Top-k Retrieval using Facility Location Analysis

Guido Zuccon[†1], Leif Azzopardi[1], Dell Zhang[2], and Jun Wang[3]
{guido, leif}@dcs.gla.ac.uk, dell.z@ieee.org, j.wang@cs.ucl.ac.uk

[1]School of Computing Science, University of Glasgow, UK
[2]DCSIS, Birkbeck, University of London, UK
[3]Department of Computing Science, University College London, UK

Abstract. The top-k retrieval problem aims to find the optimal set of k documents from a number of relevant documents given the user's query. The key issue is to balance the relevance and diversity of the top-k search results. In this paper, we address this problem using *Facility Location Analysis* taken from Operations Research, where the locations of facilities are optimally chosen according to some criteria. We show how this analysis technique is a generalization of state-of-the-art retrieval models for diversification (such as the Modern Portfolio Theory for Information Retrieval), which treat the top-k search results like "*obnoxious facilities*" that should be dispersed as far as possible from each other. However, Facility Location Analysis suggests that the top-k search results could be treated like "*desirable facilities*" to be placed as close as possible to their customers. This leads to a new top-k retrieval model where the best representatives of the relevant documents are selected. In a series of experiments conducted on two TREC diversity collections, we show that significant improvements can be made over the current state-of-the-art through this alternative treatment of the top-k retrieval problem.

1 Introduction

Information Retrieval (IR) is concerned with finding relevant documents that satisfy user information needs [12]. In many application domains, such as web search, the number of (potentially) relevant documents for a query is often quite large, while users are only interested in a few of the most important (top-k) relevant documents. In [5], Chen and Karger argued that returning more relevant documents is not always optimal. Instead, retrieving a diverse subset of the most representative relevant documents is more likely to ensure that all possible information needs of a given query are satisfied. Typically, the top-k retrieval problem is solved by using a standard retrieval model [12], which computes the relevance score of each document *individually* and then returns the k documents with the highest relevance scores. However, as the inter-relationships among the relevant documents are ignored, the top-k search results are often quite alike

[†] Now at the Australian e-Health Research Centre, CSIRO, Brisbane, Australia. Email: guido.zuccon@csiro.au

(and potentially duplicates or near duplicates). As pointed out in [5], this can be undesirable, particularly when different users are interested in different meanings or aspects of the query. In order to resolve the query ambiguity and avoid the information redundancy, it is necessary to optimize the top-k search results *collectively*. Consequently, a number of search result diversification methods have been developed [1, 17, 15, 18, 3, 14]. Here, we considered only general methods for diversifying rankings, that can be applied regardless of document collections and information used to diversify documents. We thus do not consider the approaches presented in [3, 14], because [3] can only be applied to the diversification of search results for Web search, as it uses the web-graph to diversify rankings, and because [14] uses (among others) query log and taxonomy-based features.

In this paper, we formalize the top-k retrieval problem within the unified framework of *Facility Location Analysis* (FLA) taken from Operations Research [8], as a way to account and optimize for novelty and diversity [6, 17]. We then show that the state-of-the-art techniques for search result diversification, such as Maximal Marginal Relevance [1], the Quantum Probability Ranking Principle [18], and the Modern Portfolio Theory [15], can be modelled as one type of FLA. That is, they treat the top-k search results as *obnoxious facilities* to be dispersed as far as possible from each other. On the other hand, we consider the top-k search results as *desirable facilities* to be placed as close as possible to their customers such that the top-k search results constitute the best representatives of the relevant documents. Our experiments demonstrate that this novel method outperforms current state-of-the-art methods.

2 Approaches to Top-K Retrieval

Suppose that for a given query q, the set of relevant[1] documents found by the retrieval system is $D = \{d_1, d_2, \ldots, d_n\}$. The relevance score of a document with respect to q is calculated by a function $r : D \to \mathbb{R}$. Without loss of generality, we assume that $r(d_1) \geq r(d_2) \geq \ldots \geq r(d_n)$. Furthermore, the distance[2] or dissimilarity between any two documents in D is calculated by a function $w : D \times D \to \mathbb{R}$.

