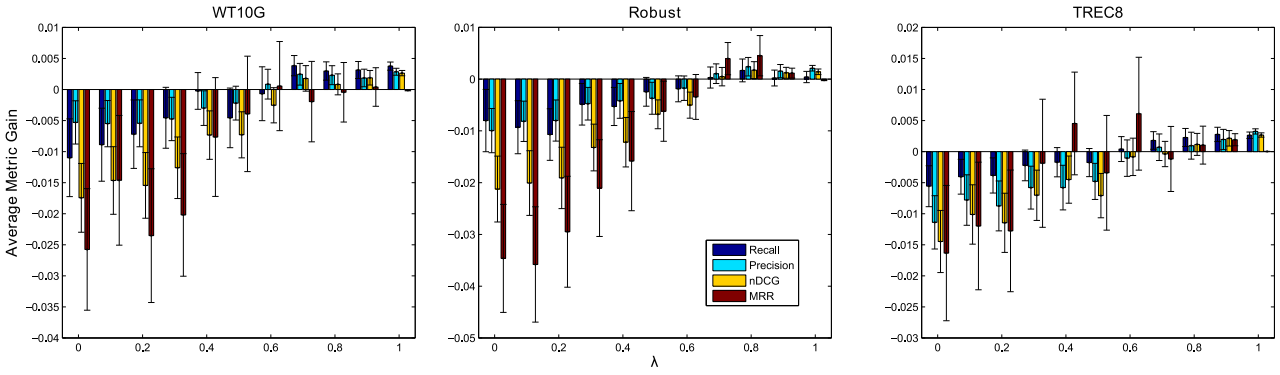

Figure 3: The average metric gain of the IES algorithm over the BM25 baseline when $M = 10$ with 95% confidence intervals. Each bar represents the average gain in a particular metric for a given value of λ, and each chart gives results for a different TREC data set. Positive values indicate gains made over the baseline algorithm, and negative values losses.

			Recall@		Precision@		nDCG@		MRR@	
			10	20	10	20	10	20	10	20
WT10G	#	IES	0.2287	0.4540^{13}	0.3560	0.3940^{135}	0.3823	0.4755^{135}	0.5900	0.5912
	1	BM25	0.2334	0.3651	0.3680	0.3210	0.3935	0.4171	0.6025	0.6053
	2	BM25-U	0.2334	0.4438	0.3680	0.3810	0.3935	0.4715	0.6025	0.6025
	3	MMR	0.1937	0.3651	0.3260	0.3210	0.3424	0.4171	0.5819	0.6053
	4	MMR-U	0.1937	0.4102	0.3260	0.3780	0.3424	0.4407	0.5819	0.5832
	5	Rocchio	0.2334	0.3999	0.3680	0.3320	0.3935	0.4392	0.6025	0.6053
Robust	#	IES	0.2308	0.3898^{13}	0.3851	0.3968^{123}	0.4594	0.5612^{135}	0.6652	0.6670
	1	BM25	0.2276	0.3607	0.3915	0.3330	0.4636	0.4706	0.6317	0.6377
	2	BM25-U	0.2276	0.3694	0.3915	0.3777	0.4636	0.5008	0.6317	0.6335
	3	MMR	0.2179	0.3607	0.3766	0.3330	0.4429	0.4706	0.6246	0.6377
	4	MMR-U	0.2179	0.3785	0.3766	0.3883	0.4429	0.5010	0.6246	0.6264
	5	Rocchio	0.2276	0.3582	0.3915	0.3638	0.4636	0.4725	0.6317	0.6377
TREC8	#	IES	0.1803	0.3768	0.4388	0.4704	0.4644	0.5228	0.6533	0.6583
	1	BM25	0.1851	0.3106	0.4551	0.3980	0.4728	0.4624	0.6404	0.6450
	2	BM25-U	0.1851	0.3644	0.4551	0.4643	0.4728	0.5168	0.6404	0.6455
	3	MMR	0.1812	0.3106	0.4388	0.3980	0.4616	0.4624	0.6472	0.6450
	4	MMR-U	0.1812	0.3640	0.4388	0.4571	0.4616	0.5106	0.6472	0.6523
	5	Rocchio	0.1851	0.3600	0.4551	0.4038	0.4728	0.4892	0.6404	0.6450

Table 1: Table of metrics at $M = 10$ (first page) and $M = 20$ (first and re-ranked second page) for each algorithm and for each data set. A superscript number refers to a metric value significantly above the value of the correspondingly numbered baseline in the table (using the Wilcoxon Signed Rank Test with p = 0.05).

