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

Jiepu Jiang, Daqing He, Shuguang Han
School of Information Sciences,
University of Pittsburgh
jiepu.jiang@gmail.com, dah44@pitt.edu, shh69@pitt.edu

ABSTRACT
We propose and experiment on an adaptive browsing model that aims at presenting users with relevant and novel results in multi-query search sessions. The model discounts a document's ranking in current search by the likelihood that it has been clearly examined by the user in previous searches within the same session. A document will be penalized to a greater extent if it appears in more of the previous searches or if it is ranked in higher positions in previous searches. We further rank documents by combining the browsing model with ad hoc search models to consider users' browsing novelty within a multi-query search session. Experiments indicate the browsing model can effectively discount the documents found and ranked in higher positions in previous searches, and shuffle novel relevant documents up to higher positions so that the ad hoc search performance will not be hurt.

Keywords
Search session; browsing; novelty; query reformulation.

1. METHODS
The TREC session track defines a search task as follows: in an ongoing multi-query search session that resolves one consistent information need (a TREC topic), how to satisfy the user's latest search query q based on q's information as well as the user's past search behaviors in the session. The dataset provides information such as the user's past queries and results, clicked documents, and turning of pages.

We use a language modeling approach for retrieval. A document d will be ranked by $P(d|q, s)$: q is the target query for search, which is also the latest search query of the ongoing session; s is the user's past search behaviors in the session. As in Eq(1), applying Bayes' theorem, we can equivalently rank documents by the product of $P(q|d, s)$ and $P(d|s)$. We further model $P(q|d, s)$ as d's relevance to the query q in the session context s, and $P(d|s)$ as the novelty of d to the user in the ongoing session.

$$P(d \mid q,s) \propto P(q \mid d,s) \cdot P(d \mid s) \tag{1}$$

1.1 Topical Relevance
Literally, $P(q|d, s)$ suggests a query generation process that q is generated from not only the document d but also the session context s. We can also explain $P(q|d, s)$ as the likelihood that the user issues a query q in the specific session context s for retrieving the document d. Apparently, this suggests an extension to the query likelihood language model (LM) framework (Ponte & Croft, 1998; Chengxiang Zhai & Lafferty, 2004).

Similarly, we can give out an extension to the KL-Divergence LM framework (Lafferty & Zhai, 2001; C Zhai & Lafferty, 2001) in multi-query search session. As in Eq(2), $P(q|d, s)$ is proportional to $P(q, s|d, s)$. Thus, we can estimate two language models $\theta_{q,s}$ and $\theta_{d,s}$, the session contextual query model and document model, and rank documents by the KL-Divergence between $\theta_{q,s}$ and $\theta_{d,s}$. We finally calculate the relevance scores by $\sum_{t \in \theta_{q,s}} P(t \mid \theta_{d,s})^{P(t|\theta_{q,s})}$, which is equivalent to $KLD(\theta_{q,s}||\theta_{d,s})$ in ranking and can be easily implemented using indri query language.

$$\begin{aligned} P(q \mid d,s) &\propto P(q,s \mid d,s) \\ &\overset{rank}{=} \sum_{t \in \theta_{q,s}} P(t \mid \theta_{d,s})^{P(t|\theta_{q,s})} \overset{rank}{=} KLD(\theta_{q,s} \,\|\, \theta_{d,s}) \end{aligned} \tag{2}$$

Although $\theta_{q,s}$ and $\theta_{d,s}$ provide us with interesting opportunities for modeling, this year we only adopt very simple methods for $\theta_{q,s}$ and $\theta_{d,s}$ so that we can focus on studying user's browsing novelty within a session. We simply estimate $\theta_{d,s}$ as θ_d, the plain document language model with Dirichlet smoothing (Chengxiang Zhai & Lafferty, 2004), as in Eq(3). As in Eq(4), we estimate $\theta_{q,s}$ by interpolating different query models: $P_{MLE}(t|q)$ and $P_{MLE}(t|q_s)$, respectively, are models estimated from the latest query q and the past queries q_s by maximum likelihood estimation (MLE); $P_{fb}(t|\theta_{q,s})$ is a relevance feedback query model.

