A Concise Integer Linear Programming Formulation for Implicit Search Result Diversification*

Hai-Tao Yu[1,5], Adam Jatowt[2], Roi Blanco[3], Hideo Joho[4], Joemon M. Jose[5], Long Chen[5], Fajie Yuan[6]
1. yuhaitao@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. hideo@slis.tsukuba.ac.jp, University of Tsukuba, Japan / RMIT University, Australia
5. {Haitao.Yu, Joemon.Jose, Long.Chen}@glasgow.ac.uk, University of Glasgow, UK
6. f.yuan.1@research.gla.ac.uk, University of Glasgow, UK

ABSTRACT

To cope with ambiguous and/or underspecified queries, *search result diversification* (SRD) is a key technique that has attracted a lot of attention. This paper focuses on *implicit SRD*, where the possible subtopics underlying a query are *unknown* beforehand. We formulate implicit SRD as a process of selecting and ranking k exemplar documents that utilizes integer linear programming (ILP). Unlike the common practice of relying on approximate methods, this formulation enables us to obtain the optimal solution of the objective function. Based on four benchmark collections, our extensive empirical experiments reveal that: (1) The factors, such as *different initial runs*, *the number of input documents*, *query types* and *the ways of computing document similarity* significantly affect the performance of diversification models. Careful examinations of these factors are highly recommended in the development of implicit SRD methods. (2) The proposed method can achieve substantially improved performance over the state-of-the-art unsupervised methods for implicit SRD.

CCS Concepts

•**Information systems** → **Information retrieval diversity;**

Keywords

Cluster-based IR; implicit SRD; integer linear programming

*This work was carried out when Hai-Tao Yu was visiting the University of Glasgow, and Hideo Joho was visiting the RMIT University.

WSDM 2017, February 06-10, 2017, Cambridge, United Kingdom

DOI: http://dx.doi.org/10.1145/3018661.3018710

1. INTRODUCTION

Accurately and efficiently providing desired information to search engine users is a problem far from being resolved. A key issue is that users often submit short queries that are ambiguous and/or underspecified. Take, for example, the common query *Harry Potter*. It may actually refer to a book or a movie. For the movie, a user may be interested in any of many possible aspects including the main characters, movie reviews and so on. However, correctly identifying users' preferences is still quite difficult and prone to errors. As a remedy, one possible solution is to apply *search result diversification* (SRD), which is characterized as providing a diversified result to maximize the likelihood that an average user can find documents relevant to her specific need. Particularly, considering the above example of *Harry Potter*, such solution should generate an optimized result list that covers the possible aspects like *book* or *movie*. According to *whether the subtopics (i.e., different information needs) underlying a query are given beforehand or not*, the task of SRD can be differentiated into *implicit SRD* and *explicit SRD*. For implicit SRD, the possible subtopics underlying a query are *unknown*. In fact, finding a group of subtopic strings that covers well the possible underlying information needs of a query is a challenging issue. Most of the time, the explicit subtopics are not available, neither the training data for supervised methods (e.g., [2, 3, 33, 22, 37]). In such scenarios, the technique of implicit SRD is used for satisfying users' search intents. Accordingly, in this work, we do not investigate methods for explicit SRD nor supervised methods for result diversification, but, instead, *we focus on implicit methods*.

The state-of-the-art methods for implicit SRD differ mainly in the following aspects: (1) how to represent diversity; (2) how to balance relevance and diversity and (3) how to generate the result list. For example, the Maximal Marginal Relevance (*MMR*) model [4] measures the diversity of a document d_i based on the maximum similarity between d_i and the previously selected documents. In order to balance the relevance and diversity, *most of the methods use a trade-off parameter* λ. Finally, for generating the desired result list, *the common practice is using the greedy strategy that follows a heuristic criterion of making the locally optimal choice at each round* [4, 28, 7, 40]. 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. A natural question arises then: *to what extent does the greedy solution affect the performance of implicit SRD*? Moreover, when conducting experimental analysis, a single weighting model (say language model with Dirichlet smoothing [38]) is commonly adopted to perform the initial retrieval. Since the initially retrieved documents (e.g., top-m documents) are then further used to test diversification models, the impact of different initial runs on these diversification models is important. Furthermore, the effect of m (i.e., the number of used documents) and query types on the performance of a diversification model is also crucial. Unfortunately, these key points are not well investigated in most of the previous studies on implicit SRD.