λ increases, performance improves until we start to see gains for the higher values. This highlights that too much exploration can lead to detrimental performance in the first page that isn't recovered by the ranking in the second page, indicating the importance of tuning the parameter correctly. On the other hand, we do see that the optimal setting for λ is typically < 1, indicating that some exploration is beneficial and leads to improved performance across all metrics.

We also observe that the different datasets display different characteristics with regard to the effect that λ has on performance. For instance, for the difficult to rank Robust data set the optimal setting is $\lambda = 0.8$, indicating that exploration is not so beneficial in this case possibly due to the lack of relevant documents in the data set for each topic. Likewise, we find that a setting of $\lambda = 0.9$ is optimal for the TREC8 data set, although there is greater variation possibly owing to the easier to rank data. Finally, the WT10G data set also showed variation, which could be a result of the explicit, graded feedback available during re-ranking. For this data set we chose $\lambda = 0.7$. The differing data sets represent different types of query and search behaviour, and illustrate how λ can be tuned to improve performance over different types of search.

6.3 Comparison with Baselines

After setting the value of λ for each data set according to the optimal values discovered in the previous section (and also setting for the MMR variants $\lambda = 0.8$ for WT10G, $\lambda = 0.8$ for Robust and $\lambda = 0.9$ for TREC8 after some initial investigation), we sought to investigate how the algorithm compared with several baseline algorithms, each of which shares a feature with the IES algorithm; the Rocchio algorithm also performs relevance feedback; the MMR diversifies results; and also some variants that use our conditional

	Recall@		Precision@		nDCG@		MRR@	
	5	10	5	10	5	10	5	10
WT10G	0.2287	0.4540^{134}	0.3560	0.3940^{1345}	0.3823	0.4755^{13}	0.5900	0.5912
Robust	0.2308	0.3898^{13}	0.3851	0.3968^{135}	0.4594	0.5612^{135}	0.6652	0.6670^{1345}
TREC8	0.1803	0.3768^{134}	0.4388	0.4704^{1345}	0.4644	0.5228^{135}	0.6533	0.6583

Table 2: Table of metrics at $M = 5$ (first page) and $M = 10$ for the IES algorithm on each data set. The superscript numbers refer to significantly improved values over the baselines, as indicated in Table 1.

model update. We repeated the experiment as before over two pages with $M = 10$, and measured the recall, precision, nDCG and MRR in order to evaluate the overall search experience of the user who engaged with multiple pages (note that we find different values for MRR@10 and MRR@20 owing to the occasions where a relevant document wasn't located in the first M documents). The results can be seen in Table 1.

We first observe that the scores for the first page ranking are generally lower than that of the baselines, which is to be expected as we sacrifice immediate payoff by choosing to explore and diversify our initial ranking. We still find that the IES algorithm generally outperforms the MMR variants on the first step, particularly for the MRR metric, indicating improved diversification. In the second page ranking we find significant gains over the non-Bayesian update baselines, and some of the baselines using the conditional model update. It is worth noting that the BM25-U variant is simply the case of the IES algorithm with $\lambda = 1$. This demonstrates that our formulation is able to correctly respond to user feedback and generate a superior ranking on the second page that is tuned to the user's information need. We refer back to Figure 3 to show that the second page gains outweigh the first page losses for the values of λ that we have chosen.

Finally, in Table 2 we see a summary of results for the same experiment where we set $M = 5$, so as to demonstrate the IES algorithm's ability to accommodate different page sizes. We observe similar behaviour to before, with the IES algorithm significantly outperforming the baselines across data sets, indicating that even with less scope to optimise the first page, and less feedback to improve the second, the algorithm can perform well.

6.4 Number of Optimised Search Pages

Throughout this paper we have simplified our formulation by setting the number of pages to $T = 2$, so for this experiment, we observe the effect of setting $T = 3$. We performed the experiment as before where we display $M = 5$ documents on each page to the user and receive feedback, which is used to inform the next page's ranking. For the $T = 3$ variant, dynamic programming is used to generate the rankings for both pages 1 and 2 (with λ nominally set to 0.5), and the document's relevancy scores are updated after receiving feedback from pages 1 and 2. We compare against the IES algorithm with $T = 2$, where after page 1 we create a ranking of $2M$ documents, split between pages 2 and 3. Finally, we compare against the baseline ranking of $3M$ documents. The results can be seen in Table 3.

We can see from the results that whilst the $T = 3$ variant still offers improved results over the baseline (except in the case of MRR), the performance is also worse than the $T = 2$ case. This is an example of too much exploration negatively affecting the results; when T is set to 3, the algorithm creates exploratory rankings in both the first and second pages, before it settles on an optimal ranking for the user on the third page. Except for on particular queries, a user engaging in exploratory search should be able to provide sufficient feedback on the first page in order to optimise the remaining pages, and so setting $T = 2$ is ideal in this situation.

