Table 1: PREVAL metrics computed over the evaluation set.

Proactive System	PREVAL-RR	PREVAL-ρ
RM-PSS	0.0235	0.4955
QP-PSS-U	0.0503	0.4980
QP-PSS-W	0.0503	0.5004

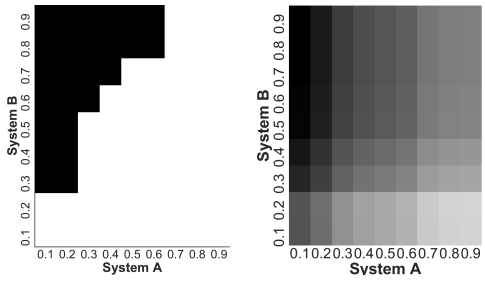

Figure 2: Comparison of PREVAL-RR values of RM-PSS (x-axis) and QP-PSS-W (y-axis), π-index values being shown alongside the axes. i) Left: A black cell indicates that $P_A > P_B$. ii) Right: Each cell plots the value of $\frac{P_A - P_B}{\max(P_A, P_B)} \in [-1, 1]$, darker shades denoting values towards 1 ($P_A > P_B$).

4.2 Experiment Settings and Results

We used a subset of the AOL query log data from the period of Mar'06 to Apr'06 for our experiments, which comprises over $6M$ queries. The first $5M$ queries (preprocessed using a Porter stemmer) were used as a training set to construct a QFG, which was then used in the QP-PSS approaches. A session length of 26 minutes, as prescribed in [6], was used as the adjacency criteria to construct the CFG. As our evaluation set to measure the relative performances of the three PSS systems, we use a random subsample of 50 query sessions. Each session of the evaluation set comprises at least 5 queries (after removing duplicate queries within a session). The average session length in our evaluation set is 8.12.

Table 1 shows the results of applying the two variants of the metric PREVAL on the test set of 50 query sessions. To compute the reward functions, M was set to 10 (see Equation 1). It can be seen that RM-PSS performs the worst in terms of both the RR and ρ-based PREVAL scores. This is to be expected since RM-PSS only uses the information from the current query and the top documents retrieved in response to it. Our metric is also able to rank QP-PSS-W better than QP-PSS-U. This is also expected since the latter considers the contribution from all predicted queries uniformly when constituting the final list of suggested documents. An interesting observation is that PREVAL-ρ is better able to distinguish between QP-PSS-W and QP-PSS-U (in terms of the relative differences between the scores) in comparison to PREVAL-RR. This shows that the rank correlation based reward acts as a better estimate for PSS effectiveness.

Figure 2 shows the relative differences between the PREVAL-RR values for systems A (RM-PSS) and B (QP-PSS-W), i.e., $\delta = \frac{P_A - P_B}{\max(P_A, P_B)}$, for a range of π-index values π_A and π_B. To report an average over sessions of varying lengths, we plot relative proportions of π_A and π_B values within a range of 0.1 to 0.9. A sample cell $(\pi_A, \pi_B) = (0.2, 0.6)$ in both plots of Figure 2 indicates that A started suggesting documents after 20% queries within a session were executed, whereas B waited until 60% of the queries had been executed before initiating proactive suggestions. Each of the $9 \times 9 = 81$ cells of Figure 2 represents an experimental observation. The left plot indicates whether RM-PSS is the winner (black cell). The right plot shows the relative differences between the PREVAL-RR values of RM-PSS and QP-PSS-W. If RM-PSS performs better, this relative difference is close to 1 (represented by darker shades), whereas lighter shades indicate that QP-PSS-W is better.

It can be seen that if QP-PSS-W starts predicting too late in comparison to RM-PSS (cells corresponding to the upper left region of the plots in Figure 2), RM-PSS is the winner. This happens because the rewards accumulated from a small number of proactive suggestions in these cases is not sufficient to outweigh the effectiveness of the early suggestions coming from RM-PSS. This in turn shows that the metric is able to prefer cold-start proactive systems. On the other hand, the system QP-PSS-W (being a more effective system) often wins in the central and the bottom-right part of the plots in Figure 2. This can be seen from the white regions in the left plot, and the lighter shades in the right plot of Figure 2, with the relative differences getting larger towards the bottom-right. This shows that the metric is not overly biased towards cold-start systems and can favour more effective systems with higher π-index values.

5 CONCLUDING REMARKS

Our analysis shows that the proposed metric PREVAL is able to fairly compare between two PSSs that become proactive at different times during search sessions and the the ranking of systems induced by PREVAL conforms to the intuitive expectation. We believe that our proposed PSS evaluation framework will help in developing effective proactive systems in the near future.

Acknowledgement. This work was supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

REFERENCES

[1] D. Billsus, D. M. Hilbert, and D. Maynes-Aminzade. Improving proactive information systems. In *Proc. of IUI'05*, pages 159–166, 2005.
[2] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: Model and applications. In *Proc. of CIKM 2008*, pages 609–618, 2008.
[3] H. Feild and J. Allan. Task-aware query recommendation. In *Proc. of SIGIR 2013*, pages 83–92, 2013.
[4] W. Kong, R. Li, J. Luo, A. Zhang, Y. Chang, and J. Allan. Predicting search intent based on pre-search context. In *Proc. of SIGIR 2015*, pages 503–512, 2015.
[5] V. Lavrenko and B. W. Croft. Relevance based language models. In *Proc. of SIGIR '01*, pages 120–127, 2001.
[6] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Discovering tasks from search engine query logs. *ACM Trans. Inf. Syst. 2013*, pages 14:1–14:43, 2013.
[7] K. Puolamäki, J. Salojärvi, E. Savia, J. Simola, and S. Kaski. Combining eye movements and collaborative filtering for proactive information retrieval. In *Proc. of SIGIR 2005*, pages 146–153, 2005.
[8] J. A. Shaw, E. A. Fox, J. A. Shaw, and E. A. Fox. Combination of multiple searches. In *TREC-2*, pages 243–252, 1994.
[9] C. Van Gysel, B. Mitra, M. Venanzi, R. Rosemarin, G. Kukla, P. Grudzien, and N. Cancedda. Reply with: Proactive recommendation of email attachments. In *Proc. of CIKM'17*, pages 327–336, 2017.
[10] T. Vuong, G. Jacucci, and T. Ruotsalo. Watching inside the screen: Digital activity monitoring for task recognition and proactive information retrieval. *Proc. of IMWUT'17*, 1(3):109:1–109:23, 2017.