PITT at TREC 2012 Session Track: Adaptive Browsing Novelty in a Search Session

Jiepu Jiang, Daqing He, Shuguang Han

University of Pittsburgh
Our focus this year
- Finding novel results for the current query

Our goal
- We expect the methods to be relatively conservative
  - with significant effects on ranking novel results to higher positions
  - without much hurt on the ad hoc search performance, e.g. nDCG@10 over all qrels (without removing duplicates)
- because currently it is unclear:
  - whether we should consider novelty issues in a search session
  - if yes, what are the proper methods?
- Anyway, it seems risky if novelty search system hurts nDCG@10
Novelty in a Search Session

- Two types of novelty issues in a search session
  - Novelty in content *(not our focus this year)*
    - Documents with very similar information are retrieved
    - Partly discussed in web track diversity task
  - Duplicate results *(new to the session track; our focus)*
    - The same webpage is returned in the results of many queries
    - Should we discount the duplicate results in current query?
Novelty in a Search Session

Should we discount duplicate results in current search?

Pros:
- It may better explain users’ query reformulation behaviors.

Cons:
- User may overlook a relevant result when browsing the result list
- User may be confused about how the system works
  - Loss of control on the search process
- Lack of proper evaluation methods and guidelines for ranking
  - remove previously shown retrieved results in the evaluation of current query? (seems too radical)
  - remove the clicked documents in previous results? (seems too conservative)
Pros on Discounting Duplicate Results in Evaluation & Ranking

- Do users reformulate to get good ad hoc search performance? (Probably No)
  - We extract 204 query reformulation pairs (reformulating from $q_{n-1}$ to $q_n$) from TREC 2011 session track sessions
  - Comparison of $P@k$ and $nDCG@k$ for the two consecutive queries

<table>
<thead>
<tr>
<th>Metric</th>
<th>mean</th>
<th>SD</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P@10$</td>
<td>0.026</td>
<td>0.275</td>
<td>0.171</td>
</tr>
<tr>
<td>$P@20$</td>
<td>0.022</td>
<td>0.212</td>
<td>0.143</td>
</tr>
<tr>
<td>$nDCG@10$</td>
<td>0.021</td>
<td>0.241</td>
<td>0.209</td>
</tr>
<tr>
<td>$nDCG@20$</td>
<td>0.019</td>
<td>0.204</td>
<td>0.180</td>
</tr>
</tbody>
</table>

- If we believe the ad hoc search evaluation metrics (e.g. $P@k$ and $nDCG@k$) are valid measures of search performance in a search session, our results indicate users are reformulating queries that are nothing better than previous ones 😞
## Pros on Discounting Duplicate Results in Evaluation & Ranking

- Do users reformulate to get novel search results? *(Probably Yes)*
  - Comparison of jaccard similarity and ranking correlation on two consecutive queries’ top results.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Jaccard Similarity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(average over topics)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10 results</td>
<td>0.357</td>
<td>0.377</td>
</tr>
<tr>
<td>Top 20 results</td>
<td>0.354</td>
<td>0.360</td>
</tr>
<tr>
<td><strong>Spearman’s ρ</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(average over topics)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top 10 results</td>
<td>0.103</td>
<td>0.609</td>
</tr>
<tr>
<td>Top 20 results</td>
<td>0.145</td>
<td>0.577</td>
</tr>
</tbody>
</table>

- Seems more persuading
- When the previous query is very effective, current query can be seemingly very “effective” by returning similar results, or even the same results.
Novelty in a Search Session

- A brief conclusion:
  - The users may need such system supports
  - Although users may lose control on the systems that explicitly discounting duplicate results, at least we can provide such support and let users decide whether to use it.
  - We may need to find a balance between the “risky” method and the “conservative” method
Overall Ranking Framework

A language modeling approach:

- q: the latest search query
- d: a document
- s: session contexts, e.g. previous queries, clicks
- \( P(q|d,s) \): topical relevance of d to q in the session s
- \( P(d|s) \): current usefulness of d given the past session context s

\[
P(d|q,s) \propto P(q|d,s) \cdot P(d|s)
\]
Topical Relevance: $P(q|d,s)$

$$P(q \mid d,s) \propto P(q,s \mid d,s) = \sum_{t \in \theta_{q,s}} P(t \mid \theta_{d,s}) P(t \mid \theta_{q,s})$$

Estimating session document models and query models

- $\theta_{d,s}$: session document model (here we downgraded to a plain document model with Dirichlet Smoothing [1])

$$P(t \mid \theta_{d,s}) \approx \hat{P}(t \mid \theta_d) = \frac{c(t,d) + \mu \cdot P(t \mid C)}{\sum_{t_i \in d} c(t_i,d) + \mu}$$
Topical Relevance: $P(q|d,s)$

$$P(q|d,s) \propto P(q,s|d,s)^{rank} = \sum_{t \in \theta_{q,s}} P(t|\theta_{d,s})P(t|\theta_{q,s})$$

Estimating session document models and query models

- $\theta_{q,s}$: interpolating different query models

$$\hat{P}(t|\theta_{q,s}) = \left(1 - \lambda_{fb}\right) \cdot \left\{ \left(1 - \lambda_{prev}\right) \cdot P_{MLE}(t|q) + \lambda_{prev} \cdot P_{MLE}(t|q_s) \right\}$$

$$+ \lambda_{fb} \cdot P_{fb}(t|\theta_{q,s})$$

- $P_{MLE}(t|q)$: current query’s MLE model (RL1 run)
- $P_{MLE}(t|q_s)$: previous queries’ MLE model (RL2 run)
- $P_{fb}(t|\theta_{q,s})$: relevance feedback query model
  - RL3: $P_{fb}(t|\theta_{q,s})$ is RL2 run’s pseudo-relevance feedback query model
  - RL4: $P_{fb}(t|\theta_{q,s})$ is the clicked-document query model
Topical Relevance: $P(q|d,s)$

- This part is nothing fancy, simply the same methods we adopted last year.
  - Similar methods have been adopted by many groups since the first year

Key to the high ad hoc search performance

- Waterloo spam filtering
  - only retrieve documents with spam scores $\geq 70$
- Well tuned weights between different query models
  - Especially the weight on previous queries
- All the parameters are in the notebook paper
Topical Relevance: \( P(q|d,s) \)

Two runs using only topical relevance

<table>
<thead>
<tr>
<th>Runs/Methods</th>
<th>Topical Relevance</th>
<th>Browsing Novelty</th>
<th>SDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PITTSHQM</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>PITTSHQMsdm</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>PITTSHQMnov</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>PITTSHQMsnov</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>RL1</th>
<th>RL2</th>
<th>RL3</th>
<th>RL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PITTSHQM</td>
<td>0.2558</td>
<td>0.3100</td>
<td>0.3221</td>
<td>0.3153</td>
</tr>
<tr>
<td>PITTSHQMsdm</td>
<td>0.2615</td>
<td>0.3071</td>
<td>0.3103</td>
<td>0.3103</td>
</tr>
<tr>
<td>PITTSHQMnov</td>
<td>0.2517</td>
<td>0.3009</td>
<td>0.3152</td>
<td>0.3070</td>
</tr>
<tr>
<td>PITTSHQMsnov</td>
<td>0.2540</td>
<td>0.2966</td>
<td>0.3009</td>
<td>0.3019</td>
</tr>
</tbody>
</table>
Topical Relevance: $P(q|d,s)$

Results (very similar to previous years’ results)

- If the RL2 query model is well tuned, it is difficult to get improvement in RL3 and RL4 query models
  - Not surprising, because RL2-4 give similar information for estimating the query language model
- RL2: previous queries (small sample; little noise)
- RL3: pseudo-relevant documents (larger sample; lots of noise)
- RL4: previous queries’ results being clicked (larger sample than RL2; less noise than RL3)

<table>
<thead>
<tr>
<th></th>
<th>RL1</th>
<th>RL2</th>
<th>RL3</th>
<th>RL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PITTSHQM</td>
<td>0.2558</td>
<td>0.3100</td>
<td>0.3221</td>
<td>0.3153</td>
</tr>
<tr>
<td>PITTSHQMSdm</td>
<td>0.2615</td>
<td>0.3071</td>
<td>0.3103</td>
<td>0.3103</td>
</tr>
</tbody>
</table>
Document Usefulness: \( P(d|s) \)

\( P(d|s) \): the probability that, after several rounds of searches (s), a document d is still informative to the user.

