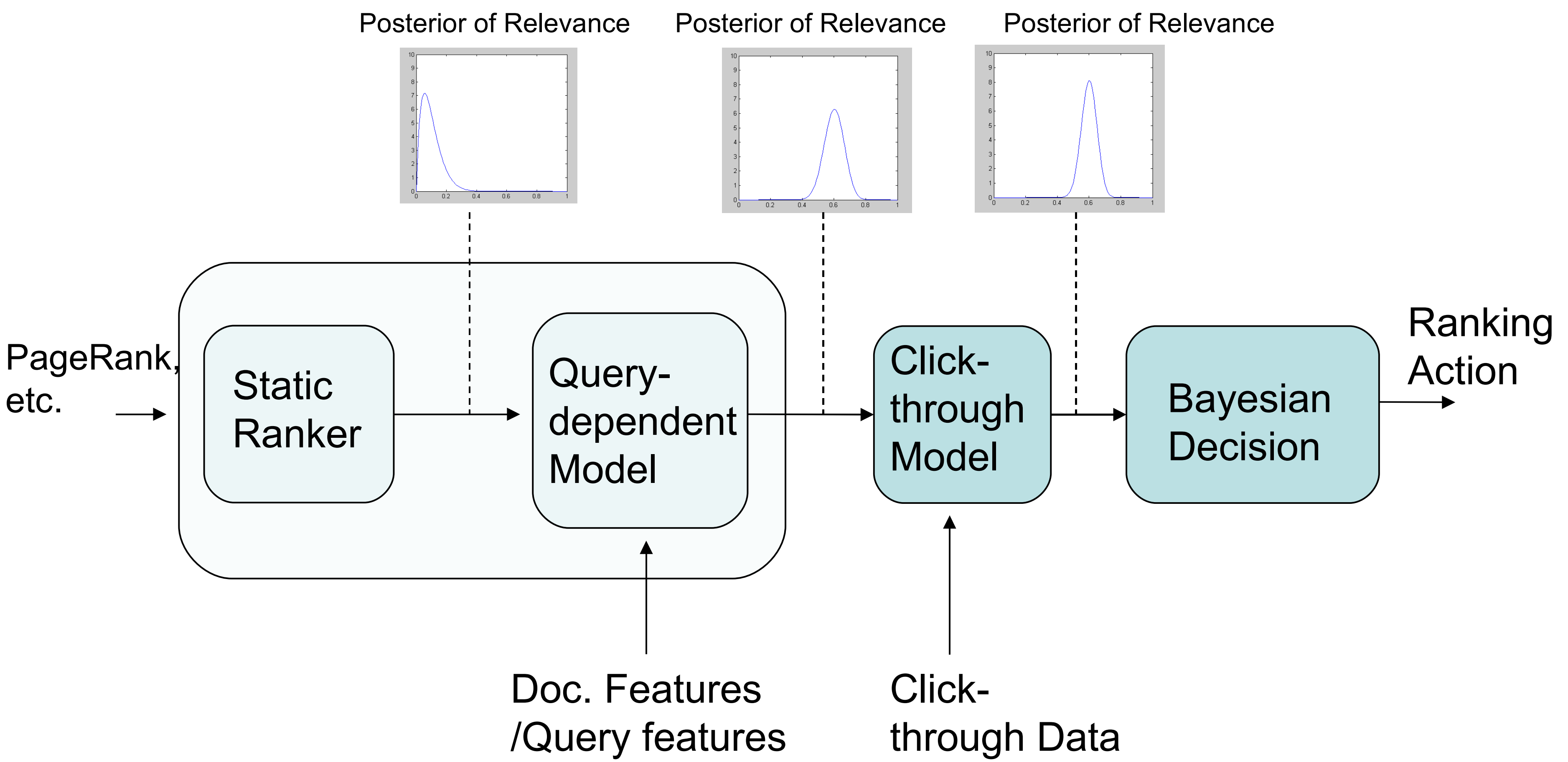
With Mike Taylor, Onno Zoeter,
Tom Minka and Steve Robertson in Microsoft, Cambridge