The task of top-k retrieval is to pick a subset $S \subset D$ of k documents that is both relevant and diverse simultaneously. To find the optimal subset, we formulate it as a *facility location problem* [8] in Operations Research — given a set of customer "locations" D, we would like to find a subset[3] $S \subset D$ to open k "facilities" there so as to optimize a graph-theoretic objective that is dependent on the cost of opening a facility at each location and also the distance between each pair of locations. The facility opening cost at location d_i is set to $-r(d_i)$

[1] It is usually not possible to know whether a document is indeed relevant to the given query, so D is actually approximated by the n documents with the highest relevance scores.

[2] The distance function $w(\cdot, \cdot)$ is not required to be a metric here.

[3] In general facility location problems, the set of potential facility locations can be different from the set of customer locations.

which reflects our preference for high relevance. Making use of different optimization objectives, Facility Location Analysis would lead to different retrieval techniques for search result diversification.

2.1 Obnoxious Facility Dispersion

One way to diversify search results, and perhaps the most intuitive, is to select dissimilar documents. In this case, the top-k search results are essentially treated as *obnoxious* facilities, such as nuclear-power plants, oil-storage tanks, and ammunition dumps. Such facilities should be dispersed as far as possible from each other so that in the case of an accident or incident at one of the facilities damage can be minimized or contained. The optimization objective of obnoxious k-facility dispersion is two-fold: (i) to minimize the total cost of opening those facilities and (ii) to maximize the spread of those facilities.

Since the facility dispersion problem is in general NP-hard even if the facility opening costs are ignored [8], it can only be solved efficiently by approximate optimization algorithms. This issue is also present when diversifying document rankings, as Carterette noted [2]. A widely used algorithm that provides an approximate solution to the facility dispersion problem is Greedy Best-First Search [13]. This algorithm is shown in Algorithm 1. For the top-k retrieval problem, the algorithm first initializes S with the single most relevant document d_1, and then sequentially adds the remaining documents in $D \setminus S$ to S one by one, until the size of S has reached k. The crucial step in the algorithm is line 3 that selects the next document according to a heuristic function:

$$h(d, S) = \lambda r(d) + (1 - \lambda) g(d, S) \ , \tag{1}$$

where the function $g : D \times 2^D \to \mathbb{R}$ estimates the effect of a document d on the overall dispersion of S, and $\lambda \in [0, 1]$ is a trade-off parameter. Finally the top-k search results in S are returned to the users in the order they have been added to S.

Algorithm 1: Greedy Best-First Search for Obnoxious Facility Dispersion.

```
  Input  : $D$, $k$, $h$
  Output: $S$
1 $S \Leftarrow \{d_1\}$ ;
2 for $i = 2, \ldots, k$ do
3     $d^* = \arg\max_{d \in D \setminus S} \ h(d, S)$ ;
4     $S \Leftarrow S \cup \{d^*\}$ ;
5 end
```

It is of interest to note that Gollapudi and Sharma proposed a set of axioms for search result diversification [7], which suggested that algorithms for obnoxious facility dispersion could be used to select the top-k search results.

Here, we follow up on their intuition and show how obnoxious facility dispersion can be connected to Maximal Marginal Relevance [1], the Quantum Probability Ranking Principle [18], and the Modern Portfolio Theory [15].

Maximal Marginal Relevance

Consider the scenario where the distance function is set such that $w(d, d') = -s(d, d')$, where s is some measure of document similarity [12]. If the dispersion function is set that:

$$g(d, S) = \min_{d' \in S} w(d, d')$$

then the heuristic function of eq. (1) becomes:

$$\begin{aligned} h(d, S) &= \lambda r(d) + (1 - \lambda) \min_{d' \in S} (-s(d, d')) \\ &= \lambda r(d) - (1 - \lambda) \max_{d' \in S} s(d, d') \ . \end{aligned} \tag{2}$$

This is exactly the ranking formula of Maximal Marginal Relevance (MMR) [1].

