Rutgers at the TREC 2011 Session Track

Chang Liu, Si Sun, Michael Cole & Nicholas J. Belkin
School of Communication and Information, Rutgers University
imliuc@gmail.com, sisun@eden.rutgers.edu, m.cole@rutgers.edu, 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, and in the papers available there (http://comminfo.rutgers.edu/imls/poodle); in this section, we give an overview of our approach and how we applied results from our previous studies to the TREC 2011 Session Track. Subsequent sections give details of how we actually did things, our results, and our conclusions about the results.

The PoODLE project aimed to develop a "personalization assistant", a client-side application which would monitor the behaviors of a single person on all that person's computing devices, including, but not limited to information-seeking behaviors, and on the basis of these data, construct a model of the person which would be used either to modify the person's queries to search engines, and/or to modify the results of the queries returned by the search engines. The most fundamental aspect of the personalization is determination of the task or goal which led the person to engage in information seeking behavior. Other aspects that are important include determination of the person's degree of knowledge of both the task and of the search topic, and the person's cognitive abilities. The intention is to make these determinations implicitly, through the evidence of past and current behaviors. To this end, the PoODLE project conducted a series of user studies/experiments, in which we controlled the tasks in which the participants were engaged, conducted psychometric tests to judge two cognitive abilities, and elicited, in various ways, estimates of the participants' knowledge of the tasks and topics of the searches they were asked to perform. In these studies, we logged a large variety of searcher behaviors, ranging from eye-tracking on search engine result pages (SERPs) and content pages, to querying and temporal behaviors of various sorts. The analysis of the data thus collected was aimed at determining associations among the different behaviors (our dependent variables) and the task, knowledge and cognitive abilities information we controlled or elicited (our independent variables), and on the influence of the independent variables and the predictive power of the dependent variables on searcher evaluation of the usefulness of documents with respect to the search task.

The end result of our PoODLE studies has been the generation of several models for prediction of document usefulness, some based solely on behavioral evidence during the searching process, and some modified according to task type, and/or knowledge. Prediction of document usefulness is then to be used as the basis for either query modification using relevance feedback, or search result re-ranking, based on similarity to predicted useful documents. Since our prediction models were generated on the basis of a relatively small number of searches (typically four searches, by each of between 32 and 40 participants, in two or three studies), on quite specific and controlled search task types, in both TREC genomic track tasks and in uncontrolled searching in the Web, and on behaviors on both SERPs and content pages, it is not clear how these models will work with the data available for the TREC Session Track. So, our general aim in this Track is to discover whether our models will work on these different task types, and with this different type of data, and if not, why not.

We addressed the Session Track tasks as follows. First, we manually classified the 76 Sessions by task facets, using the scheme and method described in section 2, based on the Session topic descriptions and narratives. This information was then used for one of our experimental runs, in which the prediction model was specific to each search task type, combined with search behaviors. For the other runs, we used our so-called "general" prediction models, which are based on different search behaviors, without reference to task type. 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 prediction of both useful and not useful documents was then used to modify the last query but one in each search session in a standard relevance feedback mode, one run with positive only feedback, one with positive and negative relevance feedback, using the Lemur system in remote mode (section 4 describes our methods in detail). The results of this modification are compared against the results of our baseline search using the last query in the search session.

2 The task classification scheme and method

2.1 Overview

The classification of task types as an important factor influencing information seeking and search has gained increasing attention recent years (Li & Belkin, 2008). Among the various ways to conceptualize search tasks, Li & Belkin (2008) proposed a holistic faceted approach which featured fifteen essential facets. Liu et al. (2010) focused Li's search task classification scheme on three facets, namely product, goal, complexity, and added another facet - level - to Li's work. Findings indicate that these task facets were associated with search behaviors including task completion time and decision time. Liu, Belkin, Cole & Gwizdka (2011) identified naming as an additional facet for usefulness prediction models based on Liu et al. (2010).

Built on the previous work introduced above, the Rutgers TREC Session Track examined the Product, Goal, Complexity, Level, and Naming of the search tasks (defined in Table 1) while holding constant the other task facets identified in Li & Belkin (2008), including Source of task; Task doer; Time (length) Process; Goal (quantity); Interdependence; and Urgency. For TREC session data, we found only tasks with factual products and no task with combined goals.