The aforementioned drawbacks motivate us to approach implicit SRD in a novel way. In this paper, we propose a concise integer linear programming (ILP) formulation for implicit SRD. Based on this formulation, the exactly optimal solution can be obtained and validated. We then compare the effectiveness of the proposed method, called *ILP4ID*, against the state-of-the-art algorithms using the standard TREC diversity collections. The experimental results prove that *ILP4ID* can improve performance over the baseline methods in terms of standard diversity metrics.

The main contributions of this paper are as follows:

1. We present a concise ILP formulation for implicit SRD which allows for the exact solution of the objective function (Eq. 6) to be obtained. On the one hand, the proposed method can lead to substantially improved performance than the sate-of-the-art unsupervised methods. The experimental results also demonstrate how much accuracy has been lost due to the usage of an approximation method (e.g., compared with the method [40]). On the other hand, the flexibility of the proposed formulation allows for further extensions by simply changing the constraints.

2. Different from prior studies, we thoroughly investigate the effects of a series of factors on the performance of a diversification model. Our main finding is that the factors, such as *different initial runs*, *the number of input documents*, *query types* and *the ways of computing document similarity* greatly affect the effectiveness of diversification models for implicit SRD.

The reminder of the paper is structured as follows. In the next section, we first survey the well-known approaches for search result diversification. In Section 3, we formulate implicit SRD as an ILP problem, then *ILP4ID* method is proposed. A series of experiments are conducted and discussed in Section 4. Finally, we conclude the paper in Section 5.

2. RELATED WORK

This work is connected to two different research areas: *data clustering* and *cluster-based information retrieval* (IR). In this section, we first provide a short description of the popular Affinity Propagation (AP) algorithm for exemplar-based clustering, which lays the groundwork for the proposed method. Then, we concisely survey the typical approaches for cluster-based IR and implicit SRD. Due to space constraints, for a detailed review of AP, we refer the reader to the work [9, 11], and to [16, 29] for an overview of cluster-based IR and search result diversification.

2.1 Affinity Propagation for Clustering

The AP algorithm [9] has been deployed and extended in many research fields, such as detecting drug sensitivity [10], image categorization [31], image segmentation [34], and so on. Under the AP algorithm, clustering is viewed as identifying a subset of exemplars (i.e., representative items) given m items. The input is a symmetric matrix U representing the pairwise similarity of each pair of items where the diagonal values of U denote the prior beliefs of the m items in how likely each item is to be selected as an exemplar. In particular, AP assigns each non-exemplar item to an exemplar item. The objective is to maximize the sum of similarities between non-exemplar items and their assigned exemplar items. To this end, the technique of *belief propagation* is used, and the solution is generated through exchanging two types of real-valued messages. However, AP does not guarantee to find the optimal solution.

Inspired by AP, we formulate the implicit SRD as a process of selecting and ranking exemplar documents, and we use the *bound-and-branch method* to obtain the optimal solution. We have empirically found that using message-passing algorithm like AP for solving the objective (Eq. 6) suffers considerably from convergence issues. On the other hand, our proposed method can be used as a complementary method for solving data clustering problems, where the exact solution is to be expected.

2.2 Cluster-based IR and Implicit SRD

We begin by introducing some notations that are 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 $s(d_i, d_j)$.

A large body of work on cluster-based approaches for IR build upon the *cluster hypothesis* [24], which states that "closely associated documents tend to be relevant to the same requests". Some cluster-based methods rely on document clusters created offline by using the entire corpus [19, 17]. The methods utilizing *query-specific document clusters* are more popular, where the clusters are generated from documents by an initial retrieval performed in response to a query. For example, [20, 15] propose to enhance the ad-hoc retrieval performance, where document clusters are used to smooth documents' representations (e.g., language models). Recently, the cluster-based retrieval paradigm has been explored in the context of search result diversification, such as [14] and [23]. Raiber and Kurland [23] studied how to incorporate various types of cluster-related information based on Markov Random Fields.