Modern Portfolio Theory Consider the scenario where the relevance function is:

$$r(d) = \omega_d \Pr(d)$$

where $\Pr(d)$ is the probability of d being relevant to the query, and ω_d is the importance weight of d's rank position in S measured by the corresponding discounting factors of nDCG [10]. The distance function can be set as the negative of the Pearson's correlation coefficient between the probability distributions of d and d' weighted by the importance of their rank positions, i.e.:

$$w(d, d') = -\omega_d \omega_{d'} \rho_{d,d'}$$

By setting the dispersion function as $g(d, S) = \sum_{d' \in S} w(d, d')$ and the trade-off parameter as $\lambda = 1/(1 + 2b\sigma_d^2)$, where b is a parameter encoding the user's risk propensity and σ_d^2 is the variance of probability distribution of d, then the heuristic function of eq. (1) can be rewritten as:

$$\begin{aligned} h(d, S) &= \lambda \omega_d \Pr(d) - \lambda (2b\sigma_d^2) \sum_{d' \in S} \omega_d \omega_{d'} \rho_{d,d'} \\ &\propto \Pr(d) - \sum_{d' \in S} 2b\sigma_d^2 \omega_{d'} \rho_{d,d'} \ . \end{aligned} \tag{3}$$

This is exactly the ranking formula of the Modern Portfolio Theory (MPT) [15].

Quantum Probability Ranking Principle

Consider the scenario where the relevance function equivalent to the probability of d being relevant to the query, i.e.:

$$r(d) = \Pr(d)$$

In order to obtain the QPRP in the framework of facility location analysis the distance function $w(d, d')$ has to be set equivalent to the quantum interference term of QPRP, i.e.:

$$w(d, d') = I(d, d')$$

where the interference $I(d, d')$ between d and d' is calculated as $-\sqrt{\Pr(d)}\sqrt{\Pr(d')}$ $\cos\theta_{d,d'}$, with $\theta_{d,d'}$ being the *phase difference* between the complex probability amplitudes associated to d and d'. Similarly to the case of PT, we set $g(d, S) = \sum_{d' \in S} w(d, d')$. Then, if $\lambda = \frac{1}{2}$, the heuristic function of eq. (1) becomes:

$$\begin{aligned} h(d, S) &= \frac{1}{2}\Pr(d) - \frac{1}{2}\sum_{d' \in S} I(d, d') \\ &\propto \Pr(d) - \sum_{d' \in S} \sqrt{\Pr(d)}\sqrt{\Pr(d')}\cos\theta_{d,d'} \ . \end{aligned} \quad (4)$$

And this equates to the Quantum Probability Ranking Principle (QPRP) ranking formula [18].

2.2 Desirable Facility Placement

In contrast to the above techniques, which subscribe to the obnoxious facility dispersion approach under FLA, for search results diversification, desirable facility placement from FLA may be more appropriate. In this setting, *desirable* facilities, such as warehouses, hospitals, and fire stations, are to be placed as close as possible to their customers. Since a customer would just go to the closest facility, there is a competitive relationship among those k-facilities. This is the underlying driving force behind diversification, i.e. it is better for each of the k-facilities to be the centers of different areas such that every customer is close to one facility. The optimization objective of desirable k-facility placement is two-fold:

1. to minimize the total cost of opening those facilities, and,
2. to minimize the distances from the customer locations to their closest facilities.

In the context of top-k retrieval, by treating search results like desirable facilities, we are in fact selecting the best representatives of the relevant documents so that the top-k search results constitute a concise summary of all the relevant information, and as such novelty and diversity naturally arise. More formally, the optimal set of top-k search results, considered as desirable facilities, is given by:

$$S^* = \arg\min_{\substack{S \subset D \\ |S| = k}} f(S) \quad (5)$$

that optimizes the objective function:

$$f(S) = -\lambda \sum_{d \in S} r(d) + (1-\lambda) \sum_{d \in D \setminus S} \left(\min_{d' \in S} w(d, d') \right) , \tag{6}$$

where $\lambda \in [0,1]$ is a trade-off parameter. This problem is an extension of the *uncapacitated facility location problem* in which the number of facilities is not bounded [8]. This formulation of the top-k retrieval problem encompasses traditional relevance-based retrieval [12] (when $\lambda = 1$) and k-medoids clustering [9] (when $\lambda = 0$).