Algorithm	Recall@15	Prec@15	nDCG@15	MRR
$T = 3$	0.3249	0.3605	0.3638	0.4194
$T = 2$	0.3584	0.3905	0.4469	0.6371
Baseline	0.2955	0.3293	0.3931	0.5961

Table 3: Table showing metric scores at rank 15 for the baseline and two variants of the IES algorithm, where T is set to 3 and 2.

7. DISCUSSION AND CONCLUSION

In this paper we have considered how to optimally solve the problem of utilising relevance feedback to improve a search ranking over two or more result pages. Our solution differs from other research in that we consider document similarity and document dependence when optimally choosing rankings. By doing this, we are able to choose more effective, exploratory rankings in the first results page that explore dissimilar documents, maximizing what can be learned from relevance feedback. This feedback can then be exploited to provide a much improved secondary ranking, in a way that is unobtrusive and intuitive to the user. We have demonstrated how this works in some toy examples, and formulated a tractable approximation that is able to practically rank documents. Using appropriate text collections, we have shown demonstrable improvements over a number of similar baselines. In addition, we have shown that exploration of documents does occur during the first results page and that this leads to improved performance in the second results page over a number of IR metrics.

Some issues that arise in the use of this algorithm include trying to determine the optimal setting for λ, which may be non-trivial, although it could be learned from search click log data by classifying the query intent [3] and associating intent with values for λ. Also, in our experiments we used BM25 scores as our underlying ranking mechanism and cosine similarity to generate the covariance matrix, it is for future work to investigate combining IES with different ranking algorithms and similarity metrics. We would also like to further investigate the effect that M plays on the optimal ranking and setting for λ; larger values of M should discourage exploration as first page rankings contain more relevant documents. Also, whilst our experiments were able to handle the binary case, our multivariate distribution assumption is a better model for graded relevance, and our

approximation methods and the sequential ranking decision can generate non optimal rankings that may cause detrimental performance.

In the future, we intend to continue developing this technique, including attempting to implement it into a live system or in the form of a browser extension. We also note that multi-page interactive retrieval is a sub-problem in the larger issue of exploratory search, in particular, designing unobtrusive user interfaces to search systems that can adapt to user's information needs. Furthermore, as previously stated we can consider our formulation in the context of POMDP and MAB research [9, 20], which may lead to a general solution that is applicable to other applications, or instead may grant access to different solution and approximation techniques.

8. REFERENCES

[1] Agichtein, E., Brill, E., and Dumais, S. Improving web search ranking by incorporating user behavior information. SIGIR '06, ACM, pp. 19–26.

[2] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. Diversifying search results. WSDM '09, ACM, pp. 5–14.

[3] Ashkan, A., Clarke, C. L., Agichtein, E., and Guo, Q. Classifying and characterizing query intent. ECIR '09, Springer-Verlag, pp. 578–586.

[4] Bellman, R. E. *Dynamic Programming*. Dover Publications, Incorporated, 2003.

[5] Billerbeck, B., and Zobel, J. Questioning query expansion: an examination of behaviour and parameters. ADC '04, Australian Computer Society, Inc., pp. 69–76.

[6] Brandt, C., Joachims, T., Yue, Y., and Bank, J. Dynamic ranked retrieval. WSDM '11, ACM, pp. 247–256.

[7] Cao, G., Nie, J.-Y., Gao, J., and Robertson, S. Selecting good expansion terms for pseudo-relevance feedback. SIGIR '08, pp. 243–250.

[8] Carbonell, J., and Goldstein, J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *In Research and Development in Information Retrieval* (1998), pp. 335–336.

[9] Cassandra, A. A survey of POMDP applications. In *Working Notes of AAAI 1998 Fall Symposium on Planning with Partially Observable Markov Decision Processes* (1998), pp. 17–24.

[10] Chen, H., and Karger, D. R. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR* (2006), pp. 429–436.

[11] Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. Novelty and diversity in information retrieval evaluation. SIGIR '08, ACM, pp. 659–666.

[12] Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. An experimental comparison of click position-bias models. WSDM '08, ACM, pp. 87–94.

[13] Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. Evaluating implicit measures to improve web search. *ACM Trans. Inf. Syst. 23* (April 2005), 147–168.

[14] Hemmje, M. A 3d based user interface for information retrieval systems. In *Proceedings of the IEEE Visualization '93 Workshop on Database Issues for Data Visualization* (1993), Springer-Verlag, pp. 194–209.