Regarding implicit SRD, in order to obtain the optimal ranked list L^*, the most intuitive way is to apply the *greedy best first strategy*. At the beginning, this strategy initializes L with the most relevant document d_1^*, and then it selects the subsequent documents one by one via a specific heuristic criterion:

$$d_j^* = \underset{d_j \in D \setminus L_{j-1}}{\operatorname{argmax}} \{\lambda r(q, d_j) + (1 - \lambda) W(d_j, L_{j-1})\} \qquad (1)$$

where $L_{j-1} = \{d_1^*, ..., d_{j-1}^*\}$, $W(d_j, L_{j-1})$ measures how far

d_j disperses w.r.t. L_{j-1}. At every round, it involves examining each document that has not been selected, computing a gain using the above heuristic criterion, and selecting the one with the maximum gain. A typical instance of this strategy is the *MMR* model [4], in which $W(d_j, L_{j-1})$ is defined as - $\max_{d_i \in L_{j-1}} s(d_i, d_j)$. In other words, the diversity under *MMR* is measured through the maximum similarity between d_j and the previously selected documents. Furthermore, Guo and Sanner [13] present a probabilistic latent view of MMR, where the need of manually tuning λ is removed. Later on, the greedy optimization of *Exp*-1-*call*@k [27] for implicit SRD was proposed. The well-known Modern Portfolio Theory (MPT) [30] model takes into account the expected relevance and relevance variance of a document, and the correlations with the already selected documents. It sequentially selects documents that maximize the following criterion

$$E(d_k) - b \cdot w_k \cdot \sigma_k^2 - 2b \sum_{i=1}^{k-1} w_i \cdot \sigma_i \cdot \sigma_k \cdot \rho_{ik} \tag{2}$$

where $E(d_k)$ is the expected relevance of d_k, and σ_k is the standard deviation, w denotes the rank-specific weigh, and ρ_{ik} denotes the correlation coefficient between d_i and d_k.

Another line of studies (referred to as *top-k retrieval* in [40, 12, 14]) for implicit SRD perform a two-step process. The first step is to select an optimal subset $S \subset D$ of k documents according to a specific objective function. At the second step, the selected documents in S are ordered in a particular way, e.g., in a decreasing order of relevance. Moreover, Gollapudi and Sharma [12] propose a set of natural axioms analyzing the properties of a diversification function. A more general model (referred to as Desirable Facility Placement *DFP*) by Zuccon et al. [40] is given as:

$$S^* = \underset{S \subset D, |S|=k}{\operatorname{argmax}} \lambda \cdot \mathcal{R}(S) + (1-\lambda) \cdot \mathcal{D}(S) \tag{3}$$

$$\mathcal{R}(S) = \sum_{d \in S} r(d) \tag{4}$$

$$\mathcal{D}(S) = \sum_{d' \in D \setminus S} \max_{d \in S} \{s(d, d')\} \tag{5}$$

where $\mathcal{R}(S)$ denotes the overall relevance. $\mathcal{D}(S)$ denotes the diversity of the selected documents, which is captured by measuring the representativeness of the selected documents w.r.t. the non-selected ones, $\lambda \in [0, 1]$ is a trade-off parameter. To obtain S^*, they use the *greedy best k strategy*. It initializes S with an arbitrary solution (e.g., 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. The process terminates after convergence or after a fixed number of iterations.

Our work is a further endeavor to the cluster-based retrieval paradigm. The studies most related to ours are [35, 40, 14, 23]. However, the ILP formulation by Yu and Ren [35] is proposed to perform explicit SRD, which requires pre-collected subtopics as the input. For implicit SRD, the methods [40, 14, 23] appeal to approximate methods for generating clusters. Our formulation of implicit SRD based on ILP allows to obtain the optimal solution, which makes it possible to investigate how much accuracy has been lost due to approximations (e.g., compared with *DFP*).

3. PROPOSED METHOD

In this section, we first describe the method *ILP4ID* proposed for implicit SRD. We then discuss the differences and connections between *ILP4ID* and the previous approaches.