The above facility location problem is also in general NP-hard, which can be proved by reduction, for example, from the set cover problem. Since we have an explicit objective function here, we choose to optimize it approximately using Greedy Local Search (GLS), a.k.a. Hill Climbing, as shown in Algorithm 2. The algorithm first initializes S with the k most relevant documents $d_1, \ldots, d_k$, and then iteratively refines S by swapping a facility location in S and a customer location in $D \setminus S$, until the process converges. Finally, the top-k search results in S are returned to users in the order of their respective contributions to $f(S)$.

Algorithm 2: Greedy Local Search for Desirable Facility Placement.

```
Input  : $D$, $k$, $f$
Output: $S$
$S \Leftarrow \{d_1, \ldots, d_k\}$ ;
repeat
    for $d \in S$ do
        for $d' \in D \setminus S$ do
            $S' \Leftarrow (S \setminus \{d\}) \cup \{d'\}$ ;
            if $f(S') > f(S)$ then
                $S \Leftarrow S'$ ;
            end
        end
    end
until $S$ does not change;
```

In section 2.1, we noted that MMR, MPT and QPRP can all be framed as solutions based on obnoxious facility dispersion (OBN) where they all employ a Best First Search (BFS) heuristic (i.e. OBN + BFS). However, for the desirable facilities placement (DES), we arrived at a solution which employs a Greedy Local Search (GLS) heuristic (i.e. DES + GLS). The focus of our empirical evaluation will be on comparing the different methods MMR, MPT and QPRP within these two solutions. To do this we can construct desirable facilities placement variants of MMR, MPT and QPRP by using the follow settings in Equation 6:

- For MMR, we set $w(d, d') = s(d, d')$.

- For MPT, we set $r(d) = \Pr(d)$ and we set the dispersion function $w(d, d') = 2b\sigma_d^2 \omega_{d'} \rho_{d,d'}$.
- For QPRP we set $r(d) = \Pr(d)$ and we set $w(d, d') = \sqrt{\Pr(d)}\sqrt{\Pr(d')} \cos \theta_{d,d'}$.

For convenience, we distinguish between the two different solutions as MMR-BFS, MPT-BFS, and QPRP-BFS for obnoxious facility dispersion (OBN) versus MMR-GLS, MPT-GLS, and QPRP-GLS for desirable facilities placement (DES), and compare these approaches in the next section.

3 Experimental Methodology

To empirically evaluate the above top-k retrieval techniques, we employed two TREC collections designed for testing diversity using three topic sets. The first is the TREC 6-7-8 subtopic retrieval collection[4] [17], consisting of the documents from the Financial Times of London (TREC Disk 4), together with the 20 retrieval topics and associated subtopic relevance assessments used in TREC 6, 7, and 8 interactive tracks. The second is the recent TREC ClueWeb collection[5], together with the two topic sets from TREC 2009 and 2010 Web tracks[6]. In our experiments on ClueWeb, all systems were restricted to retrieve documents from only part B of the ClueWeb dataset (i.e. the first 50 millions documents). Document collections have been indexed using Lemur/Indri[7], where standard stop-word removal and Porter stemming was applied. For each retrieval topic, the title of the topic was used to generate queries.

As a naïve baseline we used the standard Language Modeling (LM) approach to retrieval [16], i.e. a method *without diversification*. The ranking provided by LM was then used by all the diversity based top-k retrieval techniques to ensure that the methods were compared fairly. We computed the relevance score as

$$r(d) = \frac{logPr(q|\hat{M}(d))}{|q|}$$

where $\hat{M}(d_i)$ was the unigram language model estimated from document d_i with Dirichlet smoothing[8] [16], and $|q|$ was the number of terms in the query. The scaling factor $1/|q|$ was introduced to remove the influence of query length and facilitated setting the parameter λ across queries. We then calculated the distance between d and d' using Kullback-Leibler divergence[9]

$$s(d, d') = D_{KL}(\bar{M}(d)||\hat{M}(d'))$$

where $\bar{M}(d)$ is the unigram language model estimated from document d using the maximum likelihood estimation, and $\hat{M}(d')$ is the unigram language model