$$P(t \mid \theta_{d,s}) \approx \hat{P}(t \mid \theta_d) = \frac{c(t,d) + \mu \cdot P(t \mid C)}{\sum_{t_i \in d} c(t_i,d) + \mu} \tag{3}$$

$$\begin{aligned} \hat{P}(t \mid \theta_{q,s}) = (1-\lambda_{fb}) \cdot \left\{ (1-\lambda_{prev}) \cdot P_{MLE}(t \mid q) + \lambda_{prev} \cdot P_{MLE}(t \mid q_s) \right\} \\ + \lambda_{fb} \cdot P_{fb}(t \mid \theta_{q,s}) \end{aligned} \tag{4}$$

Specifically, we estimate different query models for RL1-4 runs. RL1 runs only use $P_{ML}(t|q)$. RL2 runs combine $P_{ML}(t|q)$ with $P_{ML}(t|q_s)$. RL3 and RL4 runs interpolate RL2 runs' models with different relevance feedback query models: for RL3 runs, $P_{fb}(t|\theta_{q,s})$ is estimated based on RL2 runs' top ranked results using RM1 relevance model (Lavrenko & Croft, 2001; Lv & Zhai, 2009); for RL4 runs, we estimate $P_{fb}(t|\theta_{q,s})$ as the mixture model of all clicked documents' MLE document models (we assign each clicked document the same weight).

Technically, the topical relevance scores are calculated using exactly the same methods we adopted last year (Jiang, Han, Wu, & He, 2011). Here we show the methods in (Jiang et al., 2011) suggests an extension to the language modeling methods for ad hoc search (C Zhai & Lafferty, 2001; Chengxiang Zhai & Lafferty, 2004) in multi-query search session.

1.2 Browsing Novelty
We model the user's browsing novelty in a multi-query session by $P(d|s)$, which can be explained as: the probability that the user, after several rounds of searches and interactions (s), will still be interested in examining d.

A document may lose its attractiveness for at least two reasons: first, it was examined by the user in past searches; second, other documents examined previously contain the same or very similar information. We focus on the first type of novelty due to the lack

of information for studying and evaluating the second type (e.g. mapping between documents and sub-topics).

We assume the following models for the user's behaviors prior to the current query q:

M1: The user examines results in a list by sequence. The user will always examine the first result in a list. After examine each result, the user has probability p to continue examining the next one, and probability $1 - p$ to stop (either to reformulate a new query for search or to terminate the current session).

M2: For each time the user examines a result, it has probability β that the result will lose its attractiveness to the user in the rest of the search session.

Here, M1 models user's browsing behaviors in a search session. We adopt the same browsing model used in rank-biased precision (RBP) (Moffat & Zobel, 2008). A similar model has been adopted in (Kanoulas, Carterette, Clough, & Sanderson, 2011) for evaluating a whole search session's performance. However, M1 differs from the model in (Kanoulas et al., 2011) in that we do not count any probability for the case that the user terminates the session prior to q (as modeled by p_{reform} in (Kanoulas et al., 2011)). We believe, in a static session dataset such as those in TREC session track, we can only observe the static session data based on the fact that the user had chosen to reformulate. Thus, it seems inconsistent for (Kanoulas et al., 2011) to consider p_{reform} in such datasets. If the user terminates the session prior to q, we will not be able to observe the static session data.

M2 is not a model on the process that a document loses its attractiveness. But M2 roughly models the effects of many complex user factors in interactive search, for example:

- **Users' browsing styles and efforts**: some users may quickly scan results, while some others may carefully examine one by one. Users of different styles may have different chances of missing important information in a document, in which case, the document does not lose its attractiveness.
- **Users' background knowledge and familiarity with the topic**: a user's background knowledge and familiarity with the topic may influence whether, after examining a result, the user can understand the major information in the result.