Click-through Data

For a given query q_i, $i \in \{1,\ldots,I\}$

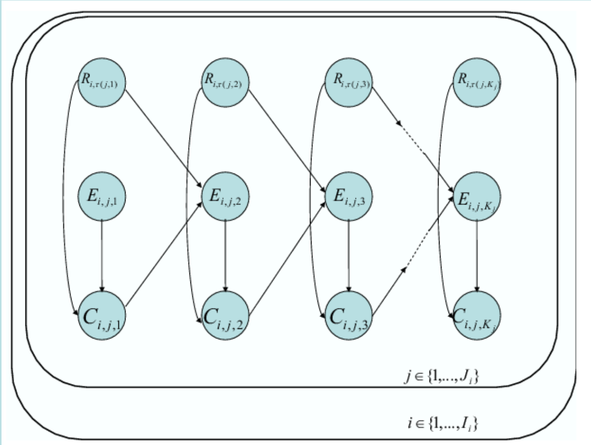


A Vision of a Ranker

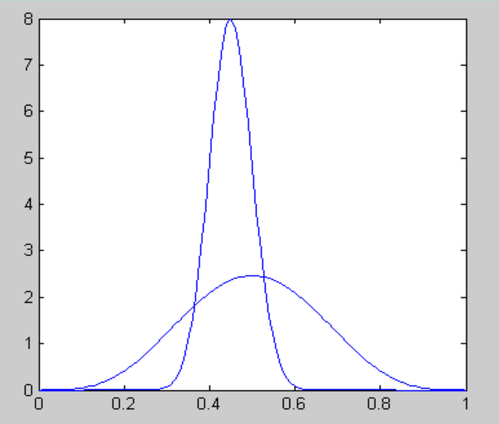


Summary of the Work

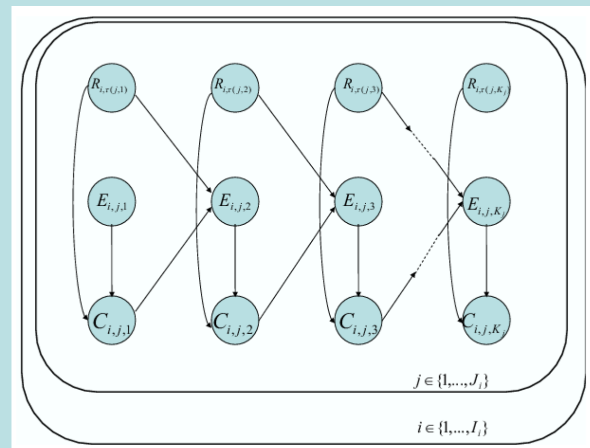
- Graphical Models for Click-through Data



- Bayesian Decision: Ranking under uncertainty



Click-through Models



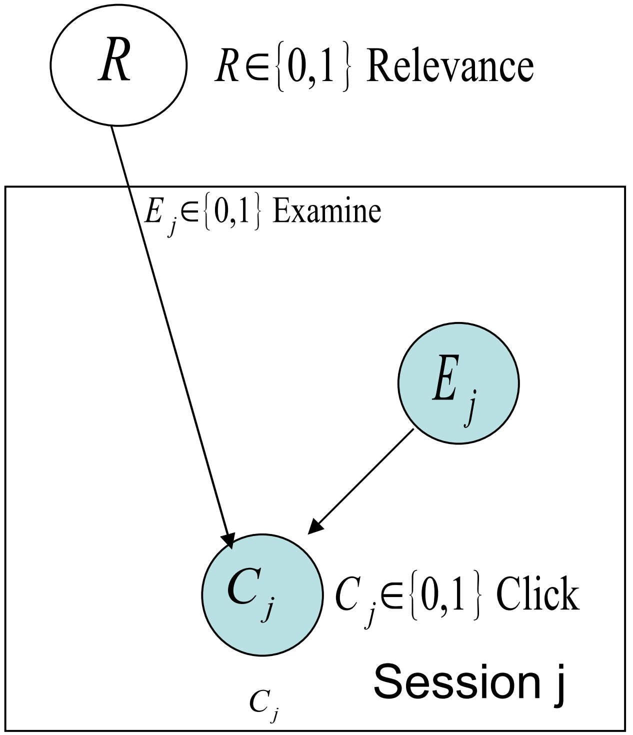
Summary

- Click-through Models
  - Model A: Click-over-examine model
  - Model B: ClickChain model:
    - models doc sequence dependency of C,E events
  - Model C: Mixture model over query types
    - Definitive (single answer) vs exploratory (multiple)

Model A: A simple Click-over-Examine model

Aim: predict relevance

For a document query pair:



1. We assume there is a binary relevance R for each of the web pages

2. For each of the displayed web pages, we observe two random variables:

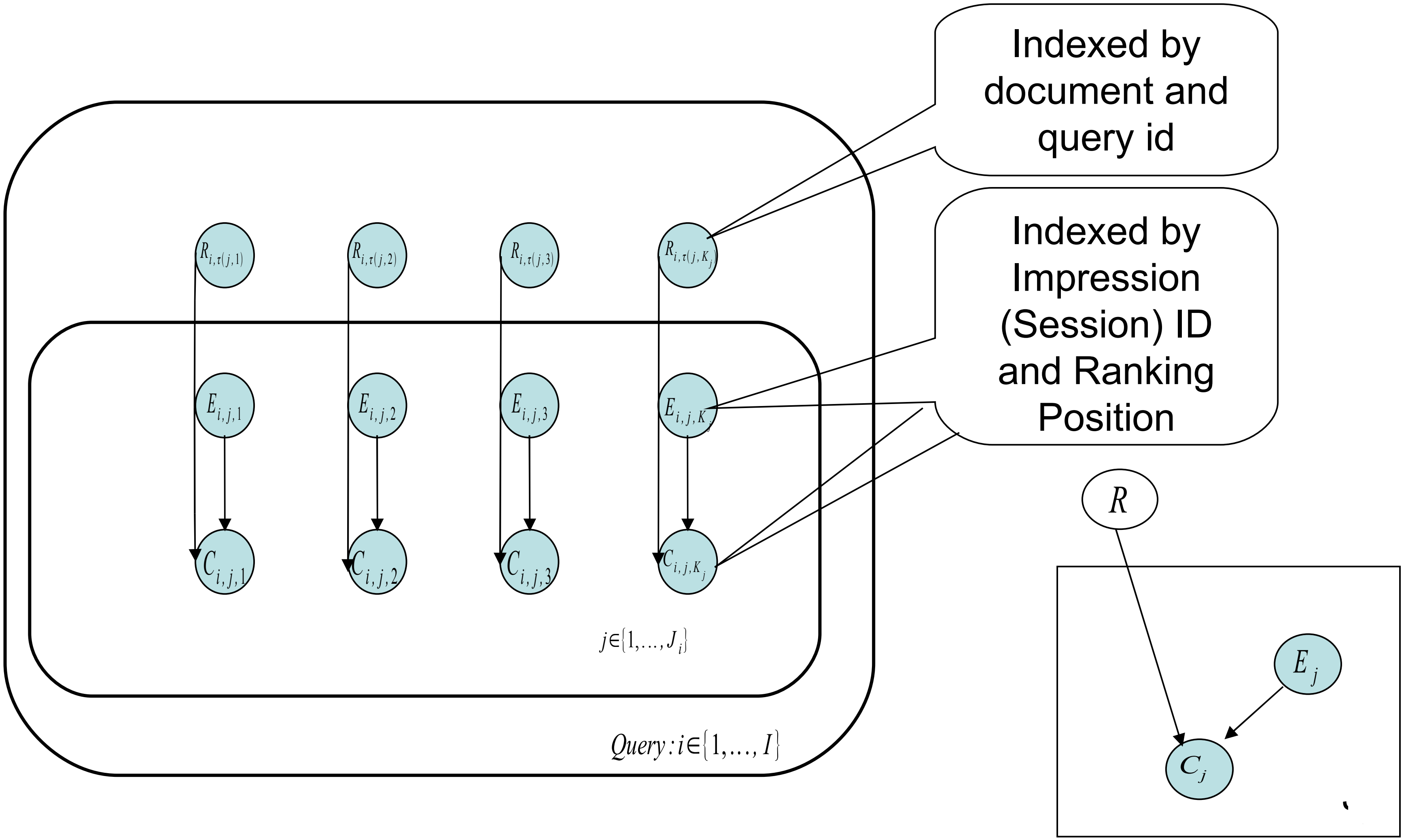
C_j: click or not click

E_j: examine or not examine

(it is reasonable to assume that the user has examined any web pages above the last click

Model A: A simple Click-over-Examine model (cont')

- Assumption: documents are judged (clicked) independently with respect to the documents above/below



Model A: A simple Click-over-Examine model (before Last Click)

- One more assumption:
  - Documents above last click are examined
  - We only consider clicks/non-clicks above the last click.
- Parameter estimation:

$$\Theta = \{p_{c,r}, p_{c,\bar{r}}\}$$

$$P(C=1|R,E) = \begin{cases} 0 & E=0 \\ p_{c,r} & R=1, \mathrm{E}=1 \\ p_{c,\bar{r}} & R=0, \mathrm{E}=1 \end{cases}$$

$$p_{c,r} = \frac{n_{c,r}}{n_{c,r} + n_{\bar{c},r}}$$

$$p_{c,\bar{r}} = \frac{n_{c,\bar{r}}}{n_{c,\bar{r}} + n_{\bar{c},\bar{r}}}$$

A simple Click-over-Examine model (4)

- Relevance Prediction:
  - Joint Probability:

$$P\left(R_{i,a}=1,\left\{C_{i,j,k},E_{i,j,k}\right\}_{j=1,k=1}^{J_i,K_j};\Theta\right)$$

$$\text{¿}P\left(R_{i,a}=1\right)\left(\prod_{\forall j,k:\tau(j,k)=a} P\left(C_{i,j,k}\middle|E_{i,j,k},R_{i,a}=1\right)P\left(E_{i,j,k}\right)\right)$$

Examine is observed. It can be safely left out

  - Posterior Probability of Relevance:

$$P\left(R_{i,a}=1\middle|\left\{C_{i,j,k},E_{i,j,k}\right\}_{j=1,k=1}^{J_i,K_j};\Theta\right)$$

$$\text{¿}P\left(R_{i,a}=1\right)\left(\prod_{\forall j,k:\tau(j,k)=a} P\left(C_{i,j,k}\middle|E_{i,j,k}=1,R_{i,a}=1\right)\right)$$

$$\text{¿}P\left(R_{i,a}=1\right)\left(p_{c,r}\right)^{n_{c,r}}\left(1-p_{c,r}\right)^{n_{\bar{c},r}}$$