[4] `http://trec.nist.gov/data/interactive.html`

[5] `http://boston.lti.cs.cmu.edu/Data/clueweb09/`

[6] `http://plg.uwaterloo.ca/~trecweb/`

[7] `http://www.lemurproject.org/lemur.php`

[8] The Dirichlet prior μ was set to 2,000.

[9] Although the Kullback-Leibler divergence is not a metric, it has been shown to work very well for measuring document dissimilarity [11].

estimated from document d' with Dirichlet smoothing. To better compare the ranking approaches independently from the distance function used, we employed the same approach used to compute $s(d, d')$ for replacing the Pearson's correlation in MPT and for approximating the quantum interference term in QPRP[10].

For each query q in the test set Q, we retrieved $n = 100$ documents with the highest relevance scores to form the set D. Then the retrieval techniques — MMR, MPT, QPRP (as OBN+BFS variants), and their DES+GLS variants — were employed to select the top $k = 20$ documents as the set S to be returned to users. Retrieval performances were measured at rank positions 5, 10 and 20 by ERR-IA [4], α-nDCG [6] and subtopic-recall (s-rec) [17], where both ERR-IA and α-nDCG were used following the standard settings employed in the TREC 2011 Web Track (i.e. in ERR-IA all intents were given the same probability, and in α-nDCG, α was set to 0.5).

Where parameters were present (e.g. b and σ^2 for MPT and λ for MMR), these were tuned so as to maximise α-nDCG@10. In particular, for λ, we explored results obtained varying the parameter in the range $[0, 1]$ with decimal steps in MMR. Additionally, for MPT we treated variance as a parameter (similarly to [18]), studying values of σ^2 in the range $[10^{-6}, 10]$ with steps of 10, while we studied values of b in the range [-10, 10] with unitary increments. No parameter estimation is required for the QPRP methods.

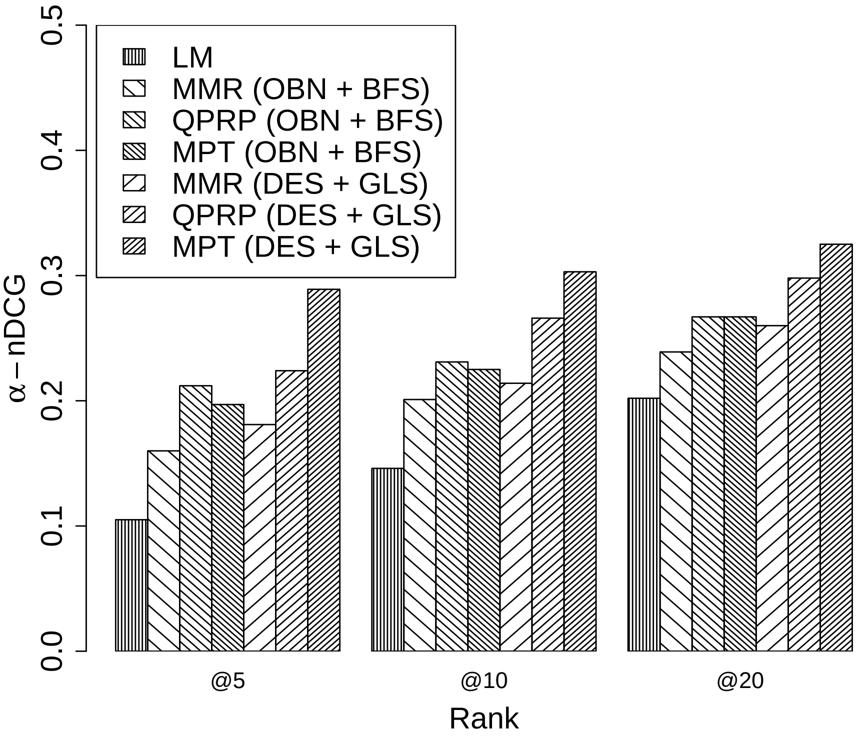
We tested the results for statistical significant differences over LM using a Wilcoxon signed-rank test with $p < 0.01$: in the tables, $*$ indicates that statistical significant differences were found. Note that we did not perform significance tests on the TREC 6-7-8 dataset because the sample size is too small (there are only 20 topics).