Table 1. Facets of task which were varied for the TREC Session Track

Facets	Values	Operational Definitions/Rules
Naming	Named	A task locating factual information to confirm or disconfirm named fact
	Unnamed	A task locating factual information to about unnamed fact
Product	Factual	A task locating facts, data, or other similar items in information systems
	Mixed	A task which produces new ideas or findings on the basis of locating facts, data, or other similar items in information systems
	Intellectual	A task which produces new ideas or findings on the basis of locating facts
Level	Document	A task for which a document as a whole is judged
	Segment	A task for which a part or parts of a document are judged
Goal (Quality)	Specific goal	A task with a goal that is explicit and measurable
	Combined goal	A task with both concrete and amorphous goal
	Amorphous goal	A task with a goal that cannot be measurable

2.1.1 Level

Judgment for task level is mainly based on questions asked in the field "Narratives". Tasks with only one question or with a second or further questions asking for more details on the first one were classified as being at Document level. For example, Session No. 28 asked "What are the origins of nicknames? Is there a religious, cultural, or ethnic factor? Not relevant are websites with searches for nicknames. Specific nicknames are not relevant". The second question is a specification of the first question. Rather than asking for more information, the second question narrows down or provide examples of answers for the first question. Session No.76 required searchers to "Give me any information on glenohumeral subluxation, including pictures". Because searchers participating in TREC were permitted to end their search when they found any piece of information on the topic, sessions whose descriptions include only one question were answerable with information from a single document.

Sessions with multiple questions in their "Narratives" usually require searching for pieces of segment information from one or multiple documents/sources. Session No.4, for example, asked for information on infections in the groin: "How are they caused? What treatments are there? Can the infections be painful? Are any contagious? Are any groin infections strictly gender related?" Four aspects of the disease including cause, treatment, symptoms, and characteristics were asked in the five questions which require capturing of segments of information.

2.1.2 Goal

A task is classified as "specific" when it specifies at least one aspect of the object in the field "Description" and "amorphous" when "all information" or "any information" is required about a subject, and does not specify any aspects in the "Description" field. Session 76, for example, asked for "any information on glenohumeral

subluxation" which makes the task goal unclear. Participants might look for definitions on the concept, discussion on the concept, or any other pieces of information related to the concept. Session No.23, on the other hand, has specific goals including searching for information on the time, reason, objective, requirements, target contestants, and prize structure of "dupont science essay contest".

2.1.3 Naming

Distinguishing between named and unnamed tasks is based on interpretation from the field "Description". For example, the task used in Session No.1 was described as "Find information about the peace corps". Searchers locate factual information about the named fact "peace corps". Session No.1 was thus classified as "Named". Session No. 22 (described as "Why do people get shoulder joint pain?"), on the other hand, asked for the causes of shoulder joint pain which is not named in the description – searchers have to infer what information is needed in order to answer this question; searching on definition of shoulder joint pain is not likely to result in retrieval of relevant information.

2.1.4 Product

Decision for the product of tasks was made based on questions listed in the field "Narratives". For example, in Session No.3, questions asked included "What are the treatments for renal cell cancer? Which treatments are experimental? What organizations are doing research for treatments?" should result in factual products: treatment, experimental treatment, organizations researching on treatments respectively. Unfortunately, none of the tasks led to mixed product, as all of them ask for information that requires no intellectual processing.

2.2 Tasks types for TREC

The task types of 76 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 classification is presented in Table 2.

Table 2. Variable facet values

Task type	**Task type (abbreviation)**	**Level**	**Goal(quality)**	**Naming**	**Product**	**number of tasks**
A	SSN	**S**egment	**S**pecific	**N**amed	Factual	**66**
B	SSU	**S**egment	**S**pecific	**U**nnamed	Factual	**4**
C	DSN	**D**ocument	**S**pecific	**N**amed	Factual	**4**
D	DAN	**D**ocument	**A**morphous	**N**amed	Factual	**2**
						76 (total)

All tasks were expected to have Factual products. A majority of the tasks (66) were at a Segment level, with Specific goal(s), and Named. The other three task types add up to 10 tasks together.

3 The prediction models

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 an implicit relevance feedback method to personalize search results. In particular, we used several prediction models generated from our previous studies to predict the usefulness of documents returned by queries through analysis of users' interactions during their search sessions, moderated by considering task type as a contextual factor.

