An In-depth Study of Implicit Search Result Diversification

Hai-Tao Yu[1], Adam Jatowt[2], Roi Blanco[3], Hideo Joho[1],

Joemon Jose[4], Long Chen[4], and Fajie Yuan[5]

[1]{yuhaitao, hideo}@slis.tsukuba.ac.jp, University of Tsukuba, Japan
[2]adam@dl.kuis.kyoto-u.ac.jp, Kyoto University, Japan
[3]rblanco@udc.es, University of A Coruña, Spain
[4]{Joemon.Jose, Long.Chen}@glasgow.ac.uk, University of Glasgow, UK
[5]f.yuan.1@research.gla.ac.uk, University of Glasgow, UK

Abstract. In this paper, we present a novel Integer Linear Programming formulation (termed *ILP4ID*) for implicit *search result diversification* (SRD). The advantage is that the exact solution can be achieved, which enables us to investigate to what extent using the greedy strategy affects the performance of implicit SRD. Specifically, a series of experiments are conducted to empirically compare the state-of-the-art methods with the proposed approach. The experimental results show that: (1) The factors, such as different initial runs and the number of input documents, greatly affect the performance of diversification models. (2) *ILP4ID* can achieve substantially improved performance over the state-of-the-art methods in terms of standard diversity metrics.

Keywords: Search Result Diversification, Integer Linear Programming

1 Introduction

Accurately and efficiently providing desired information to users is far from being resolved. A key problem is that users often submit short queries that are ambiguous and/or underspecified. As a remedy, one possible solution is to apply *search result diversification* (SRD), which is characterized as finding the optimally ranked list of documents, which maximizes the overall relevance to multiple possible intents, while minimizing the redundancy among the returned documents. Depending on *whether the subtopics underlying a query are known beforehand*, the problem of SRD can be differentiated into *implicit SRD* and *explicit SRD*. In this work, we do not investigate methods nor supervised methods for SRD, but *we focus instead on implicit methods*, where the possible subtopics underlying a query are *unknown*.

Despite the success achieved by the state-of-the-art methods, the key underlying drawback is that: the commonly used greedy strategy works well on the premise that the preceding choices are optimal or close to the optimal solution. However, in most cases, this strategy fails to guarantee the optimal solution. Moreover, the factors, such as the initial runs and the number of input documents, are not well investigated in most of the previous studies on implicit SRD.

In this paper, a novel Integer Linear Programming (ILP) formulation for implicit

SRD is proposed. Based on this formulation, the exactly optimal solution can be obtained and validated. We then compare the effectiveness of the proposed method *ILP4ID* with the state-of-the-art algorithms using the standard TREC diversity collections. The experimental results prove that *ILP4ID* can achieve improved performance over the baseline methods.

In §2, we first survey the well-known approaches for implicit SRD. In §3, the method *ILP4ID* based on ILP is proposed. A series of experiments are then conducted and discussed in §4. Finally, we conclude the paper in §5.

2 Related Work

In this section, we first give a brief survey of the typical approaches for SRD. For a detailed review, please refer to the work [3]. We begin by introducing some notations used throughout this paper. For a given query q, $D = \{d_1, ..., d_m\}$ represents the top-m documents of an initial retrieval run. $r(q, d_i)$ denotes the relevance score of a document d_i w.r.t. q. The similarity between two documents d_i and d_j is denoted as $s(d_i, d_j)$.

For implicit SRD, some approaches, such as *MMR* [1] and MPT [4], rely on the *greedy best first strategy*. At each round, it involves examining each document that has not been selected, computing a gain using a specific heuristic criterion, and selecting the one with the maximum gain. To remove the need of manually tuning the trade-off parameter λ, Sanner et al. [2] propose to perform implicit SRD through the greedy optimization of *Exp*-1-*call*@k, where a latent subtopic model is used in the sequential selection process.