According to M1 and M2, as in Eq(5), a document d can keep its attractiveness if and only if it did not lose attractiveness in any of the previous searches. In Eq(5): $R^{(i)}$ refers to the results for the ith query in the session (assuming q is the nth query); $P_{\text{examine}}(d|R^{(i)})$ is the probability that d will be examined when the user browses results $R^{(i)}$, as calculated in Eq(6); rank(d, i) is the rank of d in $R^{(i)}$.

$$P(d \mid s) = 1 - \prod_{i=1}^{n-1} \left(1 - \beta \cdot P_{\text{examine}}(d \mid R^{(i)})\right) \quad (5)$$

$$P_{\text{examine}}(d \mid R^{(i)}) = \begin{cases} p^{rank(d,i)-1} & d \in R^{(i)} \\ 0 & d \notin R^{(i)} \end{cases} \quad (6)$$

A document's attractiveness will be discounted to a greater extent if: the document appears in more of the previous queries' results; the document is at higher positions in previous queries' results; a greater value of either p or β. Let $\{d_1, d_2, \ldots, d_{10}\}$ be a result list of 10 documents. Figure 1 shows $P(d_i|s)$ for the 10 documents after the user viewed the result list $\{d_1, d_2, \ldots, d_{10}\}$ once. We use a similar model in (Jiang, He, Han, Yue, & Ni, 2012) for evaluating query reformulations in a search session.

2. EXPERIMENTS

We submitted 4 runs, as summarized in Table 1. The parameter settings are summarized in Table 2. We implement Eq(2) using Indri's query language. We build index and search on a subset of Clueweb09b dataset for only those documents that have Waterloo spam rank scores ≥ 70.

- PITTSHQM: only consider topical relevance for search; only individual terms are used.
- PITTSHQMsdm: only consider topical relevance for search; use sequential dependence model (Metzler & Croft, 2005) and combine individual terms, ordered phrases, and unordered phrases for search.
- PITTSHQMnov: consider both topical relevance and browsing novelty; only individual terms are used.
- PITTSHQMsnov: consider both topical relevance and browsing novelty; use sequential dependence model (Metzler & Croft, 2005) and combine individual terms, ordered phrases, and unordered phrases for search.

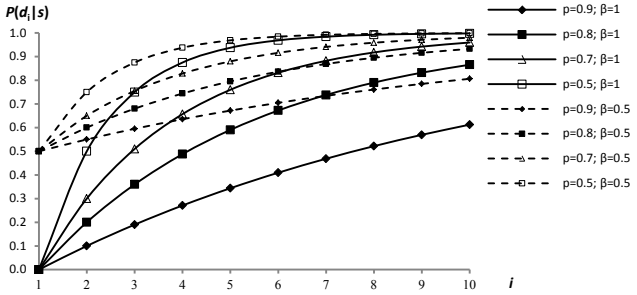

Figure 1. Discounting of results' attractiveness to the user.

Table 1. Summarization of runs.

Runs/Methods	Topical Relevance	Browsing Novelty	SDM
PITTSHQM	Y	N	N
PITTSHQMsdm	Y	N	Y
PITTSHQMnov	Y	Y	N
PITTSHQMsnov	Y	Y	Y

Table 2. Summarization of parameters.

Related Models	Parameter Settings
Document Model	μ = 3,500
Session History Query Model	λ_{prev} = 0.4
Relevance Feedback Query Model	λ_{fb} = 0.2
	# fb docs = 10
	# fb terms = 20
Sequential Dependence Model	w_{term} = 0.85
	$w_{\#\text{ow2}}$ = 0.09
	$w_{\#\text{uw8}}$ = 0.06
Browsing Novelty	p = 0.8
	β = 0.8
Waterloo Spam Rank Scores	Only retrieve >= 70

3. EVALUATION

Table 3 shows the nDCG@10 results for the 4 runs. We find very similar results to what we found last year: search performance can be improved substantially by combining the past search queries with the current query, but further applying relevance feedback query models to RL2 runs seems not helpful.