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 the Indri Retrieval System, as follows:

```
Parameters for Pseudo Relevance Feedback:
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 );
```

RL2 only considered queries in the search session. In Liu, et al. (2010), we analyzed and evaluated the effectiveness of query reformulations in different types of tasks. To categorize users' reformulation of queries in search sessions, we created an algorithm to detect the length and term changes in two successive queries and marked the type of query reformulation for all the queries excluding the last queries, since users did not reformulate queries after the last queries. The categories of reformulation are described in Liu, et al., 2010. One of our results is that after visiting and saving a useful web page, Generalization was less likely to be used while New Query was more likely to be used. In the Session Track log, there were 190 queries excluding the last queries, and we marked their reformulation types using our exiting algorithm. Among them, 51 were marked as Generalization using our algorithm, and they were regarded as "not good" queries; all other queries were regarded as "good" queries. To generate the personalized good queries, we combined users' last queries with all "good" queries in each session, and then selected only distinct query terms from them as the final queries for each session to generate the personalized results for RL2.

Our RL3 run is based on the "good" and "not good" queries selected in RL2. In RL3, we regarded all documents on search result pages (SERPs) under "good" queries as "useful documents", and all documents on SERPs under "not good" queries as "non-useful documents". These results were used to do positive and negative Relevance Feedback to generate the expanded queries.

In RL4, we considered all user interactions available in the log, and used several of them that had been shown to be included in the prediction models in our PoODLE project. In one of our previous efforts, 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 could improve 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*); time to first click after issuing a query (*time_to_first_click*); number of mouse clicks and number of keyboard activities on content pages; the total dwell time on SERPs during that query interval (*serp duration*); the proportion of time on content pages of the total dwell time during that interval (*prop_content*); the total number of content pages visited during that query interval (*content_count*); and, the difference between the dwell time on a content page, and the average dwell time on all content pages during its associated query interval (*diff_content*). 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, we only considered the other available variables in the prediction models for RL4 in our submissions.

The general model we use is described below. This prediction model was used to generate results for RL4 in submission 1: Rgposneg (general model with pos/neg RF).

rule_general:

if visit_id > 1, then it is a useful page;

if dwell time > 28.55 seconds (this is the median of dwell time) then it is a useful page;

else if time-to-first-click > 6.33 seconds and time-to-first-click <14.55 seconds [this is the median], then it is a useful page;

else non-useful pages.

We have existing prediction models for two of four types of tasks we identified in the 76 sessions in the Session track (Table 2). One is task type "SSN", which has task facets of Segment level, Specific goal, Named information objects..This model is called CPE[1]. The other is task type "DAN", which has task facets of Document level, Amorphous goal and Named information objects. This model is called OBI. These two types of tasks covered 68 sessions of the 76 TREC sessions (89.5%). We used the general model for all other types of tasks. We selected the median value of all probabilities in the results as the cutoff point, which was 0.3.

[1] These models are based on the different tasks we asked searchers to perform in our initial studies. CPE is a "copy editing", or fact-checking task, and OBI is the task of writing an advance obituary.

The specific model we use is shown below. We then performed two types of Relevance Feedback on the prediction results: positive and negative Relevance Feedback to generate submission 2; Rsposneg (specific model with positive and negative RF). For submission 3: Rspos (specific model with positive RF only), we performed only positive Relevance Feedback on the prediction results.

rule_specific:
if type = SSN, model CPE_L;
Model CPE_L: log(p/1-p) = –1.70 + 0.04 * dwelltime-0.01 * time_to_first_click + 0.23 * prop_content
if type = DAN, model OBI_L;
Model OBI_L: log(p/1-p) = – 3.27 + 0.10 * dwelltime+ 0.01 * time_to_first_click – 0.01 * diff_content
else rule_general
* cutoff point is 0.3, the median value of all probabilities in the results.
* prop_content: the proportion of time on content pages of the total dwell time during that interval.
* diff_content: the difference between the dwell time on a content page and the average dwell time on all content pages during its associated query interval

For all of the RL4 runs, the expanded query terms from the prediction models were added to the last-1 queries. The reason for this is that there is no information about the user interaction on the last queries in the log, and our models could only take account users' interactions through last-1 queries. Thus, we compared the final user query results, with the results of our modification of the last query-1.

4 Queries and runs

This section describes the construction of queries for each method in each run, how they were submitted to Lemur, and what we did to the results.

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 follows:

Parameters for Pseudo Relevance Feedback:
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);

For RL4 in our three official submissions, we did two types of Relevance Feedback (RF) on the prediction results: both positive and negative RF. Submission 1 is called *Rgposneg* (general model with pos/neg RF); submission 2 is called *Rsposneg* (task type-specific model with pos/neg RF); submission 3 is positive RF only and is called *Rspos* (task type-specific model with positive RF only).

