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
For a given query $q_i$, $i \in \{1, \ldots, I\}$, a set of ranking lists is returned. Each ranking list consists of documents indexed from $m=1$ to $m=M_i$. Click-through data is used to evaluate the ranked documents, with $k=1$ to $k=K_j$.
A Vision of a Ranker

- Static Ranker
  - PageRank, etc.

- Query-dependent Model
  - Doc. Features /Query features

- Click-through Model
  - Click-through Data

- Posterior of Relevance

- Bayesian Decision Action

- Ranking Action
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

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

\[
\begin{align*}
R_{i,j,1} & \quad R_{i,j,2} \\
R_{i,j,1} & \quad R_{i,j,2} \\
R_{i,j,1} & \quad R_{i,j,2} \\
\end{align*}
\]

Indexed by document and query id

Indexed by Impression (Session) ID and Ranking Position

\[
\begin{align*}
E_{i,j,1} & \quad E_{i,j,2} \\
E_{i,j,1} & \quad E_{i,j,2} \\
E_{i,j,1} & \quad E_{i,j,2} \\
\end{align*}
\]

\[
\begin{align*}
C_{i,j,1} & \quad C_{i,j,2} \\
C_{i,j,1} & \quad C_{i,j,2} \\
C_{i,j,1} & \quad C_{i,j,2} \\
\end{align*}
\]

Query: \( i \in \{1, \ldots, I\} \)

Impression: \( j \in \{1, \ldots, J\} \)
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:

\[
\Theta = \{p_{c,r}, p_{c,\bar{r}}\}
\]

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

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

\[
p_{c,\bar{r}} = \frac{n_{c,\bar{r}}}{n_{c,\bar{r}} + n_{\bar{c},\bar{r}}}
\]
A simple Click-over-Examine model (4)

- **Relevance Prediction:**
  - **Joint Probability:**
    \[
    P(R_{i,a} = 1, \{C_{i,j,k}, E_{i,j,k}\}_{j=1,k=1}^{J_i,K_j}; \Theta) \]
    \[
    \propto P(R_{i,a} = 1)(\prod_{\forall j,k: \tau(j,k)=a} P(C_{i,j,k} | E_{i,j,k}, R_{i,a} = 1)P(E_{i,j,k}))
    \]
  - **Posterior Probability of Relevance:**
    \[
    P(R_{i,a} = 1|\{C_{i,j,k}, E_{i,j,k}\}_{j=1,k=1}^{J_i,K_j}; \Theta) \]
    \[
    \propto P(R_{i,a} = 1)(\prod_{\forall j,k: \tau(j,k)=a} P(C_{i,j,k} | E_{i,j,k}, R_{i,a} = 1))
    \]
    \[
    \propto P(R_{i,a} = 1)(p_{c,r})^{n_{c,r}}(1-p_{c,r})^{n_{\bar{c},r}}
    \]

Examine is observed. It can be safely left out.
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:

\[
P(R_{i,a}=1|\text{Observation}; \Theta) \\
\text{i} P(R_{i,a}=1) \left( \prod_{k: \tau(j,k)=a \cap E_{i,j,k}=1} P(C_{i,j,k}|E_{i,j,k}=1, R_{i,a}=1) P(E_{i,j,k}=1) \right) \\
\left( \prod_{(\tau(j,k)=a) \cap (E_{i,j,k} \text{ is unknown})} \left( \sum_{E_{i,j,k}} P(C_{i,j,k}=0|E_{i,j,k}, R_{i,a}=1) P(E_{k}) \right) \right)
\]
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)

\[ G_i = d \]

**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:

\[
P(E_k|C_{k-1}, R_{k-1}) = \sum_G P(E_k, G|C_{k-1}, R_{k-1})
\]

\[
\leq P(E_k, G=d, C_{k-1}, R_{k-1}) P(G=d) + P(E_k, G=\bar{d}, C_{k-1}, R_{k-1}) P(G=\bar{d})
\]

\[
\leq P(E_k, G=d, C_{k-1}, R_{k-1}) P(G=d) + P(E_k, G=\bar{d}) P(G=\bar{d})
\]

• It can be learned by the EM algorithm per query
Model C: Mixture Model (Cont’)

• Preliminary Results:

<table>
<thead>
<tr>
<th>Query</th>
<th>cvs</th>
<th>levis</th>
<th>usatoday</th>
<th>mbna</th>
<th>sina</th>
<th>simslot</th>
<th>scottrade</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(G=d)</td>
<td>0.91</td>
<td>0.84</td>
<td>0.83</td>
<td>0.81</td>
<td>0.80</td>
<td>0.80</td>
<td>0.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query</th>
<th>shockwave</th>
<th>cia</th>
<th>cnn</th>
<th>aerosmith</th>
<th>army</th>
<th>alaska</th>
<th>women</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(G=d)</td>
<td>0.75</td>
<td>0.74</td>
<td>0.65</td>
<td>0.62</td>
<td>0.56</td>
<td>0.53</td>
<td>0.51</td>
</tr>
</tbody>
</table>
Model C: Mixture Model (Cont’)

• Preliminary Results:

<table>
<thead>
<tr>
<th>Model</th>
<th>P(G=1):</th>
<th>URL</th>
<th>Label</th>
<th>num. Clicks</th>
</tr>
</thead>
<tbody>
<tr>
<td>alaska</td>
<td>0.53</td>
<td><a href="http://wildlife.alaska.gov/">http://wildlife.alaska.gov/</a></td>
<td>Fair</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.uaa.alaska.edu/">http://www.uaa.alaska.edu/</a></td>
<td>Good</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://climate.gi.alaska.edu/">http://climate.gi.alaska.edu/</a></td>
<td>Bad</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.alaskaair.com/">http://www.alaskaair.com/</a></td>
<td>Good</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.travelalaska.com/">http://www.travelalaska.com/</a></td>
<td>Fair</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.alaska.com/">http://www.alaska.com/</a></td>
<td>Perfect</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.avoo.alaska.edu/">http://www.avoo.alaska.edu/</a></td>
<td>Good</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.gi.alaska.edu/">http://www.gi.alaska.edu/</a></td>
<td>Bad</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.state.ak.us/">http://www.state.ak.us/</a></td>
<td>Perfect</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.alaska.com/about/photos">http://www.alaska.com/about/photos</a></td>
<td>Good</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.dot.state.ak.us/">http://www.dot.state.ak.us/</a></td>
<td>Excellent</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>P(G=1):</th>
<th>URL</th>
<th>Label</th>
<th>num. Clicks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><a href="http://sportsillustrated.cnn.com/">http://sportsillustrated.cnn.com/</a></td>
<td>Fair</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.cnn.com/">http://www.cnn.com/</a></td>
<td>Perfect</td>
<td>12644</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://transcripts.cnn.com/TRANSCRIPTS/">http://transcripts.cnn.com/TRANSCRIPTS/</a></td>
<td>Fair</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://robots.cnn.com/">http://robots.cnn.com/</a></td>
<td>Perfect</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://money.cnn.com/">http://money.cnn.com/</a></td>
<td>Good</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://europe.cnn.com/">http://europe.cnn.com/</a></td>
<td>Excellent</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.cnn.com/WEATHER">http://www.cnn.com/WEATHER</a></td>
<td>Good</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://en.wikipedia.org/wiki/CNN">http://en.wikipedia.org/wiki/CNN</a></td>
<td>Good</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://search.cnn.com/">http://search.cnn.com/</a></td>
<td>Fair</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://money.cnn.com/news/">http://money.cnn.com/news/</a></td>
<td>Good</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.cnn.com/TECH/">http://www.cnn.com/TECH/</a></td>
<td>Excellent</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.cnn.com/ASIA/">http://www.cnn.com/ASIA/</a></td>
<td>Excellent</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://asia.cnn.com/">http://asia.cnn.com/</a></td>
<td>Good</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://cnnstudentnews.cnn.com/fy">http://cnnstudentnews.cnn.com/fy</a></td>
<td>Excellent</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.cnn.com/WORLD/">http://www.cnn.com/WORLD/</a></td>
<td>Good</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>P(G=1):</th>
<th>URL</th>
<th>Label</th>
<th>num. Clicks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><a href="http://www.cvs.com/">http://www.cvs.com/</a></td>
<td>Good</td>
<td>676</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="http://www.nongnu.org/cvs/">http://www.nongnu.org/cvs/</a></td>
<td>Bad</td>
<td>2</td>
</tr>
</tbody>
</table>
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:

\[ s_i = \text{mean}(\theta_i) - a \cdot \text{var}(\theta_i), \]

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