ESF ELIAS Long Term Exchange Report

Colin Wilkie

March - May 2015

Abstract

Over the month of April 2015, the author participated in an ESF ELIAS funded exchange from the University of Glasgow to the Vienna University of Technology (Technische Universität Wien). The exhcange was arranged with Dr. Andreas Rauber who is in works within the Department of Software Technology and Interactive Systems. The exchange lasted 4 weeks, however, the author spent the week prior to the exchange attending ECIR which was also taking place in TU Wien.

1 Purpose of the Visit

The purpose of this exchange grant was to allow the candidate to meet with and participate in projects with researchers working in the field of retrievability (and related topics) outside of the University of Glasgow. The Vienna University of Technology was chosen as the destination for this visit due to the presence of Dr. Andreas Rauber, an Associate Professor at the Department of Software Technology and Interactive Systems (IFS). Dr. Rauber had previously supervised Shariq Bashir, a researcher who was very active in research utilising retrievability, and subsequently, was heavily involved in several publications on the effects and uses of retrievability (particularly in prior art/ patent retrieval) [4, 3, 5, 7, 6]. Dr. Rauber currently supervises several students in the IFS group working on a range of topics which include Aziz Taha, who works on hubness in retrieval, Serwah Sabetghadam who investigates reachability, and many other who work on various topics in which retrievability could be applied.

Retrievability is a document centric evaluation measure that was introduced by Azzopardi and Vinay to provide an alternative perspective on how to evaluate systems [1]. The measure can estimate how *likey* a document is to be retrieved from the collection given a retrieval system. The results can then also be used to estimate the retrievability bias of a system by aggregating the individual document retrievability score with a measure like the Gini Coefficient [8] to gain an overview of the bias of the system. However, calculating the retrievability of documents has traditionally required very large numbers of queries to be issued to a system to get an estimate of document retrievability. Doing so on modern, large collections is proving extremely time consuming to the point that these evaluations cannot possibly be performed on these large collections

without substantial computing resources. As these resources are not available to everybody, an important line of retrievability research centres around the efficient estimation of retrievability. To date, only 2 papers have investigated this efficient estimation and they have approached it in two different ways. Wilkie and Azzopardi took a very simple approach to improving efficiency by performing a study that investigated how large a query set had to be to accurately estimate both the retrievability of each document in the collection as well as the overall retrievability bias of the system (computed by estimating the Gini Coefficient over the retrievability scores of the documents in the collection) [12]. In this work, it was found that when using very biased systems (e.g. TF-IDF), only 40% of the queries extracted were required to get an accurate estimation. However, when the system employed was not very biased, removing even 10% of queries resulted in significant differences in both retrievability bias and in individual document retrievability. Work by Bashir also investigated more efficient ways to estimate retrievability by analysing document features [2]. This work found that the retrievability rank of a document could be estimated reasonably well but the retrievability bias could not be estimated accurately.

Given that Dr. Rauber has extensive knowledge of retrievability and that he has students working on related concepts, the main purpose of this visit was to begin work on an investigation into how to more efficiently calculate the retrievability scores of documents in a collection, a task the previous work has not been able to do. To better understand how to evaluate retrievability, we first examine how similar measures (namely hubness and reachability) are estimated. This way, we can use facets of the estimation process for other measures to begin to estimate retrievability.

2 Work Performed During Visit and Initial Results

During the exchange to TU Wien, all the work performed centred around retrievability and investigated topics such as query length, query expansion and estimation. Due to the nature of the experiments required in this line of research, it was known that results would not be ready until after the exchange was completed. As such, the author continued work on papers that results were available for to submit to top tier IR conferences upon their return. Two short papers were written from experiment results completed either before travelling to Vienna or while based there.

2.1 Retrievability and Query Length

To compute the retrievability of documents in collections, an important aspect of the process is the generation of a large query set to get a large, unbiased sample of queries for the collection. These queries are issued in turn and the results used to compute the document retrievability. Experiments were run to determine what impact the length of the queries in the set has on the retrievability estimate