From the prediction of the usefulness of documents (as described in section 3), we calculated the term frequency for each term in the corpus of useful documents, and in the corpus of non-useful documents. The observed (i.e. predicted) term frequency was then discounted by the prior of the expectation of appearance in a random document in the language using the Brown corpus.

With respect to the number of useful and non-useful terms for query expansion, we used the approach described in TREC-6 RU (Belkin et al, 1998), in which a negative RF system was implemented. The number of suggested feedback terms was determined by the formula:

5n + 5, where n = number of judged documents to a maximum of 25 suggested terms.

The query was parsed as a weighted sum, using the default weighting for RF term addition for positive terms, and adding the negative terms under the InQuery "NOT" operator, with 0.6 weight.

Two relevance feedback methods were implemented:

Positive relevance feedback only. In the runs with positive RF only, the predicted "useful" documents were used to calculate the term frequency and the top 25 terms were selected to be useful terms and then expanded with the last-1 queries in the session.

Both positive and negative relevance feedback. In the runs with both positive and negative RF, the predicted "useful" documents were used to calculate the term frequency for "useful" terms and the top 15 terms were selected to be "useful" terms; the predicted "non-useful" documents were used to calculate the term frequency for "non-useful" terms and the top 10 terms were selected to be "non-useful" terms. We then combined the last-1 queries with the 25 "useful" terms (with weight 1.0), and the 10 "non-useful" terms (with weight 0.6) using the Indri query language.

If the session had only useful documents clicked, then only the useful documents were considered to select "useful" terms to accomplish positive RF. If the session had only non-useful documents clicked, then only the non-useful documents were considered for selecting "non-useful" terms to do negative RF. If the session contained no clicked documents, then the SERP documents were used to supply the documents for the corpus of "non-useful" documents.

5 Results

5.1 Mean of our results

We first calculated the average performance of each of our models on all measures. For most measures, the performance of RL2 was better than RL1, RL3 was better than RL2, and RL4 was better than RL3. Among the three RL4 models, Rgposneg and Rgpos (prediction based on the complete general model) was better than those generated from task-specific models (Rsposneg and Rspos). Comparing between Rgposneg and Rgpos, Rgpos (general model with only positive relevance feedback) performed a bit better than Rgposneg (general model with both positive and negative relevance feedback); and comparing between Rsposneg and Rspos, we found that Rspos (task-specific model with only positive relevance feedback) was better than Rsposneg (task-specific model with both positive and negative relevance feedback). Comparing between Tables 3 and 4, we see that our models performed better when all subtopics were considered for evaluation than when only subtopics in last query were evaluated.

Table 3. TREC 2011 allsubtopics Evaluation

Run Measure	RL1 (Baseline)	RL2	RL3	Rgposneg RL4	Rgpos RL4	Rsposneg RL4	Rspos RL4
err	0.1915	0.2131	0.2417	0.2647	0.2712	0.2468	0.2534
err@10	0.1779	0.1999	0.2367	0.2592	0.2656	0.2400	0.2465
nerr	0.2952	0.3346	0.3896	0.4360	0.4471	0.3930	0.4043
nerr@10	0.2738	0.3151	0.3846	0.4307	0.4415	0.3840	0.3950
ndcg	0.2939	0.3213	0.2160	0.2418	0.2503	0.2401	0.2488
ndcg@10	0.1970	0.2297	0.3030	0.3395	0.3442	0.3053	0.3100
ap	0.0868	0.1002	0.0778	0.0865	0.0890	0.0833	0.0858
gap	0.0807	0.0943	0.0753	0.0885	0.0911	0.0851	0.0877

Table 4. TREC 2011 lastquerysubtopics Evaluation

Run Measure	RL1 (Baseline)	RL2	RL3	Rgposneg RL4	Rgpos RL4	Rsposneg RL4	Rspos RL4
err	0.1135	0.1325	0.1407	0.1430	0.1459	0.1664	0.1732
err@10	0.0990	0.1200	0.1363	0.1371	0.1400	0.1597	0.1663
nerr	0.1730	0.2039	0.1972	0.2240	0.2295	0.2510	0.2626
nerr@10	0.1482	0.1833	0.1895	0.2144	0.2199	0.2408	0.2521
ndcg	0.2824	0.2953	0.1652	0.1864	0.1902	0.1955	0.2044
ndcg@10	0.1069	0.1257	0.1489	0.1531	0.1538	0.1760	0.1808
ap	0.0731	0.0735	0.0622	0.0701	0.0708	0.0674	0.0700
gap	0.0679	0.0700	0.0596	0.0675	0.0684	0.0663	0.0691