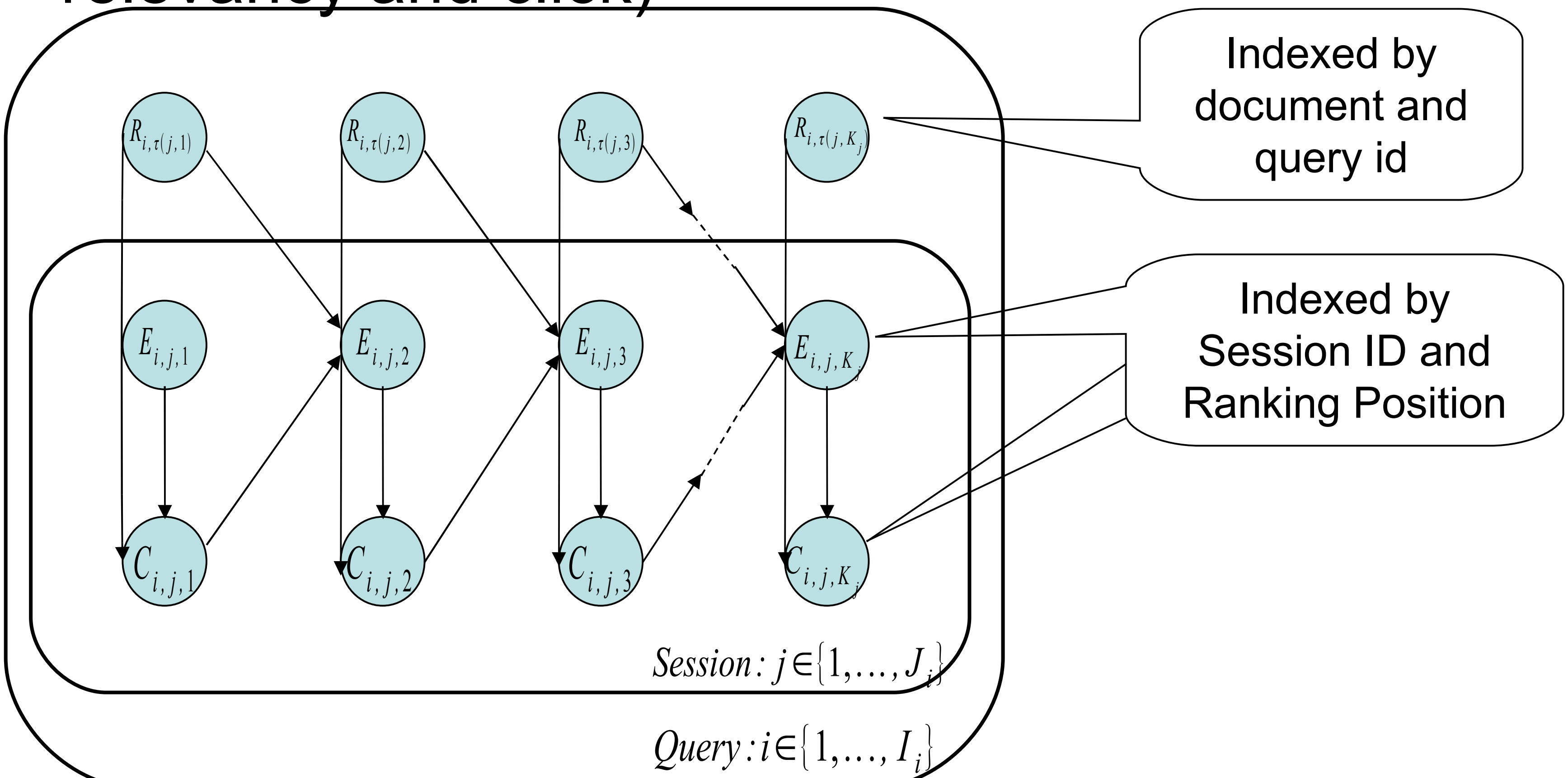
A simple Click-over-Examine model (5)

- For the documents after last clicks, we do not have the observation of the examines
- We may treat them as missing data and marginalize them out
- Thus, the Posterior Probability of Relevance:

$$
\begin{aligned}
&P(R_{i,a}=1 \mid Observation;\Theta) \\
&\text{¿}\, P(R_{i,a}=1)\Big(\prod_{k:\tau(j,k)=a \cap E_{i,j,k}=1} P(C_{i,j,k} \mid E_{i,j,k}=1, R_{i,a}=1)\, P(E_{i,j,k}=1)\Big) \\
&\Big(\prod_{(\tau(j,k)=a)\cap(E_{i,j,k}\ \text{is unknown})} \Big(\sum_{E_{i,j,k}} P(C_{i,j,k}=0 \mid E_{i,j,k}, R_{i,a}=1)\, P(E_k)\Big)\Big)
\end{aligned}
$$