[15] Hofmann, K., Whiteson, S., and de Rijke, M. Balancing exploration and exploitation in learning to rank online. In *Proceedings of the 33rd European conference on Advances in information retrieval* (2011), ECIR'11, pp. 251–263.

[16] Joachims, T. Optimizing search engines using clickthrough data. KDD '02, ACM, pp. 133–142.

[17] Koenemann, J., and Belkin, N. J. A case for interaction: a study of interactive information retrieval behavior and effectiveness. CHI '96, ACM, pp. 205–212.

[18] Kules, B., and Capra, R. Visualizing stages during an exploratory search session. HCIR '11.

[19] Morita, M., and Shinoda, Y. Information filtering based on user behavior analysis and best match text retrieval. SIGIR '94, Springer-Verlag New York, Inc., pp. 272–281.

[20] Pandey, S., Chakrabarti, D., and Agarwal, D. Multi-armed bandit problems with dependent arms. ICML '07, ACM, pp. 721–728.

[21] Radlinski, F., Kleinberg, R., and Joachims, T. Learning diverse rankings with multi-armed bandits. ICML '08, ACM, pp. 784–791.

[22] Radlinski, F., Szummer, M., and Craswell, N. Inferring query intent from reformulations and clicks. WWW '10, pp. 1171–1172.

[23] Raman, K., Joachims, T., and Shivaswamy, P. Structured learning of two-level dynamic rankings. CIKM '11, ACM, pp. 291–296.

[24] Robertson, S., and Hull, D. A. The trec-9 filtering track final report. In *TREC* (2001), pp. 25–40.

[25] Robertson, S. E. The Probability Ranking Principle in IR. *Journal of Documentation 33*, 4 (1977), 294–304.

[26] Rocchio, J. *Relevance Feedback in Information Retrieval*. 1971, pp. 313–323.

[27] Rui, Y., Huang, T. S., Ortega, M., and Mehrotra, S. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. Circuits Syst. Video Techn. 8*, 5 (1998), 644–655.

[28] Ruthven, I. Re-examining the potential effectiveness of interactive query expansion. SIGIR '03, ACM, pp. 213–220.

[29] Shen, X., Tan, B., and Zhai, C. Context-sensitive information retrieval using implicit feedback. SIGIR '05, ACM, pp. 43–50.

[30] Shen, X., Tan, B., and Zhai, C. Implicit user modeling for personalized search. CIKM '05, ACM, pp. 824–831.

[31] Shen, X., and Zhai, C. Active feedback in ad hoc information retrieval. SIGIR '05, ACM, pp. 59–66.

[32] Sloan, M., and Wang, J. Iterative expectation for multi period information retrieval. In *Workshop on Web Search Click Data* (2013), WSCD'13.

[33] Spink, A., Jansen, B. J., and Ozmultu, C. H. Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy 10*, 4 (2000), 317–328.

[34] Stephen, R., and Hugo, Z. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval 3*, 4 (2009).

[35] Sugiyama, K., Hatano, K., and Yoshikawa, M. Adaptive web search based on user profile constructed without any effort from users. WWW '04, ACM, pp. 675–684.

[36] Voorhees, E. M. The cluster hypothesis revisited. SIGIR '85, pp. 188–196.

[37] Wang, J., and Zhu, J. Portfolio theory of information retrieval. SIGIR' 09, ACM, pp. 115–122.

[38] Wang, J., and Zhu, J. On statistical analysis and optimization of information retrieval effectiveness metrics. SIGIR '10, pp. 226–233.

[39] White, R. W., and Ruthven, I. A study of interface support mechanisms for interactive information retrieval. *J. Am. Soc. Inf. Sci. Technol. 57*, 7 (May 2006), 933–948.

[40] White, R. W., Ruthven, I., and Jose, J. M. A study of factors affecting the utility of implicit relevance feedback. ACM Press, pp. 35–42.

[41] Xu, Z., and Akella, R. Active relevance feedback for difficult queries. CIKM '08, ACM, pp. 459–468.

[42] Yan, R., Hauptmann, A. G., and Jin, R. Negative pseudo-relevance feedback in content-based video retrieval. MULTIMEDIA '03, ACM, pp. 343–346.

[43] Zamir, O., and Etzioni, O. Grouper: a dynamic clustering interface to web search results. WWW '99, Elsevier North-Holland, Inc., pp. 1361–1374.

[44] Zhang, L., and Zhang, Y. Interactive retrieval based on faceted feedback. SIGIR '10, ACM, pp. 363–370.