5.2 Improvement over baseline

We then further examined the improvement of our models over our baseline, and between each other. We also compared our results with minimum, median and maximum of all TREC results. We use ERR (Expected Reciprocal Rank), and ndcg@10 for evaluation of our results. ERR was selected because it was based on the "cascade" user model, which is similar to the personalization models we adopted in our method, and ndcg@10 because it is the "basic" evaluation measure for the track.

When calculating the improvement of each of our models over our baseline, we calculated both absolute and percent improvement. In the calculation of percent improvement, the sessions whose baseline measure was 0 (i.e., no relevant documents retrieved) were excluded because we are unsure how to calculate the percent improvement for those sessions.

The average improvements on ERR and ndcg@10 are shown in Table 5. When all subtopics were considered for evaluation, all models achieve some improvement, and the models which were based on all interactions in the session (RL4) achieved much more improvement than RL2 and RL3. Among them, RL4_rgpos had highest mean improvement on both ERR and ndcg@10. When comparing the improvement using Wilcox tests, we did not find significant difference among them.

Table 5. Improvement over baseline model (allsubtopics Evaluation)

Model to be compared with the baseline	ERR		ndcg@10	
	absolute improvement	percent improvement	absolute improvement	percent improvement
RL2	0.02	1.05	0.03	0.77
RL3	0.05	3.68	0.11	2.75
RL4 rgposneg	0.07	6.58	0.14	1.79
RL4 rgpos	0.08	8.04	0.15	2.21
RL4 rsposneg	0.06	6.68	0.11	2.21
RL4 rspos	0.06	6.70	0.11	2.23

When the subtopics of the last query were considered for evaluation, all models achieve some improvement. Among them, RL4_rspos and RL4_rsposneg had highest mean improvement. When comparing the improvement using Wilcox tests, we did not find significant difference among them either.

Table 6. Improvement over baseline model (lastquerysubtopics Evaluation)

Model to be compared with the baseline	ERR		ndcg@10	
	absolute improvement	percent improvement	absolute improvement	percent improvement
RL2	0.02	1.16	0.02	0.66
RL3	0.03	1.60	0.04	1.75
RL4 rgposneg	0.03	6.11	0.05	0.99
RL4 rgpos	0.04	8.9	0.05	1.91
RL4 rsposneg	0.05	7.30	0.07	1.07
RL4 rspos	0.06	7.32	0.07	1.10

5.3 Performance between our different models

When comparing the performance between our different models, we also adopted the similar method of improvement, which is to calculate the improvement from one model to another.

Table 7 shows improvements when the all subtopics were considered for the evaluation on ERR. It demonstrats that all our RL4 models had some improvement over RL2 and RL3. The absolute improvement shows that the improvement from RL4 rgpos over RL2 was the greatest. The percent improvement showed that the improvement from RL4 rgpos over RL4 rgposneg was the greatest.

Table 7. Comparison of performance on err between our different models (allsubtopics, absolute and percent)

	over RL2	over RL3	over RL4 rgposneg	over RL4 rgpos	over RL4 rsposneg
RL3	0.03 (1.69)				
RL4 rgposneg	0.05 (5.04)	0.02 (26.73)			
RL4 rgpos	0.06 (5.86)	0.03 (3.11)	0.01 (55.86)		
RL4 rsposneg	0.03(4.91)	0.01 (11.44)	0.02(0.95)	0.02 (11.07)	
RL4 rspos	0.04 (5.75)	0.01 (11.44)	0.01(0.95)	0.01 (11.08)	-0.01 (-0.00)

Table 8 shows improvements when the all subtopics were considered for the evaluation using ndcg@10. It also demonstrats that all our RL4 models had some improvement over RL2 and RL3. The absolute improvement shows that the improvement from RL4 rgpos and RL4 rgponeg over RL2 was the greatest, and the percent improvement showed RL4 rgpos achieved better improvement over RL2 than RL4 rgposneg.