4 Results and Discussion

Table 1 reports the results obtained for LM, OBN+BFS methods (MMR, MPT, QPRP) and their corresponding DES+GLS variants on the ClueWeb topic sets, while Table 2 contains the results from TREC 6-7-8 collections. Included in each table is the percentage increase of the LM Baseline. An overview of the results from ClueWeb is shown in Figure 1 which shows a bar plot for the alpha-nDCG results for ranks 5, 10 and 20. As expected the OBN+BFS methods outperform the LM baseline quite considerably across all the measures shown for all of the collection/topic sets. Also, for the DES+GLS methods, substantial improvements are witness over the baseline, and in most cases the DES+GLS heuristic outperforms the OBN+BFS heuristics. Of particular note, is that the DES+GLS variant of MPT clearly and consistently outperforms all other methods. For example, on TREC 2009, MPT-BFS obtains alpha-DCG of 0.181 while MPT-GLS scores 0.282. This is quite a substantial increase in performance. Further significance testing reveals that these increases of the DES+GLS variants were significantly better than the OBN+BFS variants (denoted by †).

[10] Note that a similar approach was taken in [18]: we refer to that work for the details.

Fig. 1. The top-k retrieval performances measured by α-nDCG at various rank positions on the TREC Clueweb collections (2009 and 2010).



In terms of the different methods, we would expect the OBN+BFS and DES+GLS variants of MPT to perform the best as the method encodes two parameters for tuning. We also observe that this is generally the case. However, the QPRP methods (OBN+BFS and DES+GLS) do not require any within method parameterization, making the solution much more robust. With respect to performance we can also see that the QPRP-GLS variant not only significantly outperforms the QPRP-BFS variant, but it also out performs the MPT-BFS variant. That is, despite MPT having two additional parameters to tune, the QPRP-GLS method is comparable, if not better across all the measures (and across collections/topic sets). This means that the QPRP-GLS is a practical and viable solution to undertaken, while the MPT-GLS method results in even further improvements, such improvements are only possible if the parameters of MPT can be reliably estimated.

These findings suggest that the heuristic employed by the desirable facilities placement solution creates a better ranking due to the optimization being performed on the set of documents rather than just on the next document to

Table 1. The top-k retrieval performances measured by ERR-IA, α-nDCG and s-recall on the TREC 2009 and TREC 2010 topic sets. For each measure, the best result is reported in bold. We used the Wilcoxon signed-rank test with $p < 0.01$ for testing statistical significance: $*$ indicates that statistical significant differences against the LM baseline were found, while † indicates that statistical significant differences between OBN + BFS and DES + GLS were found.

TREC 2009										
		ERR-IA			**alpha-DCG**			**s-rec**		
		@5	@10	@20	@5	@10	@20	@5	@10	@20
LM		0.065	0.084	0.097	0.105	0.150	0.207	0.171	0.273	0.435
OBN + BFS	**MMR**	0.108* 66.38%	0.126* 50.94%	0.134* 38.38%	0.164* 56.53%	0.208* 38.33%	0.243* 17.46%	0.244* 42.50%	0.360* 31.71%	0.465 6.97%
	QPRP	0.136* 109.43%	0.149* 78.19%	0.157* 62.81%	0.196* 87.51%	0.222* 48.09%	0.262* 26.49%	0.244* 42.88%	0.343* 25.37%	0.475 9.27%
	MPT	0.127* 95.27%	0.141* 68.19%	0.150* 55.54%	0.181* 73.16%	0.216* 43.55%	0.263* 26.80%	0.238* 39.18%	0.357* 30.73%	0.506* 16.32%
DES + GLS	**MMR**	0.127*† 96.02%	0.141*† 68.38%	0.151*† 56.15%	0.182*† 73.57%	0.215*† 43.55%	0.265*† 27.88%	0.238* 39.18%	0.361* 32.20%	0.506* 16.32%
	QPRP	0.142*† 119.25%	0.160*† 91.26%	0.169*† 75.11%	0.210*† 100.27%	0.255*† 70.03%	0.300*† 44.56%	0.300*† 75.24%	**0.438***† 60.24%	**0.556***† 27.89%
	MPT	**0.203***† 212.59%	**0.213***† 155.19%	**0.218***† 125.42%	**0.282***† 169.37%	**0.303***† 101.51%	**0.323***† 55.78%	**0.346***† 102.34%	0.431*† 57.80%	0.490*† 12.72%

