Rutgers at the TREC 2012 Session Track

Chang Liu, Michael Cole, Eun Baik & Nicholas J. Belkin

School of Communication and Information, Rutgers University

imliuc@gmail.com, m.cole@rutgers.edu, eunjungbaik@gmail.com, belkin@rutgers.edu

1 Introduction

At Rutgers, we approached the Session Track task as an issue of personalization, based on both the behaviors exhibited by the searcher during the course of an information seeking episode, and a classification of the task that led the person to engage in information-seeking behavior. Our general approach is described in detail at the Web site of our project (http://comminfo.rutgers.edu/imls/poodle) and in the papers available there. In the TREC 2011 Session Track, we tested preliminary results of predictive models of document usefulness using recursive partitioning models learned from user studies of task session information behaviors. In this year's TREC Session Track, we tested predictive models of document usefulness based on user behaviors by using logistic regression. This was combined with predictive models of task type derived from a multinomial logistic regression model learned from the 2012 Session Track data.

After an overview of our approach we provide details of how we actually did things, our results, and our conclusions about the results. The Session Track tasks were addressed by first classifying the track task sessions into four types of tasks, using the scheme and method described in section 2. This classification was based on the Session topic descriptions and narratives. Task classification was performed both manually and automatically. We distinguish the two in our results submissions as RutgersHu (human) and RutgersM (machine). The task classifications were used in our experimental runs, where the document usefulness prediction model depended on the identified search task type along with the observed search behaviors. Since the Session Track data did not allow us to incorporate evidence from behaviors on content pages, we used only data associated with SERPs and various temporal characteristics, such as dwell time on content pages, and time between queries (section 3 describes the models and data in detail). The predicted useful documents were then used to modify the last query but one in each search session using the useful documents to supply terms in a standard relevance feedback mode using the Lemur system in remote mode (section 4 describes our methods in detail). Unlike our work in the 2011 Session Track, this year we used only positive feedback in the query expansion.

2 The task classification scheme and method

There are several ways to conceptualize search tasks. Li & Belkin (2008) proposed a holistic faceted approach which features fifteen essential facets. Liu et al. (2010) and Liu, Belkin, Cole & Gwizdka (2011) identified several additional facets to extend Li & Bekin's classification scheme. Liu, Belkin & Cole (2012) present the task facets controlled during our user experiments on behavior-based prediction of document usefulness and some results using those predictive models on TREC Session Track 2011 data. The specific task facets controlled were the Product, Goal, Complexity, Level, and

Naming of the search tasks. The other task facets identified in Li & Belkin (2008) were not manipulated, including Source of task; Task doer; Time (length) Process; Goal (quantity); Interdependence; and Urgency. We generated specific predictive models of document usefulness for each task type, and then compared the differences among these specific models to examine the task facet effect on the specific models of document usefulness.

1.1 Tasks classification for TREC

Our results identified two task facets that influenced the predictors and predictive rules: “Product” and “Goal quality”. These two facets were used to classify the TREC 2012 Session Track topics. The “Product” facet has three values: intellectual, factual and image. Given the nature of the data in the Session Track, we used just the intellectual and factual values to provide the controls in the search tasks. The difference between Intellectual and Factual tasks is that Intellectual tasks produce new ideas or findings (e.g. learn about a topic or make decision based on information collected), while Factual tasks only involve locating facts, data and other information items.

The "Goal quality" facet has two values: specific goal(s) and amorphous goal(s). Goal quality is very similar to the dimension that Ingwersen & Järvelin (2005) proposed as 'well-defined' and 'ill-defined' information need. Tasks with specific goals have well-defined information needs, while in tasks with amorphous goals, the information need is ill-defined. Tasks with an amorphous goal might require users to redefine the topic or identify specific aspects of the subject themselves. Note that we simplify here in terms of facet values; for goal quality, in particular, the values we have specified are more properly viewed as the poles of a dimension of specificity or clarity in the searcher's understanding of the goal, than as a binary distinction. Using these two facets, we would have four types of tasks, as shown in Table 1.

Table 1. Task type classification

	Goal Quality (specific vs. amorphous)		
	Specific	**Amorphous**	
Product (factual vs. intellectual)	**Factual**	**Type A: Known-item search** Factual tasks with specific goal(s)	**Type B: Known-subject search** Factual tasks with amorphous goal(s)
	Intellectual	**Type C: Interpretive search** Intellectual tasks with specific goal(s)	**Type D: Exploratory search** Intellectual tasks with amorphous goal(s)

1.2 Manual classification of tasks types for TREC

The task types of 98 sessions were manually classified by two doctoral students independently according to the classification scheme introduced above. An initial classification was produced after the two coders compared notes and discussed to reach an agreement. A third coder (faculty) confirmed and made minor revisions to the discussion results, which were agreed upon by all three coders. The final manual classification of task types for TREC 2012 Session Track is presented in Table 2.