The Desirable Facility Placement (*DFP*) model [6] is formulated as:

$$S^* = \max_{S \subset D, |S|=k} \lambda \cdot \sum_{d \in S} r(d) + (1-\lambda) \cdot \sum_{d' \in D \setminus S} \max_{d \in S} \{s(d, d')\} \tag{1}$$

where $\mathcal{R}(S) = \sum_{d \in S} r(d)$ denotes the overall relevance. $\mathcal{D}(S) = \sum_{d' \in D \setminus S} \max_{d \in S} \{s(d, d')\}$ denotes the diversity of the selected documents. For obtaining S^*, they initialize S with the k most relevant documents, and then iteratively refines S by swapping a document in S with another one in $D \setminus S$. At each round, interchanges are made only when the current solution can be improved. Finally, the selected documents are ordered according to the contribution to Eq. 1.

Instead of solving the target problem approximately, we formulate implicit SRD as an ILP problem. Moreover, the effects of different initial runs and the number of used documents on the diversification models have been explored. This study is complementary to the work by Yu and Ren [5], where explicit subtopics are required.

3 ILP Formulation for Implicit SRD

We formulate implicit SRD as a process of selecting and ranking k exemplar documents from the top-m documents. We expect to maximize not only the overall relevance of the k exemplar documents w.r.t. a query, but also the *representativeness* of the exemplar documents w.r.t. the non-selected documents. The ILP formulation of selecting k exemplar documents is given as:

$$\max_{\mathbf{x}} \lambda \cdot (m\text{-}k) \cdot \sum_{i=1}^{m} x_{ii} \cdot r(q, d_i) + (1\text{-}\lambda) \cdot k \cdot \sum_{i=1}^{m} \sum_{j=1:j \neq i}^{m} x_{ij} \cdot s(d_i, d_j) \tag{2}$$

$$s.t.\ x_{ij} \in \{0,1\},\ i \in \{1,...,m\}, j \in \{1,...,m\} \tag{3}$$

$$\sum_{i=1}^{m} x_{ii} = k \tag{4}$$

$$\sum_{j=1}^{m} x_{ij} = 1, i \in \{1,...,m\} \tag{5}$$

$$x_{jj} - x_{ij} \geq 0, i \in \{1,...,m\}, j \in \{1,...,m\} \tag{6}$$

In particular, the binary square matrix $\mathbf{x} = [x_{ij}]_{m \times m}$ is defined as: $m = |D|$, x_{ii} indicates whether document d_i is selected, and $x_{ij:i \neq j}$ indicates whether document d_i chooses document d_j as its exemplar. Restriction by Eq. 4 guarantees that k documents are selected. Restriction by Eq. 5 means that each document must have one representative exemplar. The constraint given by Eq. 6 enforces that if there is one document d_i selecting d_j as its exemplar, then d_j must be an exemplar. $\mathcal{R}^{'}(\mathbf{x}) = \sum_{i=1}^{m} x_{ii} \cdot r(q, d_i)$ depicts the overall relevance of the selected exemplar documents. $\mathcal{D}^{'}(\mathbf{x}) = \sum_{i=1}^{m} \sum_{j=1:j \neq i}^{m} x_{ij} \cdot s(d_i, d_j)$ denotes diversity. In view of the fact that there are k numbers (each number is in $[0,1]$) in the relevance part $\mathcal{R}^{'}(\mathbf{x})$, and m-k numbers (each number is in $[0,1]$) in the diversity part $\mathcal{D}^{'}(\mathbf{x})$. The coefficients m-k and k are added in order to avoid possible skewness issues, especially when $m \gg k$. Finally, the two parts are combined through the parameter λ.

Although solving arbitrary ILPs is an NP-hard problem, modern ILP solvers can find the optimal solution for moderately large optimization problems in reasonable time. We use the free solver GLPK in this study. Once the k exemplar documents are selected, they are further ranked in a decreasing order of their respective contributions to objective function given by Eq. 2. We denote the proposed approach as *ILP4ID*, namely, *a novel Integer Linear Programming method for implicit SRD*.

Looking back at *DFP* given by Eq. 1, if we view S as the set of exemplar documents, and $D \setminus S$ as the complementary set of non-selected documents, the calculation of $\max_{d \in S}\{s(d, d^{'})\}$ can be then interpreted as selecting the most representative exemplar $d \in S$ for $d^{'} \in D \setminus S$. Thus $\mathcal{D}(S)$ is equivalent to $\mathcal{D}^{'}(\mathbf{x})$. Therefore, *DFP* is a special case of *ILP4ID* when the coefficients m-k and k are not used. Since *ILP4ID* is able to obtain the exact solution w.r.t. the formulated objective function, and *DFP* relies on an approximate algorithm, thus *ILP4ID* can be regarded as the *theoretical upper-bound* of *DFP*.