Some intuitions:
- The higher rank of d in previous results, the more likely d has been examined and is useless for user.
- The more previous queries returned d, the more likely d has been examined and is useless for user.
- The user may overlook a document d in browsing.
Document Usefulness: $P(d|s)$

User Model: RBP [1] browsing model
- $q_i$: the $i^{th}$ query in the session;
- $R^{(i)}$: the results of $q_i$.
- The user always examines the first document in $R^{(i)}$.
- After examine a document, the user has:
  - Probability $p$ to continue to examine the next document in $R(i)$
  - Probability $1-p$ to stop examining (either to reformulate or to leave the current session): but for the session track data, we always assume the user will reformulate.

Document Usefulness: $P(d|s)$

- $p$: the probability to continue to examine the next document in $R(i)$
- $P_{\text{examine}}(d|R^{(i)})$: the probability that the user had examined a document $d$ when browsing $R^{(i)}$
- rank($d,i$): the rank of $d$ in results $R^{(i)}$

$$P_{\text{examine}}(d \mid R^{(i)}) = \begin{cases} p^{\text{rank}(d,i)-1} & d \in R^{(i)} \\ 0 & d \not\in R^{(i)} \end{cases}$$

$P_{\text{examine}}(d|R^{(i)})$ depends on $p$ and rank($d,i$), models the intuition:
- The higher rank of $d$ in previous results, the more likely $d$ has been examined and is useless for user
Document Usefulness: $P(d|s)$

User Models: Browsing Novelty [2]

- For each time the user examines a document, it has the probability $\beta$ that the user can understand the information of the document and will not need to see the document again in the same session.
- After a series of searches ($s$), the probability that a document can keep its utility is $P(d|s)$:

$$P(d \mid s) = 1 - \prod_{i=1}^{n-1} \left(1 - \beta \cdot P_{\text{examine}}(d \mid R^{(i)})\right)$$

Document Usefulness: $P(d|s)$

User Models: Browsing Novelty [2]
- Models the other two intuitions:
  - The more previous queries returned $d$, the more likely $d$ has been examined and is useless for user
  - The user may overlook a document $d$ in browsing.
- $P_{\text{examine}}(d|R^{(i)})$ may be replaced by other browsing models

$$P(d \mid s) = 1 - \prod_{i=1}^{n-1} \left(1 - \beta \cdot P_{\text{examine}}(d \mid R^{(i)})\right)$$

Document Usefulness: $P(d|s)$

- The parameter $\beta$ is simply set to a constant value here
- $\beta$ may be further modeled to consider some complex factors:
  - User factors
    - Reading style
    - Careful/careless
    - Users’ background knowledge and familiarity to the topic
  - Session factors
    - Search tasks: exploratory search may have lower $\beta$
    - Search stages: $\beta$ can change during different search stages
  - System & Collection factors
    - Search interface etc.
    - Attractiveness of results
Document Usefulness: $P(d|s)$

Two runs considering both novelty and topical relevance

<table>
<thead>
<tr>
<th>Runs/Methods</th>
<th>Topical Relevance</th>
<th>Browsing Novelty</th>
<th>SDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>PITTSHQMQ</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>PITTSHQMQsdm</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>PITTSHQMQnov</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>PITTSHQMSnov</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Parameters:

$p = 0.8$, $\beta = 0.8$ for all runs in all sessions

<table>
<thead>
<tr>
<th></th>
<th>RL1</th>
<th>RL2</th>
<th>RL3</th>
<th>RL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PITTSHQMQ</td>
<td>0.2558</td>
<td>0.3100</td>
<td>0.3221</td>
<td>0.3153</td>
</tr>
<tr>
<td>PITTSHQMQnov</td>
<td>0.2517</td>
<td>0.3009</td>
<td>0.3152</td>
<td>0.3070</td>
</tr>
<tr>
<td>PITTSHQMQsdm</td>
<td>0.2615</td>
<td>0.3071</td>
<td>0.3103</td>
<td>0.3103</td>
</tr>
<tr>
<td>PITTSHQMSnov</td>
<td>0.2540</td>
<td>0.2966</td>
<td>0.3009</td>
<td>0.3019</td>
</tr>
</tbody>
</table>
Our mistake: our RL1 runs for PITTSHQMQnov and PITTSHQMQmsnov actually used RL2 information
- because $P(q|s)$ used RL2 information

Evaluation (without removing duplicates)
- Discounting duplicate documents slightly hurt the nDCG@10 results
- But for all the runs, the differences are insignificant

<table>
<thead>
<tr>
<th>Using all qrels for evaluation (without removing duplicates)</th>
<th>RL1</th>
<th>RL2</th>
<th>RL3</th>
<th>RL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PITTSHQMQ</td>
<td>0.2558</td>
<td>0.3100</td>
<td>0.3221</td>
<td>0.3153</td>
</tr>
<tr>
<td>PITTSHQMQnov</td>
<td>0.2517</td>
<td>0.3009</td>
<td>0.3152</td>
<td>0.3070</td>
</tr>
<tr>
<td>PITTSHQMQsdm</td>
<td>0.2615</td>
<td>0.3071</td>
<td>0.3103</td>
<td>0.3103</td>
</tr>
<tr>
<td>PITTSHQMQmsnov</td>
<td>0.2540</td>
<td>0.2966</td>
<td>0.3009</td>
<td>0.3019</td>
</tr>
</tbody>
</table>
Document Usefulness: $P(d|s)$

Evaluation (removing all shown duplicates)
- nDCG@10 significantly improved about 7%-10%
- Still large improvements of RL2-4 over RL1

Using all qrels for evaluation (without removing duplicates)

<table>
<thead>
<tr>
<th></th>
<th>RL1</th>
<th>RL2</th>
<th>RL3</th>
<th>RL4</th>
</tr>
</thead>
<tbody>
<tr>
<td>PITTSQM</td>
<td>0.2314</td>
<td>0.2746</td>
<td>0.2877</td>
<td>0.2781</td>
</tr>
<tr>
<td>PITTSQMnov</td>
<td>0.2500*</td>
<td>0.3001*</td>
<td>0.3146*</td>
<td>0.3063*</td>
</tr>
<tr>
<td>PITTSQMsdm</td>
<td>0.2344</td>
<td>0.2650</td>
<td>0.2698</td>
<td>0.2696</td>
</tr>
<tr>
<td>PITTSQMsnov</td>
<td>0.2498</td>
<td>0.2916*</td>
<td>0.2959*</td>
<td>0.2959*</td>
</tr>
</tbody>
</table>
Some preliminary conclusions:

- We can consider novelty issues in a search session without hurting ad hoc search performance
  - On average, it seems there is no much risk of providing users with such system

- Novelty may not be an essential issue in interactive search
  - It seems users can by themselves reformulate very different queries
  - The most fundamental way of improving a system seems still to be aiming at high ad hoc search performance
  - But ….
Some Suggestions on Evaluation

The two novelty evaluation methods this year:

- Discount the relevance of clicked documents in previous results to 0.
  - May be too conservative
  - $P(\text{understand} | \text{clicked})$ may be high, but $P(\text{clicked} | \text{understand})$ may be low

- Discount the relevance of all showed documents in previous searches to 0.
  - May be too radical
  - Some shown results are not examined
  - User may overlook a document at browsing
  - User may be not confident or clear about the information in a document after examine
Some Suggestions on Evaluation

Suggestion 1: Collecting time-sensitive qrels (a model free approach)
Suggestion 2: Estimating the session context sensitive qrels
- Thanks!
- Questions?