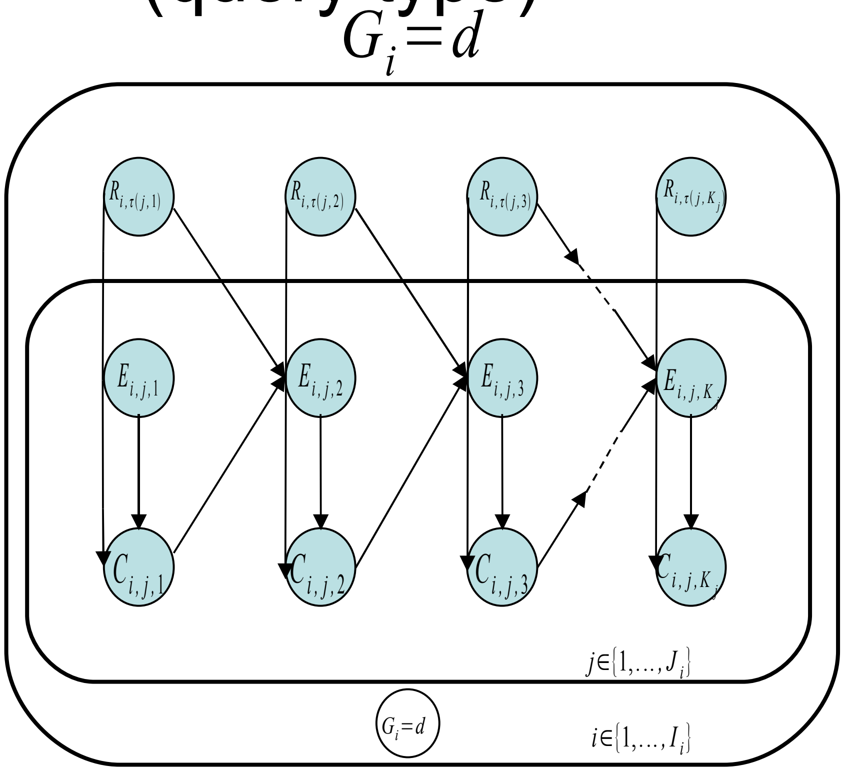
Model B: ClickChain Model

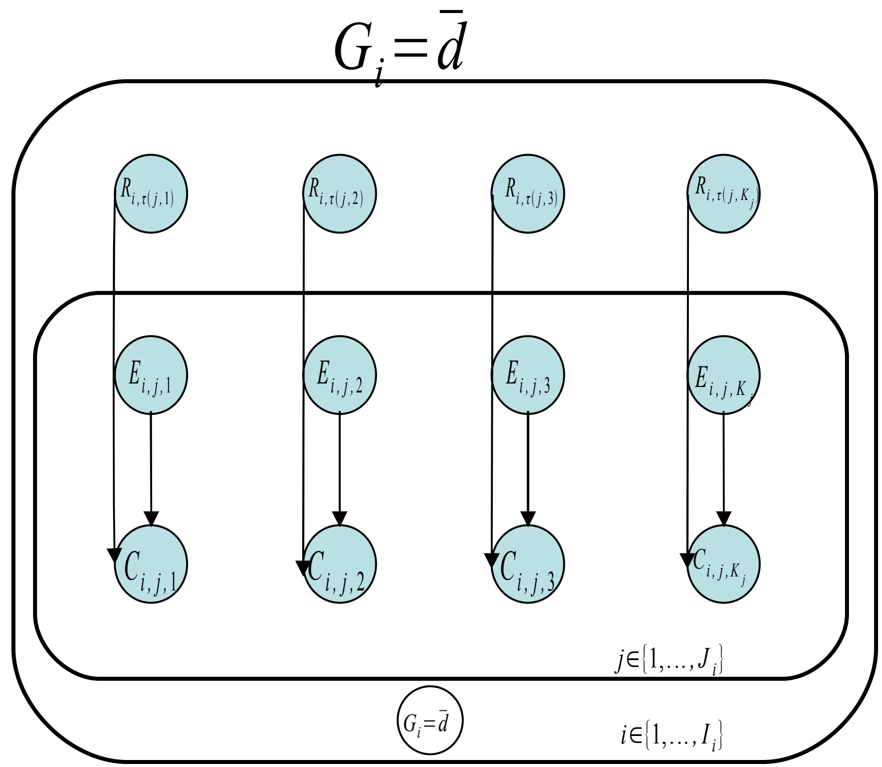
- A more realistic model: ClickChain
- whether a user examines current document or not is dependent on the previous document (its relevancy and click)



Model C: Mixture Model

- However, this dependency relies on user goal (query type)





Definitive query (one answer from the query):

We expect higher dependency

Exploratory query (multiple relevance web pages):

We expect less dependency

Model C: Mixture Model(Cont')