3.1 ILP Formulation for Implicit SRD

In this section, we formulate the task of implicit SRD as a process of selecting and ranking k exemplar documents from the top-m documents of an initial retrieval. We call a document as *exemplar* if it is selected to represent a group of documents based on some measure of similarity. On the one hand, we expect to maximize the overall relevance of the k exemplar documents w.r.t. a query. On the other hand, we wish to maximize the *representativeness* of the exemplar documents w.r.t. the non-selected documents. This motivation follows the aforesaid cluster hypothesis [24]. Intuitively, if the selected exemplars concisely represent the entire set of documents, the novelty and diversity will naturally arise.

To clearly describe the way of identifying the expected k exemplar documents, we introduce the binary square matrix $\mathbf{x} = [x_{ij}]_{m \times m}$ such that $m = |D|$, x_{ii} indicates whether document d_i is selected as an exemplar or not, and $x_{ij:i \neq j}$ indicates whether document d_i "chooses" document d_j as its exemplar. The process of selecting k exemplar documents is then expressed as the following ILP problem:

$$\max_{\mathbf{x}} \lambda \cdot (m\text{-}k) \cdot \mathcal{R}^{'}(\mathbf{x}) + (1\text{-}\lambda) \cdot k \cdot \mathcal{D}^{'}(\mathbf{x}) \tag{6}$$

$$\mathcal{R}^{'}(\mathbf{x}) = \sum_{i=1}^{m} x_{ii} \cdot r(q, d_i) \tag{7}$$

$$\mathcal{D}^{'}(\mathbf{x}) = \sum_{i=1}^{m} \sum_{j=1:j \neq i}^{m} x_{ij} \cdot s(d_i, d_j) \tag{8}$$

$$s.t.\ x_{ij} \in \{0, 1\},\ i \in \{1, ..., m\}, j \in \{1, ..., m\} \tag{9}$$

$$\sum_{i=1}^{m} x_{ii} = k \tag{10}$$

$$\sum_{j=1}^{m} x_{ij} = 1, i \in \{1, ..., m\} \tag{11}$$

$$x_{jj} - x_{ij} \geq 0, i \in \{1, ..., m\}, j \in \{1, ..., m\} \tag{12}$$

In particular, the restriction given by Eq. 10 guarantees that k documents are selected. The restriction by Eq. 11 means that each document must have only one representative exemplar. The constraint given by Eq. 12 enforces that if there is one document d_i selecting d_j as its exemplar (i.e., $x_{ij} = 1$), then d_j must be an exemplar (i.e., $x_{jj} = 1$). $\mathcal{R}^{'}(\mathbf{x})$ depicts the overall relevance of the selected exemplar documents. $\mathcal{D}^{'}(\mathbf{x})$ denotes diversity. In other words, the diversity is expressed through selecting documents that represent the intrinsic diverse information revealed by the input documents. 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 as shown in Eq. 6. Once the k exemplar documents are selected, they are further ranked in the decreasing order of their respective contributions to objective function given by Eq. 6. We denote the proposed approach as *ILP4ID*,

namely, *a concise integer linear programming approach for implicit SRD*.

A number of successful ILP formulations have been developed for natural language processing tasks, such as semantic role labelling [26], syntactic parsing [21] and summarisation [32]. Yet, the ILP formulation we present is, to the best of our knowledge, the first one for implicit SRD. In fact, the above ILP formulation is quite flexible, and different variants can be derived by simply changing the constraints. For example, when removing the constraint by Eq. 10, the relevance expression (by Eq. 7) and the coefficients m-k and k in Eq. 6, the above formulation boils down to an equivalent ILP formulation of AP. It would be interesting to make an in-depth comparison between AP and its ILP formulation in the future, which helps to know *to what extent AP diverges from the optimal solution*.

3.2 Connections with Prior Models

Looking back at the model *DFP* given by Eqs. 3, 4 and 5, if we view S as the set of exemplar documents, and $D \setminus S$ as the complementary set of non-selected documents, calculating $\max_{d \in S}\{s(d, d^{'})\}$ can be then interpreted as selecting the most representative exemplar $d \in S$ for $d^{'} \in D \setminus S$. Thus $\mathcal{D}(S)$ is essentially equivalent to $\mathcal{D}^{'}(\mathbf{x})$. In addition, $\mathcal{R}(S)$ is also equivalent to $\mathcal{R}^{'}(\mathbf{x})$. Therefore, *DFP* can be viewed as 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, its performance can be regarded as the *upper-bound* of formulations of this kind.