Table 2. Manual classification of tasks types for TREC

Task type	Task type	Goal(quality)	Product	number of topics	number of sessions
A	**Known-item search**	Specific	Factual	19	40
B	**Known-subject search**	Amorphous	Factual	10	18
C	**Interpretive search**	Specific	Intellectual	9	17
D	**Exploratory search**	Amorphous	Intellectual	10	23
Total				48	98

1.3 Automatic classification of tasks types for TREC

In our previous studies, we generated predictive models of task type using behavioral measures during the search sessions and after the search sessions are completed. In the automatic classification of task types for TREC 2012 Session Track, we simply applied the predictive models of task types as learned from our previous user experiment. The task descriptions in that user experiment were presented in Liu, Belkin, and Cole (2012), and the four task types can be labeled as the four task types we described in Table 1. Task type classification 1. The predictive models of task types are listed below:

$$oddsfortypeB = ln\left(\frac{ProbabilityoftypeB}{ProbablityoftypeA}\right)$$
$$= -1.91 - 0.061 * mean.content.dwell.unique + 0.051 * number.content - 0.004$$
$$* task.completion.time + 0.277 * number.serp + 0.719 * number.serp.per.query$$

$$oddsfortypeC = ln\left(\frac{ProbabilityoftypeC}{ProbablityoftypeA}\right)$$
$$= -0.769 - 0.255 * mean.content.dwell.unique - 0.021 * number.content + 0.005$$
$$* task.completion.time + 0.132 * number.serp + 0.662 * number.serp.per.query$$

$$oddsfortypeD = ln\left(\frac{ProbabilityoftypeD}{ProbablityoftypeA}\right)$$
$$= 0.2403 - 0.073 * mean.content.dwell.unique + 0.08 * number.content - 0.002$$
$$* task.completion.time + 0.021 * number.serp - 0.661 * number.serp.per.query$$

$$odds\ for\ type\ A = 1 - oddsfortypeB - oddsfortypeC - oddsfortypeD$$

These predictive models provided an automatic classification of task types for TREC 2012 Session Track as presented in Table 3. The results show that only two types of tasks were identified by our model: Type A and Type C. Among 98 sessions, 60 were predicted as type A, and 38 were predicted as type C.

Table 3. Automatic classification of tasks types for TREC

Task type	Task type	Goal(quality)	Product	number of sessions
A	**Known-item search**	**S**pecific	Factual	60
B	**Known-subject search**	**A**morphous	Factual	0
C	**Interpretive search**	**S**pecific	Intellectual	38
D	**Exploratory search**	**A**morphous	Intellectual	0
Total				98

3 The prediction models of document usefulness

In this section we provide a description of how we arrived at the models that were used for prediction of document usefulness, and a specification of the models themselves.

Our group implemented implicit relevance feedback to personalize search result content. In particular, we used several prediction models generated from our previous studies to predict the usefulness of the returned documents. This was accomplished through analysis of users' interactions during their search sessions, considering task type as a contextual factor.

RL1 is our baseline run, which used Pseudo Relevance Feedback on the last query issued by users in each session. We used the default parameters for the Indri Retrieval System, as follows:

Parameters for Pseudo Relevance Feedback (RL1)

int fbDocs = _param.get("fbDocs" , 10);

int fbTerms = _param.get("fbTerms" , 10);

double fbOrigWt = _param.get("fbOrigWeight", 0.5);

double mu = _param.get("fbMu", 0);

In RL4, we considered all user interactions available in the log, and used those that were also variables in the prediction models. In Liu, Belkin, Cole and Gwizdka (2011), we examined multiple user interactions on both content pages and search result pages, with respect to document usefulness and task type, and generated several prediction models of document usefulness. Our results demonstrated that combining multiple behaviors on content pages and search result pages can improve the prediction of useful documents. In addition, the specific prediction models for each type of task demonstrated improved prediction results.