- The query type:

$$P(E_k|C_{k-1},R_{k-1})=\sum_G P(E_k,G|C_{k-1},R_{k-1})$$
$$\text{¿}\,P(E_k,|G=d,C_{k-1},R_{k-1})P(G=d)+P(E_k|G=\bar{d},C_{k-1},R_{k-1})P(G=\bar{d})$$
$$\text{¿}\,P(E_k,|G=d,C_{k-1},R_{k-1})P(G=d)+P(E_k|G=\bar{d})P(G=\bar{d})$$

- It can be learned by the EM algorithm per query

Model C: Mixture Model (Cont')

- Preliminary Results:

P(G=d)	0.91	0.84	0.83	0.81	0.80	0.80	0.78
Query	cvs	levis	**usatoday**	mbna	sina	simslot	scottrade

P(G=d)	0.75	0.74	0.65	0.62	0.56	0.53	0.51
Query	shockwa ve	cia	cnn	aerosmith	army	alaska	**women**

Model C: Mixture Model (Cont')

- Preliminary Results:

alaska P(G=1):0.53

URL	Label	num. Clicks
http://wildlife.alaska.gov/	Fair	2
http://511.alaska.gov/	Good	4
http://www.uaa.alaska.edu/	Bad	1
http://climate.gi.alaska.edu/	Good	5
http://www.alaskaair.com/	Bad	8
http://www.travelalaska.com/	Fair	40
http://www.alaska.com/	Perfect	29
http://www.avo.alaska.edu/	Good	3
http://www.gi.alaska.edu/	Bad	3
http://www.state.ak.us/	Perfect	97
http://www.alaska.com/about/photos	Good	3
http://www.dot.state.ak.us/	Excellent	1

cnn P(G=1):0.65

URL	Label	num. Clicks
http://sportsillustrated.cnn.com/	Fair	23
http://www.cnn.com/	Perfect	12644
http://transcripts.cnn.com/TRANSCRIPTS/	Fair	16
http://robots.cnn.com/	Perfect	15
http://money.cnn.com/	Good	70
http://europe.cnn.com/	Excellent	26
http://edition.cnn.com/	Excellent	248
http://www.cnn.com/WEATHER	Good	113
http://en.wikipedia.org/wiki/CNN	Good	1
http://search.cnn.com/	Fair	24
http://money.cnn.com/news/	Good	1
http://edition.cnn.com/WEATHER	Fair	7
http://www.cnn.com/WORLD/	Excellent	48
http://www.cnn.com/TECH/	Excellent	3
http://edition.cnn.com/ASIA/	Good	8
http://asia.cnn.com/	Excellent	5
http://cnnstudentnews.cnn.com/fy	Good	4

levis P(G=1):0.84

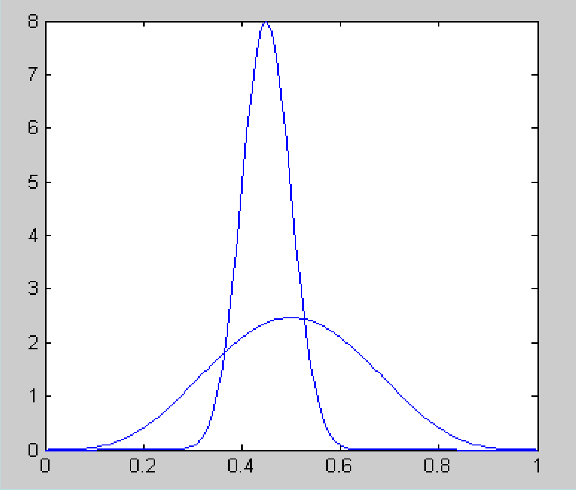
URL	Label	num. Clicks
http://www.levis.com/	Perfect	66
http://www.levis.info/	Bad	1
http://www.levistrauss.com/	Excellent	2