TREC 2010										
		ERR-IA			**alpha-DCG**			**s-rec**		
		@5	@10	@20	@5	@10	@20	@5	@10	@20
LM		0.080	0.095	0.110	0.105	0.142	0.198	0.150	0.266	0.409
OBN + BFS	**MMR**	0.119* 49.84%	0.135* 41.87%	0.145* 32.06%	0.157* 49.00%	0.194* 36.78%	0.236* 19.16%	0.219* 46.00%	0.343* 28.91%	0.467* 14.17%
	QPRP	0.174* 118.68%	0.183* 91.42%	0.189* 72.15%	0.227* 115.85%	0.240* 69.60%	0.271* 37.14%	0.279* 86.00%	0.371* 39.42%	0.478* 16.86%
	MPT	0.178* 123.72%	0.172* 80.37%	0.187* 70.09%	0.213* 102.55%	0.235* 65.77%	0.272* 37.59%	0.249* 66.00%	**0.460*** 72.72%	0.512* 25.08%
DES + GLS	**MMR**	0.139*† 74.70%	0.155*† 62.55%	0.168*† 52.81%	0.181*† 72.12%	0.213*† 50.26%	0.255* 28.99%	0.231* 54.00%	0.363* 36.30%	0.501* 22.39%
	QPRP	0.194*† 143.83%	0.199*† 108.69%	0.201*† 82.83%	0.238*† 126.32%	0.276*† 94.70%	0.297*† 50.23%	0.321*† 114.00%	0.415*† 55.82%	**0.542***† 32.41%
	MPT	**0.235***† 195.10%	**0.242***† 154.29%	**0.248***† 125.75%	**0.296***† 181.52%	**0.304***† 114.36%	**0.327***† 65.42%	**0.373***† 148.44%	0.453* 69.96%	0.529* 29.15%

be ranked as it is the case for the OBN+BFS solution. While, the results show that substantial improvements can be made in terms of retrieval performance, it should be noted that this does come at a small computational cost. However, given that only the top-k results need to be optimized then increase in computational cost is marginal, making this approach practical in an online setting.

Table 2. The top-k retrieval performances measured by ERR-IA, α-nDCG and s-recall on the TREC 6-7-8.

TREC 6-7-8										
		ERR-IA			**alpha-DCG**			**s-rec**		
		@5	@10	@20	@5	@10	@20	@5	@10	@20
LM		0.170	0.179	0.189	0.489	0.485	0.523	0.323	0.404	0.578
OBN + BFS	**MMR**	0.176	0.184	0.194	0.499	0.492	0.528	0.334	0.410	0.578
		3.61%	3.12%	2.79%	1.97%	1.45%	0.99%	3.44%	1.38%	0.00%
	QPRP	0.192	0.200	0.207	0.585	0.558	0.562	0.385	0.468	0.602
		12.97%	11.78%	9.67%	19.60%	15.13%	7.48%	19.29%	15.85%	4.25%
	MPT	0.193	0.204	0.212	0.558	0.562	0.568	0.373	0.488	0.595
		13.50%	14.50%	12.32%	14.03%	15.92%	8.71%	15.64%	20.91%	3.05%
DES + GLS	**MMR**	0.193	0.205	0.212	0.557	0.565	0.569	0.375	**0.493**	**0.609**
		13.36%	14.94%	12.29%	13.78%	16.55%	8.85%	16.32%	21.94%	5.41%
	QPRP	0.197	0.206	0.214	0.571	0.564	0.571	0.393	0.488	0.601
		16.10%	15.63%	13.29%	16.77%	16.34%	9.11%	21.91%	20.91%	4.01%
	MPT	**0.234**	**0.238**	**0.242**	**0.633**	**0.585**	**0.577**	**0.409**	0.465	0.580
		37.45%	33.47%	28.00%	29.33%	20.63%	10.38%	26.68%	15.02%	0.38%