Our experiments and comparison focus on the adaptive browsing models (PITTSHQM vs. PITTSHQMnov, and PITTSHQMsdm vs.

PITTSHQMsnov). As indicated in Table 3, there is no observable effect of the browsing novelty model on nDCG@10. Thus, we further calculate the rank correlation between results using only topical relevance scores and those using adaptive browsing model.

Figure 2 shows the rank correlation (Kendall's tau) for the top 10 results between PITTSHQM.RL1 and PITTSHQMnov.RL1. In 55 out of 98 sessions, the top 10 results' rankings are affected by the browsing model. Although the top 10 results' rankings for the 55 sessions do change to some degree (with average tau 0.71), the nDCG@10 for the 55 sessions did not change (with the average change in nDCG@10 only -0.007).

We further analyze results and find the reason: the browsing model will discount relevant documents ranked in high positions of previous searches, and shuffle some novel relevant documents to higher positions so that the nDCG@10 scores will not be affected much. The effects are most observable on some topics that the search system is very effective in finding relevant documents. For example, for session 47, 48, 46, 51, and 8, the nDCG@10 scores are 1 (using only relevance scores for search). After applying browsing model, nDCG@10 scores are still 1, while the ranking correlations are, respectively, only 0.20, 0.20, 0.38, 0.64, and 0.71, which indicate the results do changed significantly on rankings.

Table 4 shows the shuffling of results for session 47. A relevant document "clueweb09-enwp01-63-10556" was ranked at the top position in the results of the first query in the session. The document will be discounted to very low positions so that other relevant documents can be shuffled to higher positions.

The results also question whether using and ad hoc search metrics can validly evaluate systems' performance in a session. For example, in the extreme case, a system can be seemingly very effective by simply returning only relevant documents found by the user in previous searches. Such system may achieve very high nDCG@10 scores but is almost useless.

Thus, we also suggest two alternative evaluation methods for the static session search task:

1. An interactive qrels collecting and evaluation method. This method is enlightened by the interactive search and judge method for collecting qrels (Cormack, Palmer, & Clarke, 1998). We can ask the user to search freely in an interactive search system, saving each relevant document when the user ensures the document is relevant, until the user believes there is no more relevant documents in the collection. Finally, we can collect the user's whole search history as a static search session, and the time-sensitive qrels: each relevant document is associated with the time in the session it is marked by the user as relevant. We can extract one query and the query's previous behaviors from the collected static session for testing systems, and evaluate the systems using the qrels collected later than the test query.
2. Using existing dataset and qrels, but modeling on novelty in evaluation metrics, as the *irel*-series metrics we proposed in (Jiang et al., 2012).

Table 3. nDCG@10 of submitted runs.

	RL1	RL2	RL3	RL4
PITTSHQM	0.2558	0.3100	0.3221	0.3153
PITTSHQMsdm	0.2615	0.3071	0.3103	0.3103
PITTSHQMnov	0.2517	0.3009	0.3152	0.3070
PITTSHQMsnov	0.2540	0.2966	0.3009	0.3019

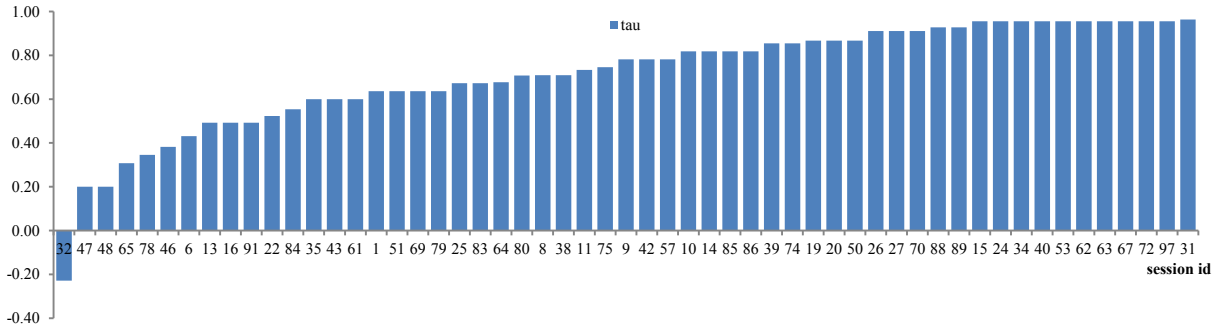

Figure 2. Correlation of top 10 results between PITTSHQM.RL1 and PITTSHQMnov.RL1 (Kendall's tau).