Moreover, *MMR*, *MPT* and *QPRP* can be rewritten as different variants of *ILP4ID* since the study [6] has shown that they can be rewritten as different variants of *DFP*. However, *ILP4ID* is not the upper-bound of *MMR*, *MPT* and *QPRP*. Because the space of feasible solutions for *ILP4ID* and *DFP* relying on a two-step diversification is different from the one for *MMR* or *MPT* or *QPRP*, which generates the ranked list of documents in a greedy manner.

4 Experiments

4.1 Experimental Setup

The four test collections released in the diversity tasks of TREC Web Track from 2009 to 2012 are adopted (50 queries per each year). Queries numbered 95 and 100

are discarded due to the lack of judgment data. The evaluation metrics we adopt are nERR-IA and α-nDCG, where nERR-IA is used as the main measure as in TREC Web Track. The metric scores are computed using the top-20 ranked documents and the officially released script *ndeval* with the default settings. The ClueWeb09-T09B is indexed via the Terrier 4.0 platform. The language model with Dirichlet smoothing (denoted as *DLM*) and *BM25* are deployed to generate the initial run.

In this study, the models *MMR* [1], *MPT* [4], 1-*call*@k [2] and *DFP* [6] introduced in §2 are used as baseline methods. In particular, for 1-*call*@k, we follow the setting as [2]. For *MPT*, the relevance variance between two documents is approximated by the variance with respect to their term occurrences. For *DFP*, the iteration threshold is set to 1000. For *MMR*, *MPT*, *MPT* and the proposed model *ILP4ID*, we calculate the similarity between a pair of documents based on the Jensen-Shannon Divergence. The relevance values returned by *DLM* and *BM25* are then normalized to the range $[0, 1]$ using the MinMax normalization.

4.2 Experimental Evaluation

Optimization Effectiveness We first validate the superiority of *ILP4ID* over *DFP* in solving the formulated objective function. In particular, we set $\lambda = 0$ (for $\lambda \neq 0$, the results can be compared analogously), both *DFP* and *ILP4ID* work the same, namely selecting k exemplar documents. For a specific topic, we compute the representativeness (denoted as $\mathcal{D}$) of the subset S of k exemplar documents, which is defined as $\mathcal{D}'(\mathbf{x})$ in §3. The higher the representativeness is, the more effective the adopted algorithm is. Finally, for each topic, we compute the difference between $\mathcal{D}_{ILP4ID}$ and $\mathcal{D}_{DFP}$. As an illustration, we use the top-100 documents of the initial retrieval by *BM25*. Fig. 1 shows the performance of *DFP* and *ILP4ID* in finding the best k exemplars, where the x-axis represents the 198 topics, and the y-axis represents the representativeness difference (i.e., $\mathcal{D}_{ILP4ID} - \mathcal{D}_{DFP}$).

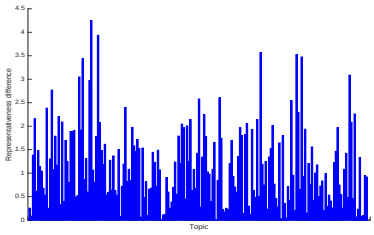

Fig. 1. Optimization effectiveness comparison.

From Fig. 1, we see that $\mathcal{D}_{ILP4ID} - \mathcal{D}_{DFP} \geq 0$ for all topics. Because *ILP4ID* always returns the exact solution for each topic, while *DFP* can not guarantee to find the optimal solution due to the adopted approximation algorithm. Since the process of selecting exemplar documents plays a fundamental role for implicit SRD, the effectiveness of DFP is therefore impacted, which is shown in the next section.