User behavioral measures in our prediction models include:

- dwell time on content pages,
- number of times a page has been visited in one search episode (visit_id),
- number of mouse clicks and number of keyboard activities on content pages,
- the total number of content pages visited during that query interval (content_count);
- the total dwell time on content pages during that query interval (content_sum);
- the total dwell time on SERPs during that query interval (serp_sum), and

- the average dwell time on each SERP during that query interval (serp_mean).

Among these behavioral measures, users' interactions on content pages (i.e. number of mouse movements and keyboard activities) are not available in the interaction log of Session Track. Therefore, our submission for RL4 was based on only the available variables in the Session data.

The specific models we used are as follows.

Type A: Known-item search

$$ln\left(\frac{p}{1-p}\right) = -1.59 + 0.03 * dwelltime + 0.07 * visit.id$$

Type B: Known-subject search

$$ln\left(\frac{p}{1-p}\right) = -2.41 + 0.08 * dwelltime + 0.84 * visit.id$$

Type C: Interpretive search

$$ln\left(\frac{p}{1-p}\right) = -0.76 + 0.06 * dwelltime + 0.07\ \text{content\_count} - 0.005 * \text{content\_sum}$$

Type D: Exploratory search

$$ln\left(\frac{p}{1-p}\right) = -1.73 + 0.05 * dwelltime + 0.76 * visit.id - 0.04 * serp.mean + 0.02 * serp.sum$$

These models predict document usefulness. From those documents we selected the most informative terms and used them for query expansion. For both of the RL4 runs, the expanded query terms from the prediction models were added to the penultimate (last-1) queries. The reason for this is that the logs do not contain the user interactions that followed the last query, and so our behavior-based models could not be applied.

4 Queries and runs

RL1 is our baseline run, which used Pseudo Relevance Feedback on the last queries users issued in each session. We used the default parameters in Indri Retrieval System, as described in section 3.

For RL4 in our two submissions, we performed Positive Relevance Feedback using the document usefulness prediction results. This was accomplished by taking the predicted useful documents and calculating the term frequency for each term in the corpus of useful documents for that task session. Each term frequency was then discounted by the prior expectation of the appearance of the term, using the Brown corpus as the English language frequency reference. The terms were stopped using the SMART project stopwords but not stemmed. The top 25 terms in the resulting ranking were used to expand the last-1 query in the session. If the session contained no clicked documents or had only non-useful documents clicked, then we did not conduct any relevance feedback, which means no change to users' original queries.

5 Results

5.1 Average performance of models over all performance measures

We start by comparing the average of the RutgersHu (human-assigned task types) runs with the RutgersM (machine-assigned task types) against one another and the baseline run (RL1). On average, both methods performed better on the mean of all the measures than the baseline run. The RutgersHu method performed somewhat better than RutgerM.

Mean of all measures: RL1 =0.1329 RutgersHu=0.1711 RutgersM=0.1701

5.2 Model performance by evaluation measures

We then compared the improvement of our models (RutgersHu and RutgersM) with our Baseline (RL1) on all the measures. Table 4 shows that both the models performed better than the baseline for err and normalized. In the case of normalized DCG, our models performed better for the top ten (44.10% and 30.42%), but worse for nDCG over all ranks by RutgersHu model (-1.21%). In terms of average precision (ap), our models performed better for the top ten (38.5% and 38.56%), but did not improve much for average precision over all ranks (3.11% and 12.64%). The Human and Machine models had similar performance gains/losses against the baseline run for all performance measures. This provides evidence that the task type assignment model disagreements with the human-assigned task types do not result in materially less effective performance.

Table 4. Performance by measure (all interaction sessions)

	RL1	RutgersHu	RutgersHu (absolute improvement over RL1)	RutgersHu (percentage improvement over RL1)	RutgersM	RutgersM (absolute improvement over RL1)	RutgersM (percentage improvement over RL1)
err	0.0968	0.1246	0.0278	**28.74%**	0.1112	0.0144	**14.91%**
err@10	0.0830	0.1137	0.0307	**37.00%**	0.0997	0.0167	**20.15%**
nerr	0.1605	0.2148	0.0543	**33.84%**	0.1983	0.0378	**23.53%**
nerr@10	0.1351	0.1947	0.0596	**44.10%**	0.1762	0.0411	**30.42%**
ndcg	0.1947	0.1923	-0.0024	**-1.21%**	0.2489	0.0542	**27.86%**
ndcg@10	0.1074	0.1606	0.0533	**49.65%**	0.1505	0.0431	**40.19%**
ap	0.0799	0.0824	0.0025	**3.11%**	0.0900	0.0101	**12.64%**
ap@10	0.2062	0.2856	0.0794	**38.50%**	0.2857	0.0795	**38.56%**

5.3 Comparison to 2011 Session Track performance

Recall that our work this year extended the positive feedback models used in last year's Session Track, in particular they were a test of the document usefulness prediction models that were specialized to task type. As shown in Table 5, compared to the 2011 Session Track, the baseline run performance was about the same for many measures, except for ndcg@10 and average precision where this year (2012)'s result were quite a bit better. Generally our models this year performed about the same as the runs last year, including average precision.