Table 4. Shuffling of results in session #47 after applying browsing novelty model.

q_1 = "pseudocyesis"		q_2 = "pseudocyesis epidemiology"		q = "pseudocyesis history"				
				PITTSHQM.RL1			PITTSHQMnov.RL1	
1	enwp01-63-10556	1	enwp01-23-15772	1	enwp01-63-10556	↓	2→1	enwp00-68-14496
2	en0038-44-08898	2	enwp00-88-14910	2	enwp00-68-14496	↑	3→2	enwp02-13-04273
3	en0013-47-24913	3	en0060-14-21952	3	enwp02-13-04273	↑	4→3	enwp01-83-08322
4	en0121-70-04288	4	en0006-59-10549	4	enwp01-83-08322	↑	8→4	enwp00-86-21481
5	en0047-21-02636	5	en0009-11-14983	5	enwp01-56-06800	↓	10→5	enwp00-94-21656
6	enwp01-80-10554	6	en0011-66-21877	6	enwp01-66-10938	↓	9→6	enwp00-98-19091
7	en0123-83-35172	7	en0074-17-31531	7	enwp01-51-08462	=	7→7	enwp01-51-08462
8	en0063-23-33834	8	en0005-88-05908	8	enwp00-86-21481	↑	5→8	enwp01-56-06800
9	en0065-33-00328	9	en0004-33-02114	9	enwp00-98-19091	↑	6→9	enwp01-66-10938
10	en0092-76-41724	10	en0013-29-10622	10	enwp00-94-21656	↑	12→10	enwp02-21-21481
				…			…	
				12	enwp02-21-21481	↑	1→36	enwp01-63-10556

4. REFERENCES

[1] Cormack, G. V., Palmer, C. R., & Clarke, C. L. A. (1998). Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 282–289). New York, NY, USA: ACM. doi:10.1145/290941.291009

[2] Jiang, J., Han, S., Wu, J., & He, D. (2011). Pitt at TREC 2011 session track. In *Proceedings of the 20th Text REtrieval Conference, (TREC 2011)*.

[3] Jiang, J., He, D., Han, S., Yue, Z., & Ni, C. (2012). Contextual Evaluation of Query Reformulations in a Search Session by User Simulation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM'12)*. ACM.

[4] Kanoulas, E., Carterette, B., Clough, P. D., & Sanderson, M. (2011). Evaluating multi-query sessions. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1053–1062). New York, NY, USA: ACM. doi:10.1145/2009916.2010056

[5] Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 111–119). New York, NY, USA: ACM. doi:10.1145/383952.383970

[6] Lavrenko, V., & Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 120–127). New York, NY, USA: ACM. doi:10.1145/383952.383972

[7] Lv, Y., & Zhai, C. (2009). A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 1895–1898). New York, NY, USA: ACM. doi:10.1145/1645953.1646259

[8] Metzler, D., & Croft, W. B. (2005). A Markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 472–479). New York, NY, USA: ACM. doi:10.1145/1076034.1076115

[9] Moffat, A., & Zobel, J. (2008). Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, *27*(1), 2:1–2:27. doi:10.1145/1416950.1416952

[10] Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–281). New York, NY, USA: ACM. doi:10.1145/290941.291008

[11] Zhai, C, & Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management* (pp. 403–410).

[12] Zhai, Chengxiang, & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, *22*(2), 179–214. doi:10.1145/984321.984322

[13]