Implicit SRD Performance We use 10-fold cross-validation to tune the trade-off parameters, namely b for *MPT* and λ for *MMR*, *DFP* and *ILP4ID*. Particularly, λ is tuned in the range $[0, 1]$ with a step of 0.1. For b, the range is $[-10, 10]$ with a step of 1. The metric nERRIA@20 is used to determine the best result. Table 1 shows how

MMR, *MPT*, *DFP*, 1-*call*@k and *ILP4ID* vary when we change the initial runs (i.e., *BM25* and *DLM*) and the number of input documents (i.e., top-m documents of the initial run, where $m \in \{30, 50, 100\}$). Based on the Wilcoxon signed-rank test with $p < 0.05$, the superscripts $* \diamond \dagger$ indicate statistically significant difference to the best result of each setting, respectively.

Table 1. The performances of each model based on the initial run with *BM25* (columns 2-4) and the initial run with *DLM* (columns 6-8), respectively. The best result of each setting is indicted in bold.

m	Model	nERR-IA@20	α-nDCG@20	Model	nERR-IA@20	α-nDCG@20
	$BM25$	$0.2168^{\diamond\dagger}$	$0.2784^{*\diamond\dagger}$	DLM	$0.1596^{*\diamond\dagger}$	$0.2235^{*\diamond}$
30	$MMR(\lambda = 0.64)$	0.2257	0.2888	$MMR(\lambda = 0.1)$	0.1595^{*}	0.226^{*}
	$MPT(b = 10)$	0.2078^{*}	0.2701^{*}	$MPT(b = 10)$	0.176^{*}	0.2464
	$DFP(\lambda = 0.65)$	0.2285^{*}	0.2916^{*}	$DFP(\lambda = 0.4)$	**0.2177**	**0.2626**
	1-*call*@k	0.1918^{*}	0.2632^{*}	1-*call*@k	0.1873^{*}	0.248^{*}
	$ILP4ID(\lambda = 0.79)$	**0.2387**	**0.2995**	$ILP4ID(\lambda = 0.57)$	0.2107^{*}	0.2578^{*}
50	$MMR(\lambda = 0.62)$	0.2247	0.288	$MMR(\lambda = 0)$	$0.1353^{\diamond}$	$0.1983^{\diamond}$
	$MPT(b = 10)$	$0.1889^{\diamond}$	$0.2409^{\diamond}$	$MPT(b = 10)$	0.1823	**0.2542**
	$DFP(\lambda = 0.65)$	0.2522	0.3111	$DFP(\lambda = 0.4)$	0.197	$0.2394^{\diamond}$
	1-*call*@k	$0.1783^{\diamond}$	$0.2458^{\diamond}$	1-*call*@k	$0.1663^{\diamond}$	$0.2233^{\diamond}$
	$ILP4ID(\lambda = 0.78)$	**0.2565**	**0.3112**	$ILP4ID(\lambda = 0.57)$	**0.2026**	0.2445
100	$MMR(\lambda = 0.67)$	$0.2276^{\dagger}$	0.2917	$MMR(\lambda = 0)$	$0.1107^{\dagger}$	$0.1515^{\dagger}$
	$MPT(b = 10)$	$0.1646^{\dagger}$	$0.2059^{\dagger}$	$MPT(b = 10)$	0.161	**0.2227**
	$DFP(\lambda = 0.68)$	0.2489	0.3094	$DFP(\lambda = 0.4)$	**0.1836**	0.2181
	1-*call*@k	$0.1543^{\dagger}$	$0.2109^{\dagger}$	1-*call*@k	$0.1535^{\dagger}$	$0.1988^{\dagger}$
	$ILP4ID(\lambda = 0.78)$	**0.2618**	**0.3157**	$ILP4ID(\lambda = 0.56)$	$0.1731^{\dagger}$	0.2114

At first glance, we see that *BM25* substantially outperforms *DLM*. Moreover, given the better initial run by *BM25*, all the models tend to show better performance than that based on the initial run with *DLM*.

A closer look at the results (columns 2-4) shows that *MPT* and 1-*call*@k exhibit poor performance, which even does not enhance the naive-baseline results with $BM25$. For *MMR*, DFP and $ILP4ID$, they show a positive effect of deploying a diversification model. Moreover, the proposed model *ILP4ID* outperforms all the other models in terms of both nERR-IA@20 and α-nDCG@20 across different cutoff-values of used documents. When using the top-100 documents, the improvements in terms of nERR-IA@20 over $BM25$, *MMR*, *MPT*, *DFP* and 1-*call*@k are 20.76%, 15.03%, 59.05%, 5.18% and 69.67%, respectively.