5 Conclusions and Future Work

In this paper, we approached the top-k retrieval problem as if it was analogous to Facility Location Analysis in Operations Research where there are two main approaches depending on the type of facilities to be placed:

- Obnoxious Facility Dispersion – which we noted characterized existing search result diversification techniques such as MMR, MPT and QPPR; and uses a Best First Search heuristic (i.e. OBN+BFS); and,
- Desirable Facility Placement – which resulted in the development of an alternative and novel approach to the top-k problem, that used a Greedy Local Search heuristic and produced DES+GLS variants of MMR, MPT and QPRP.

The findings from our experiments demonstrate that this novel alternative treatment of the top-k based on desirable facility placement results in substantially and significantly better performance over the obnoxious facility dispersion based methods (i.e. DES+GLS over OBN+BFS essentially). These results back up the intuition that relevant items are desirable in nature, and that the most representative set of relevant items should be found. While we have witness significant and substantial increases of performance it would be of further interest to examine different heuristics to either decrease the computational complexity and improving the solutions efficiency by switching from GLS to BFS for the DES solution, or to try to further improve performance by exploring more complex facility location models (such as the capacitated facility location problem [8]) and to investigate more advanced optimization methods (such as simulated annealing and genetic algorithm [13]). Indeed, it would also be interesting to examine other heuristics for the OBN solutions too. Another direction of further research would be to use these solutions with methods that also draw upon external evidence (such as [14] and [3]) to determine whether even greater improvements to performance are possible using these problem formulations.

References

1. J. G. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st ACM SIGIR*, pages 335–336, Melb., Australia, 1998.
2. B. Carterette. An analysis of np-completeness in novelty and diversity ranking. *Information Retrieval*, 14:89–106, February 2011.
3. P. Chandar and B. Carterette. Diversification of search results using webgraphs. In *Proceeding of the 33rd international ACM SIGIR*, pages 869–870, 2010.
4. O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceeding of the 18th ACM CIKM*, pages 621–630, 2009.
5. H. Chen and D. R. Karger. Less is more: Probabilistic models for retrieving fewer relevant documents. In *Proceedings of the 29th ACM SIGIR*, pages 429–436, Seattle, WA, USA, 2006.
6. C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st ACM SIGIR*, pages 659–666, Singapore, 2008.
7. S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th WWW*, pages 381–390, Madrid, Spain, 2009.
8. T. F. Gonzalez, editor. *Handbook of Approximation Algorithms and Metaheuristics*. 2007.
9. J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2005.
10. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.
11. O. Kurland and L. Lee. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of the 28th ACM SIGIR*, pages 306–313, Salvador, Brazil, 2005.
12. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
13. S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2009.
14. R. L. T. Santos, C. Macdonald, and I. Ounis. Intent-aware search result diversification. In *Proceedings of the 34th ACM SIGIR*, pages 595–604, 2011.
15. J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd ACM SIGIR*, pages 115–122, Boston, MA, USA, 2009.
16. C. Zhai. *Statistical Language Models for Information Retrieval*. Morgan and Claypool, 2008.
17. C. Zhai, W. W. Cohen, and J. D. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th ACM SIGIR*, pages 10–17, TO, Canada, 2003.
18. G. Zuccon and L. Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In *Proceedings of the 32nd BCS-IRSG ECIR*, pages 357–369, Milton Keynes, UK, 2010.