as in previous work, different query generation methods leading to different lengths of queries have been used but nobody had compared the effects of these changes [11, 7]. The intuition being that increasing the length of queries may result in less bias due to the fact more terms would allow more documents to be retrieved. We automatically generated a query from the title of every document in the collection meaning, ideally, every document would have a unique query. However, in reality some titles were very short and so we padded these short titles with terms from the document to expand it to 5 words. Also, several documents share the same, or very similar, titles meaning duplicate queries were extracted. Finally, some documents simply did not have titles so in reality we generated queries for roughly 75% of the collection. We also computed performance based on a success at measure, as we knew what document was used to compute what query, we could generate a Qrel file with the c̈orrectd̈ocument for each query.The main findings in the experiments performed were as follows: Using longer queries does result in an overall decrease in retrievability bias however, this is at a rate of diminishing returns. For example, moving from 1 to 2 term queries has a large impact on reducing bias, while increasing from 4 to 5 terms has very little noticeable effect. When associating this with performance we found that increasing performance by increasing the number of query terms has a trade-off between how easy it is to generate a new meaningful term against how much of a performance gain will be observed. This paper is currently under review for a top tier IR conference.

2.2 Retrievability and Query Expansion

Work previously completed by Bashir and Rauber investigated how to utilise the theory of retrievability to improve pseudo-relevance feedback (PRF) using clustering [5]. In this work, Bashir and Rauber used a clustering approach based on the retrievability of documents to create a new method of PRF for query expansion (QE). In doing so, the authors improved the performance of the QE methods used. Similarly, Pickens *et al* created a new retrieval mechanism for QE called the reverted index [9]. The reverted index was created by computing the retrievability of documents using a large set of single term basis queries and calculating which terms in a document made the document retrievable. The results were then stored like an inverted index where each term was associated with the documents that were retrievable when issuing that term. This was then used for QE where the reverted index would be queried to compute which terms should be used for QE by extracting terms that make documents retrievable. Both of these studies analyse QE in terms of their performance in traditional TREC metrics, however, neither of them evaluate the impact that QE has on retrievability bias and document retrievability. Therefore, experiments were created to analyse how standard QE methods impacted retrievability bias and how this relates to performance, similar to studies by Wilkie and Azzopardi relating retrievability and performance [11, 13]. These experiments compared a number of retrieval models (BM25,TF.IDF and DPH) using two QE methods (Bo1 and KL) on both their performance and their retrievability bias when parameters

associated with the QE methods were altered (namely number of documents to extract terms from, number of terms extracted and Rocchio's beta). Preliminary findings indicated that the number of terms used has the biggest impact on the performance and retrievability bias. Results provide evidence that increasing either, the number of terms or the number of documents improves performance but also increases bias, meaning a more biased search leads to better results. Altering Rocchio's beta, we found that higher values improve performance but again at the expense of bias. This finding is somewhat intuitive as the terms extracted should be focussing the search on progressively smaller sets of documents if the terms are accurate. Further experiments are being run to conclude how the parameters interact with one another when we alter 2 or more at a time, for instance increasing the number of terms extracted and increasing the number of documents to extract the terms from. This paper is currently under review for a top tier IR conference.

3 Further Collaboration & Publication

Collaborations with multiple students at the Vienna University of Technology is currently ongoing and will hopefully lead to several publications at top tier conferences in IR. These collaborations were started during the exchange but due to the nature of the experiments needed to obtain results, results will not be ready for publication until later in the year.

3.1 Reachability

Work with Sabetghadam involves examining the reachability of nodes in a graph, a follow up on her 2015 ECIR paper [10]. Reachability within a graph of documents denotes how easily a document can be reached given an algorithm which steps through the graph. Similar to how retrievability denotes how easily the document can be retrieved, reachbability approaches the same problem in a different context. The collection used in Sabetghadams' experiments is a graph of wikipedia documents and the corresponding images. As some images appear in multiple pages, the documents which contain the same images are intrinsically linked in the graph. Work by Sabetghadam investigated how recall changes as reachability is improved. This has been done by increasing the number of edges within the graph by using semantic links. We have now proposed an improved retrieval model which not only takes relevance into account but also includes the retrievability of documents as well. We hypothesise that by promoting documents with low retrievability, recall can be improved similar to how it has been done by Sabetghadam.

Further work involving Sabetghadam involves further investigating the cause of the results obtained her previous work. Mainly, we wish to evaluate why, when performing a reachability analysis, a subset of the documents within the graph always appears regardless of which point the analysis begins from. We seek to uncover whether these documents are highly linked (i.e. they share the

same pictures thus making them more reachable) and if these documents are also highly retrievable. In the case that the documents that are highly reachable are also highly retrievable, alternate means of retrieval could be employed to improve recall by promoting documents which are not highly retrievable.