Given a poor initial run with *DLM* (columns 6-8), for *MMR*, $\lambda = 0$ (i.e., using top-50 or top-100 documents) indicates that at each step *MMR* selects a document merely based on its similarity with the previously selected documents. When using only the top-30 documents, all models (except *MMR*) outperform *DLM* that does not take into account the feature of diversification. The improvements of *MPT*, DFP, 1-*call*@k and $ILP4ID$ over *DLM* in terms of nERR-IA@20 are 10.28%, 36.4%, 17.36% and 32.02%, respectively. However, when we increase the number of used documents of the initial retrieval, *MPT* shows a slightly improved performance when using the top-50 documents, but the other models consistently show decreased performance. For *MMR*, the results are even worse than *DLM*. These consistent variations imply that there are many *noisy documents* within the extended set of documents.

For MPT, $b = 10$ indicates that MPT performs a risk-aversion ranking, namely an unreliably-estimated document (with big variance) should be ranked at lower positions.

Given the above observations, we explain them as follows: Even though 1-*call*@k requires no need to fine-tune the trade-off parameter λ, the experimental results show that 1-*call*@k is not as competitive as the methods like *MPT*, DFP and $ILP4ID$, es-

pecially when more documents are used. The reason is that: for 1-*call*@k, both relevant and non-relevant documents of the input are used to train a latent subtopic model, thus it greatly suffers from the noisy information. Both *MMR* and *MPT* rely on the best first strategy, the advantage of which is that it is simple and computationally efficient. However, at a particular round, the document with the maximum gain via a specific heuristic criterion may cause error propagation. For example, for *MMR*, a long and highly relevant document may also include some noisy information. Once noisy information is included, the diversity score of a document measured by its maximum similarity w.r.t. the previously selected documents would not be precise enough. This well explains why *MMR* and *MPT* commonly show an impacted performance with the increase of the number of used documents. *DFP* can alleviate the aforesaid problem based on the swapping process. Namely, it iteratively refines S by swapping a document in S with another unselected document whenever the current solution can be improved. However, *DFP* is based on the hill climbing algorithm. A potential problem is that hill climbing may not necessarily find the global maximum, but may instead converge to a local maximum. *ILP4ID* casts the implicit SRD task as an ILP problem. Thanks to this, *ILP4ID* is able to simultaneously consider all the candidate documents and globally identify the optimal subset. The aforementioned issues are then avoided, allowing *ILP4ID* to be more robust to the noisy documents.

To summarize, *ILP4ID* substantially outperforms the baseline methods in most reference comparisons. Furthermore, the factors like different initial runs and the number of input documents greatly affect the performance of a diversification model.

5 Conclusion and Future Work

In this paper, we present a novel method based on ILP to solve the problem of implicit SRD, which can achieve substantially improved performance when compared to state-of-the-art baseline methods. This also demonstrates the impact of optimization strategy on the performance of implicit SRD. In the future, besides examining the efficiency, we plan to investigate the potential effects of factors, such as query types and the ways of computing document similarity, on the performance of diversification models, in order to effectively solve the problem of implicit SRD.

References

[1] Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st SIGIR. pp. 335–336 (1998)

[2] Sanner, S., Guo, S., Graepel, T., Kharazmi, S., Karimi, S.: Diverse retrieval via greedy optimization of expected 1-call@k in a latent subtopic relevance model. In: Proceedings of the 20th CIKM. pp. 1977–1980 (2011)

[3] Santos, R.L.T., Macdonald, C., Ounis, I.: Search result diversification. Foundations and Trends in Information Retrieval 9(1), 1–90 (2015)

[4] Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: Proceedings of the 32nd SIGIR. pp. 115–122 (2009)

[5] Yu, H., Ren, F.: Search result diversification via filling up multiple knapsacks. In: Proceedings of the 23rd CIKM. pp. 609–618 (2014)

[6] Zuccon, G., Azzopardi, L., Zhang, D., Wang, J.: Top-k retrieval using facility location analysis. In: Proceedings of the 34th ECIR. pp. 305–316 (2012)