cvs P(G=1):0.91

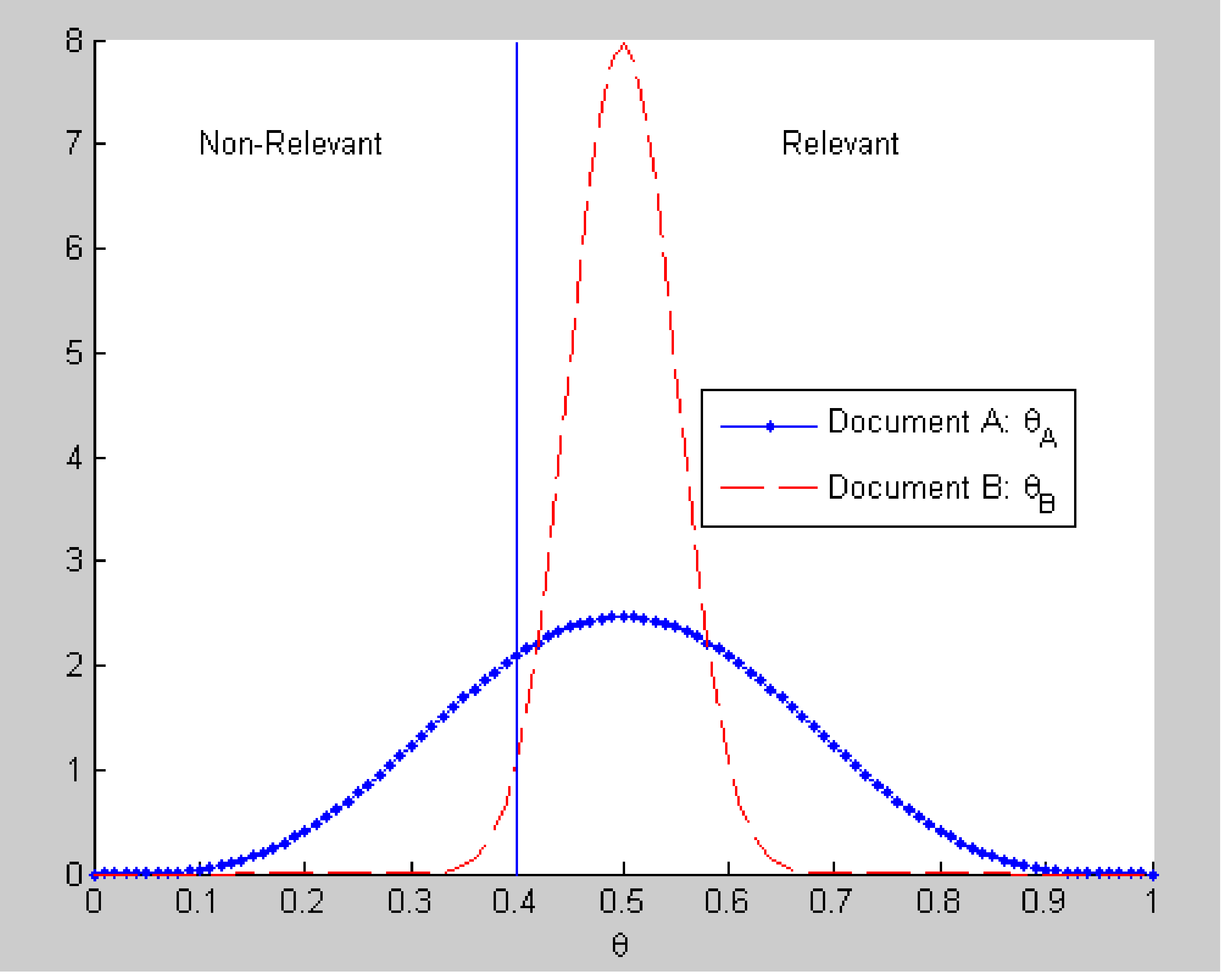
URL	Label	num. Clicks
http://www.cvs.com/	Good	676
http://www.nongnu.org/cvs/	Bad	2

Ranking under Uncertainty

- Next step is to rank web pages on the basis of the estimated probability of relevance



Ranking under Uncertainty

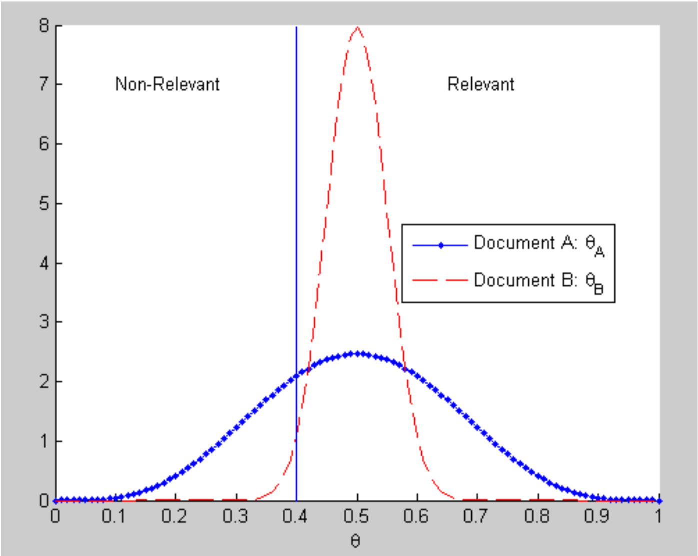


Ranking under Uncertainty (2)

- What is the optimal ranking score given the posterior distribution of the relevance?
- Empirically we found that optimal ranking score:

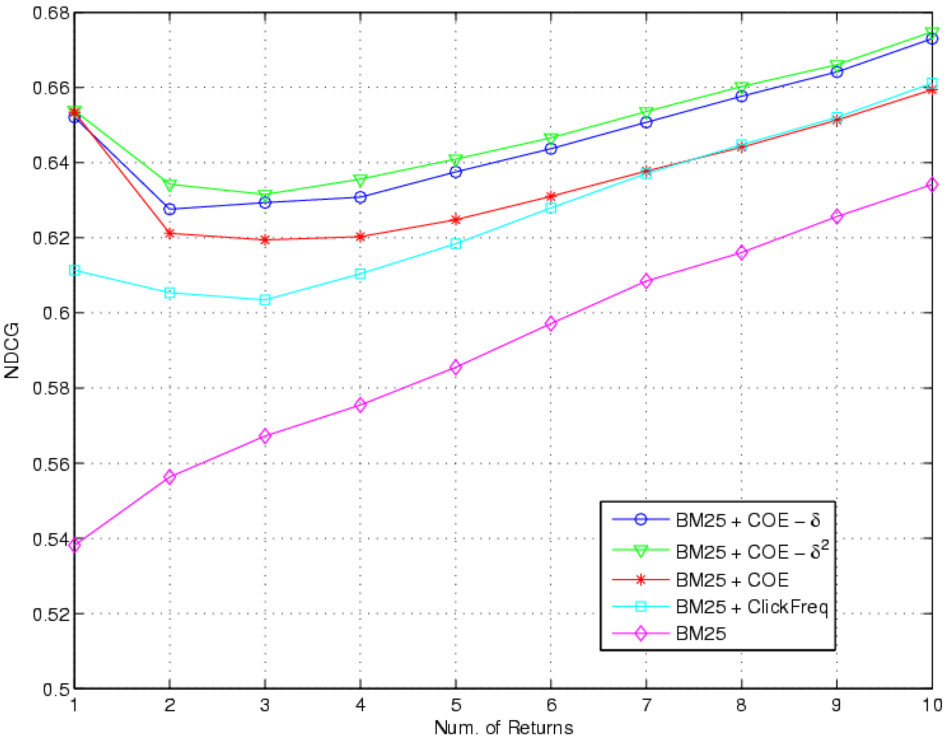
$$s_i = mean(\theta_i) - a \cdot \mathrm{var}(\theta_i),$$

where a is a parameter

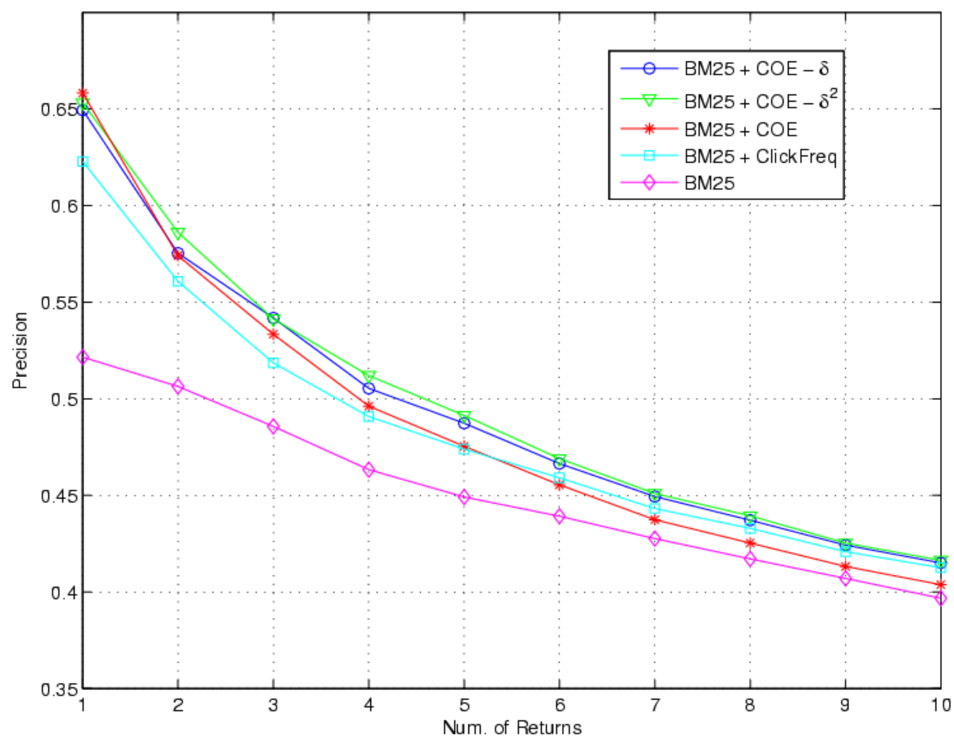


- The theoretical foundation can be found in our SIGIR09 paper

Experiments



(a) NDCG



(b) Precision

Future Work

- Click fraud
- Use of richer feature set:
  - Dwell time
- Network analysis
  - Modelling click-through data as a complex network

The quest for assets

- The Good:
  - Knowing what web pages have been clicked given queries.
- The Bad:
  - The data is rather static. We need to have a feedback loop. Observing users' response towards the updated model prediction.
- Wanted:
  - Cross-referencing between algorithmic search and sponsored search.

Thanks