Moreover, the study by Zuccon et al. [40] also shows that there are close connections between *DFP* and the models like *MMR* [4], *MPT* [30] and Quantum Probability Ranking Principle (QPRP) [39]. Namely, *MMR*, *MPT* and *QPRP* can be rewritten as different variants of *DFP* (the reader can refer to [40] for detailed derivation). Analogously, *MMR*, *MPT* and *QPRP* can also be rewritten as different variants of *ILP4ID*. The detailed derivation can be obtained based on the work [40]. However, it should be noted that: the space of feasible solutions for *ILP4ID* and *DFP* is different from the one for *MMR* or *MPT* or *QPRP*. This is because both *ILP4ID* and *DFP* rely on a two-step diversification, while *MMR*, *MPT* and *QPRP* directly generate the ranked list of documents in a greedy manner.

4. EXPERIMENTS

In this section we report a series of experiments conducted to evaluate the performance of the proposed method by comparing it to the state-of-the-art implicit diversification approaches. In the following, we first detail the test collections and the topics as well as the evaluation metrics used in the experiments. We then describe the configuration of each method to be evaluated, including the parameter setting and the ways of computing relevance scores, document similarity, etc. Finally, we describe the experimental results.

4.1 Test Collections and Metrics

Four standard test collections released in the diversity tasks of TREC Web Track from 2009 to 2012 are adopted for the experiments (50 queries per each year). Each query is structured as a set of a representative subtopics. Moreover, each query is further categorized as either "faceted" or "ambiguous" [5]. Queries numbered 95 and 100 in TREC 2010 are discarded due to the lack of judgment data, resulting in 198 queries being finally used. The evaluation metrics we adopt are nERR-IA (normalized Intent-Aware Expected Reciprocal Rank) [1] and α-nDCG (novelty-biased Discounted Cumulative Gain) [6], where nERR-IA is used as the main effectiveness measure in this study same as in TREC Web Track. In particular, the performance is evaluated using the top-20 ranked documents and the officially released script *ndeval* with the default settings.

The ClueWeb09 Category B collection is indexed with the Terrier 4.0 platform[1]. Two ad-hoc weighting models are deployed for investigating the effect of initial runs, i.e., *language model with Dirichlet smoothing* [38] (denoted as *DLM*) and *BM25* [25] based on the default setting of Terrier 4.0.

4.2 Baselines and Model Configuration

The models *MMR* [4], *MPT* [30], 1-*call*@k [27] and *DFP* [40] introduced in Section 2 are used as baseline methods. Similar to 1-*call*@k, He et al. [14] have also used the Latent Dirichlet Allocation (LDA) topic model for document clustering, while Raiber and Kurland [23] have utilized a supervised method (i.e., SVM^{rank}) to utilize the cluster information. Due to these reasons, [14] and [23] are not compared in this study. When it comes to 1-*call*@k, we follow the same setting as in [27]. The LDA model (α=2.0, β=0.5) is trained based on the top-m results for each query and the obtained subtopic distributions are used for the similarity and diversity computation. In particular, the topic number is set to: 15 (when $m \leq 100$), 20 (when $100< m \leq 300$), 25 (when $300< m \leq 500$) and 40 (when $500< m \leq 1000$). 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 $1,000$) and the proposed model *ILP4ID*, the k is set to 20.

For *MMR*, *DFP* and *ILP4ID*, we calculate the similarity between a pair of documents in two ways. One is the Jensen-Shannon Divergence (denoted as JSD) between document language models (e.g., DLM), which is a symmetric and smoothed version of KL divergence. The other is the cosine similarity based on tf-idf weight vectors (denoted as COS). The relevance values returned by *DLM* and *BM25* are then normalized to the range $[0, 1]$ using the MinMax normalization [18]. Using the same methods to compute both the relevance score and the document-to-document similarity in all the studied approaches enables us to conduct a fair comparison when investigating the impact of a specific component (e.g., the adopted optimization strategy) on the performance.