Table 5. Comparison of retrieval performance between this year (2012) and last year (2011)

RL1 (baseline)	2011	2012	absolute improvement by rgpos (2011)	absolute improvement by rspos (2011)	absolute improvement by RutgersHu (2012)	absolute improvement by RutgersM (2012)
err	0.1135	0.0968	0.0324	0.0597	0.0278	0.0144
err@10	0.0990	0.0830	0.041	0.0673	0.0307	0.0167
nerr	0.1730	0.1605	0.0565	0.0896	0.0543	0.0378
nerr@10	0.1482	0.1351	0.0717	0.1039	0.0596	0.0411
ndcg	0.2824	0.1947	-0.0922	-0.0780	-0.0024	0.0542
ndcg@10	0.1069	0.1074	0.0469	0.0739	0.0533	0.0431
ap	0.0731	0.0799	-0.0023	-0.0031	0.0025	0.0101

5.4 Comparison by task type

Since we used task-specific prediction models of document usefulness for implicit relevance feedback, we are interested in the retrieval performance in each type of task. Table 6 compares retrieval performance between the baseline and our manual model and the improvement of our manual model over the baseline by task type.

Table 6. Comparison of retrieval performance between baseline and our manual model by task type

	A (N=40)			B (N=18)			C (N=17)			D (N=23)		
	RL1	Hu. RL4	improvement	RL1	Hu. RL4	improvement	RL1	Hu. RL4	improvement	RL1	Hu. RL4	improvement
err	0.100	0.101	**1%**	0.081	0.097	**20%**	0.058	0.100	**74%**	0.134	0.209	**56%**
err@10	0.086	0.089	**3%**	0.068	0.086	**26%**	0.041	0.089	**117%**	0.122	0.201	**65%**
nerr	0.164	0.172	**5%**	0.160	0.212	**32%**	0.102	0.163	**59%**	0.200	0.336	**68%**
nerr@10	0.137	0.151	**10%**	0.135	0.189	**40%**	0.073	0.142	**94%**	0.179	0.320	**78%**
ndcg	0.219	0.192	**-12%**	0.155	0.182	**17%**	0.134	0.123	**-9%**	0.230	0.254	**10%**
ndcg@10	0.109	0.130	**20%**	0.097	0.156	**62%**	0.043	0.102	**139%**	0.164	0.265	**62%**
ap	0.091	0.074	**-18%**	0.055	0.079	**43%**	0.042	0.046	**9%**	0.109	0.128	**18%**
ap@10	0.225	0.235	**4%**	0.194	0.267	**37%**	0.100	0.247	**147%**	0.264	0.423	**60%**
Average of all measures	0.141	0.143	**1.6%**	0.118	0.159	**35%**	0.074	0.126	**79%**	0.175	0.267	**52%**

First of all, it is shown that our manual model improved over the baseline on nearly all the measures of the retrieval performance, except nDCG and average precision (ap). This is similar to our general results.

Secondly, we compared the retrieval performance by the baseline model (RL1) among different types of tasks. The results show that type D (Exploratory search) tasks achieved best performance improvement over the baseline, followed by type A (Known-item search) and type B (Known-subject search). Type C (Interpretive search) had the worst performance in the baseline model, which used pseudo relevance feedback technique to refine users' queries. This result indicates that pseudo relevance feedback may not be a good technique for Interpretive search (type C) tasks.

Thirdly, we found the amount of improvement was different for different types of tasks. Our model achieved the greatest improvements over the baseline for task type C, Interpretive search (Intellectual search with specific goals), on some measures (err, err@10, nerr, nerr@10), and the improvement was even greater than 100%. Our model also achieved much improvement over the baseline in task type D (about 50%), and in task type B (about 30%). As we just pointed out, type D tasks also had the best performance in the baseline run (RL1), and our behavioral model improved the retrieval results even more, and the absolute retrieval performance was best in type D tasks among all tasks. Comparatively, our model achieved least improvement for task type A, Known-item search (Factual search with specific goals), with improvement less than 20%.