Table 8. Comparison of performance on ndcg@10 between our different models (allsubtopics, absolute and percent)

	over RL2	over RL3	over RL4 rgposneg	over RL4 rgpos	over RL4 rsposneg
RL3	0.07 (1.72)				
RL4 rgposneg	0.11 (1.45)	0.04 (0.92)			
RL4 rgpos	0.11 (1.91)	0.04 (1.09)	0 (0.44)		
RL4 rsposneg	0.08 (1.29)	0(0.75)	0.03 (0.62)	0.03 (0.97)	
RL4 rspos	0.08 (1.29)	0.01 (0.75)	0.03 (0.62)	0.03 (1.03)	0 (0)

Table 9 shows improvements when the subtopics of last queries were considered for evaluation using ERR. It demonstrats again that all our RL4 models had some improvement over RL2 and RL3, and the most improvement was achieved from RL4rspos over RL2.

Table 9. Comparison of performance on err between our different models (lastquerysubtopics, absolute and percent)

	over RL2	over RL3	over RL4 rgposneg	over RL4 rgpos	over RL4 rsposneg
RL3	0.01 (1.42)				
RL4 rgposneg	0.01 (4.25)	0.00 (28.04)			
RL4 rgpos	0.02 (11.99)	0.01 (61.51)	0.01(8.21)		
RL4 rsposneg	0.03 (4.70)	0.03(10.39)	0.02 (1.52)	0.02 (19.31)	
RL4 rspos	0.04 (6.23)	0.02(10.39)	0.03 (1.52)	0.02 (19.35)	0.01 (0)

Table 10 showed improvements when the subtopics of last queries were considered for the evaluation on ndcg@10. It demonstrated again that all our RL4 models had some improvement over RL2 and RL3, and the models based on the specific model (RL4 rsposneg, and RL4 rspos) achieved more improvements than the other two models based on the general model.

Table 10. Comparison of performance on ndcg@10 between our different models (lastquerysubtopics, absolute and percent)

	over RL2	over RL3	over RL4 rgposneg	over RL4 rgpos	over RL4 rsposneg
RL3	0.02 (1.68)				
RL4 rgposneg	0.03 (0.91)	0 (1.2)			
RL4 rgpos	0.03 (1.17)	0.01 (0.68)	0 (0.35)		
RL4 rsposneg	0.05 (1.32)	0.03 (1.72)	0.02 (0.79)	0.02 (0.09)	
RL4 rspos	0.05 (1.32)	0.03 (1.73)	0.03 (0.79)	0.03 (0.01)	0 (0)

6 Discussion

To recap, we used standard Indri techniques, including pseudo-relevance feedback based on the results of the last query but one in each session to modify the final query, as our baseline performance. For our experimental runs, we used the document usefulness prediction models that were developed from our PoODLE data for Indri relevance feedback to modify the last query but one. We found that, in general, and evaluated by ERR and ndcg@10, all our prediction models led to consistent improvement over our baseline results, and that performance improved monotonically as more data was used by the models (RL4 > RL3 > RL2). However, the absolute results of our techniques are not especially great when compared to median and maximum results for the Track as a whole. However, our baseline technique, to which we applied our prediction models, was itself rather low, compared to the overall Track baselines. Because the data that we used for our improvement algorithms should be applicable to any general retrieval engine, one might expect that our levels of improvement would be applicable to techniques with much higher baseline performance, resulting in higher absolute performance levels. It is also the case that our usefulness prediction models were used as input to quite standard, and rather simple relevance feedback techniques, and that more sophisticated use of the models could result in better overall performance improvement.

It is of some interest that our "general" prediction model led to better performance improvement than out task-specific models. One reason for this result could be that our general prediction model does not depend upon "client-side" data, such as activity on SERPs and content pages, which was unavailable, whereas the task-specific prediction models depend upon such data.

7 Conclusion

Our results have shown that the 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. This positive "transfer" effect leads us to believe that the models we have developed could be used for personalization of retrieval in a variety of searching circumstances, and that we could expect even greater performance benefit when the richer, client-side data that our prediction models depend upon.

8 Acknowledgements

Thanks to David Pane at CMU, who helped us greatly and generously in performing the Indri runs. The research that led to this work was funded by the IMLS, under grant number 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., 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.pdf.

Liu J., Cole M., Liu C., Bierig R., Gwizdka J., Belkin NJ, Zhang J, Zhang. X. (2010). Search behaviors in different task types. In *Proceedings of ACM-IEEE Computer Society Joint Conference on Digital Libraries (JCDL)* 2010. Goldcoast, Australia, June 21-25, 2010.