3.2 Hubness

Work with Aziz Taha investigating the correlation between hubness and retrievability is also currently ongoing. Hubness and retrievability are intrinsically related as they both seek to analyse the same problem. However, hubness has generally been computed in tasks like musical similarity, until very recently, where there is a large number of features and thus a very high dimensionality to work within. Hubness describes a document in a highly dimensional space that is warped slightly towards the centre, meaning it is much more likely to be deemed similar to the query document due to its position in space and not necessarily because of it similarity. Hubness can then be corrected by removing these hub documents, allowing for more similar documents in the collection to have a better chance of being retrieved. Applying hubness to text retrieval should correlate highly with retrievability as documents that are very retrievable are expected to be highly connected in a graph when all the documents are nodes and the edges linking them are terms, thus creating a very high dimensional space. Experiments with Taha are being conducted to understand this link and if we could use aspects of hubness to gain an accurate estimate of retrievability. The relationship could also be expanded in further experiments where we will prune documents in the collection, similar to how Azzopardi and Vinay have previously [1]. However, unlike how Azzopardi and Vinay removed the least retrievable documents in the collection, we intend to remove the most retrievable documents in the collection (i.e. the hubs). In doing so, we create the opportunity to retrieve less retrievable documents which may increase performance by removing very general documents which match large numbers of queries but are not relevant to the information need.

Clearly, hubness and reachability are related also and so it will be very interesting to link the three measures together and find what makes them different from one another.

4 Conclusions

To conclude, the exchange visit has been very profitable to both the author of this report and several members of the community at TU Wien. This exchange has not only brought about several exciting new strands of research but has also strengthened connections between the University of Glasgow and TU Wien. Plans are underway to bring some of the students from TU Wien to the University of Glasgow to participate in other research projects and to do talks to the research group in the university.

Further work is required before the research being performed will be publishable but this work should be completed within a reasonable time frame and results published to top IR venues around the world. The connections forged will last beyond these initial projects and will hopefully lead to further collaborations between Glasgow and Vienna.

References

[1] L. Azzopardi and V. Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *Proc. of the 17th ACM CIKM*, pages 561–570, 2008.

[2] S. Bashir. Estimating retrievability ranks of documents using document features. *Neurocomput.*, 123:216–232, Jan. 2014.

[3] S. Bashir and A. Rauber. Analyzing document retrievability in patent retrieval settings. In *Database and Expert Systems Applications*, pages 753–760. 2009.

[4] S. Bashir and A. Rauber. Identification of low/high retrievable patents using content-based features. In *Proceedings of the 2Nd International Workshop on Patent Information Retrieval*, PaIR '09, pages 9–16, New York, NY, USA, 2009. ACM.

[5] S. Bashir and A. Rauber. Improving retrievability of patents with cluster-based pseudo-relevance feedback documents selection. In *Proc. of the 18th ACM CIKM*, pages 1863–1866, 2009.

[6] S. Bashir and A. Rauber. Improving retrievability & recall by automatic corpus partitioning. In *Trans. on large-scale data & knowledge-centered sys. II*, pages 122–140. 2010.

[7] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *Proc. of the 32nd ECIR*, pages 457–470, 2010.

[8] J. Gastwirth. The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, 54:306–316, 1972.

[9] J. Pickens, M. Cooper, and G. Golovchinsky. Reverted indexing for feedback and expansion. In *Proc. of the 19th ACM CIKM*, pages 1049–1058, 2010.

[10] S. Sabetghadam, M. Lupu, R. Bierig, and A. Rauber. Reachability analysis of graph modelled collections. In A. Hanbury, G. Kazai, A. Rauber, and N. Fuhr, editors, *Advances in Information Retrieval*, volume 9022 of *Lecture Notes in Computer Science*, pages 370–381. Springer International Publishing, 2015.

[11] C. Wilkie and L. Azzopardi. Best and fairest: An empirical analysis of retrieval system bias. *Advances in Information Retrieval*, 2014.

[12] C. Wilkie and L. Azzopardi. Efficiently estimating retrievability bias. In *ECIR'14*, pages 720–726, 2014.

[13] C. Wilkie and L. Azzopardi. A retrievability analysis: Exploring the relationship between retrieval bias and retrieval performance. In *Proc. of the 23rd ACM CIKM*, pages 81–90, 2014.