6 Discussion

Our baseline run used Indri's default pseudo-relevance feedback on the penultimate query in each session. The experimental runs used our document usefulness prediction models that were developed from our PoODLE user studies of information behaviors when users engage different types of tasks. This year we applied the logistic regression predictive models on the new interaction sessions, which had much more diversity in task types as compared to last year. We expected to see improvements in the performance of the models over baseline and different retrieval performance in different types of tasks.

Generally, however the performance improvements over baseline were rather similar to last year's results. The baseline performance was a bit better this year than last. The 2011 Session Track baseline run was rather low when compared to the other systems. This year the baseline run was at the median of the runs in the Session Track. Last year the specific model did not really face a classification problem as the interaction sessions were overwhelmingly of a single type, and we found that our model generally improved our baseline. This year there was greater diversity and the performance of the models was also shown to improve the baseline, as measured by nearly all the measures of retrieval performance, except nDCG and Average Precision on all ranks. This is encouraging from the perspective that the performance improvement of our document usefulness prediction models is confirmed with a new, more diverse interaction data set. Even though the baseline performance was better this year we saw roughly the same absolute gains as last year, so this is an encouraging sign that our technique will provide a performance boost to any general retrieval engine. It is important to note that the usefulness prediction models were used as input to a simple relevance feedback technique. The models can be used in more sophisticated ways that could result in greater improvements in overall system performance.

More interestingly, we found a task type effect on the retrieval performance by implementing our predictive models. As shown in our results, our model achieved most improvement over the baseline in task type C, the Interpretive Search, Intellectual search with specific goals; while our model achieved least improvement over the baseline in task type A, the Known-item Search, Factual search with specific goals. But we also notice that the baseline for Interpretive search (task type C) was much lower than that in other task types. This may indicate that the pseudo relevance feedback could not work well for Interpretive search, and our model was able to improve the retrieval performance greatly. However, our model did not have much improvement over the baseline in Known-item search (task type A). This

may indicate that our model could work as well as pseudo relevance feedback in Known-item search type of tasks. We also found that even though the baseline for task type D, Exploratory search, was the best compared with other types of tasks, our model was still able to improve the baseline by about 50%. Therefore, it is very important to implement our model rather than pseudo relevance feedback (the baseline), especially for tasks that are not Known-item search type of tasks.

7 Conclusion

Our results have shown that the task-specific document usefulness prediction models which were developed from radically different search sessions than those represented in the TREC Session Track, nevertheless led to consistently improved performance over a reasonable baseline that did not take account of session-level information. We also found our model was able to improve retrieval performance over pseudo relevance feedback in task types that are not Known-item search. Since the current search systems work best for Known-item search, our results leads us to believe that it is very important to detect the type of tasks users are engaging in when using search systems, and the models we have developed could be used for personalization of retrieval with more practical value.

8 Acknowledgments

The research that provided the data for the user models in this work was funded by IMLS grant LG-06-07-0105-07. We thank all of the members of the PoODLE research team, without whose efforts this work could not have been accomplished.

9 References

Belkin, N. J., Carballo, J. P., Cool, C., Lin, S., Park, S. Y., Rieh, S. Y., et al. (1998). Rutgers' TREC-6 interactive track experience. *Proceedings of the Sixth Text REtrieval Conference*, 597-610.

Li, Y. & Belkin, N.J. (2008). A faceted approach to conceptualizing tasks in information seeking. Inf. Process. Manage. 44, 6 (November 2008), 1822-1837. DOI=10.1016/j.ipm.2008.07.005 http://dx.doi.org/10.1016/j.ipm.2008.07.005

Liu, C., Gwizdka, J., Liu, J., Xu, T., and Belkin, N.J. (2010). Analysis and evaluation of query reformulations in different task types. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47* (ASIS&T '10), Vol. 47. American Society for Information Science, Silver Springs, MD, USA, Article 17 , 10 pages.

Liu, C., Belkin, N.J. & Cole, M. (2012). Personalization of Search Results Using Interaction Behaviors in Search Sessions. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*. ACM, New York, NY, USA, 205-214.

Liu, C., Belkin, N.J., Cole, M., Gwizdka, J. (2011). Personalization of Information Retrieval in Different Types of Tasks. Presented at the *Workshop on Enriching Information Retrieval (ENIR 2011), July 28, 2011, Beijing, China. http://select.cs.cmu.edu/meetings/enir2011/papers/liu-belkin-cole-gwizdka.pd*