4.3 Experimental Evaluation

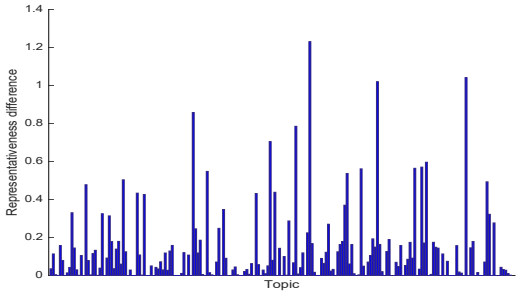
In the following experiments, we first compare the optimization effectiveness between DFP and $ILP4ID$. We then describe the differences of the used initial runs by *DLM* and *BM25*. Later, we investigate the models from different perspectives, including the effectiveness and efficiency.

4.3.1 Optimization Effectiveness

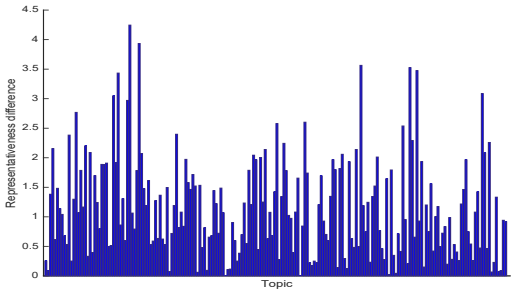
Before investigating the effectiveness of the aforementioned methods in performing implicit SRD, we first validate the

[1]http://terrier.org/

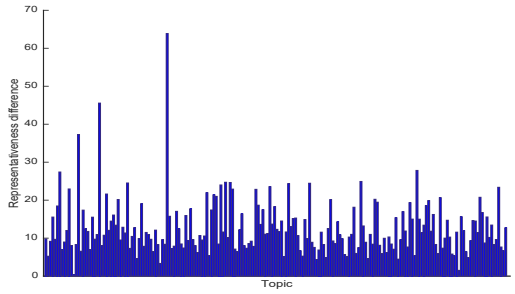
superiority of *ILP4ID* over *DFP* when solving the formulated objective function (Eq. 3 and Eq. 6). In particular, we set $\lambda = 0$ for both *DFP* and *ILP4ID*, and remove the coefficient k for *ILP4ID*. Thanks to this, *DFP* and *ILP4ID* are enforced to work in the same way, namely by selecting k exemplar documents without ranking. For a specific topic, we then compute the representativeness (denoted as $\mathcal{D}$) of the subset S of k exemplar documents, which is defined as $\mathcal{D}(S)$ in Eq. 5. 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 $\mathcal{D}_{DFP}$ that is the difference between the representativeness by *ILP4ID* and the one by *DFP*. As an illustration, we use the top-50, 100 and 500 documents of the initial retrieval by *BM25*, respectively. Fig. 1 shows the performance of *DFP* and *ILP4ID* in finding the best k exemplars, where the x-axis represents the 198 queries, and the y-axis represents the difference of the representativeness (i.e., $\mathcal{D}_{ILP4ID} - \mathcal{D}_{DFP}$).



(a) Using top-50 documents.



(b) Using top-100 documents.



(c) Using top-500 documents.

Figure 1: Optimization effectiveness comparison between *DFP* and *ILP4ID*.

From Fig. 1, we can clearly observe that regardless of how many documents are used, $\mathcal{D}_{ILP4ID} - \mathcal{D}_{DFP} \geq 0$ for all the queries. When the number of documents increases, so does the representativeness difference values. Specifically, the total sum of difference values in Fig. 1(a) is 24.73, the total sum in Fig. 1(b) is 228.36 and the total sum in Fig. 1(c) is 2,499.28. Thus, it is reasonable to say that *DFP* is feasible for small tasks. But for a moderately larger task, the solution obtained by *DFP* significantly diverges from the optimal solution w.r.t. the objective formulation. This is because *DFP* obtains the solution based on an approximation algorithm (i.e., the hill climbing algorithm), while *ILP4ID* finds the exact solution based on the branch-and-bound algorithm. *ILP4ID* always returns the exact solution, while *DFP* can not guarantee to find the optimal solution. Fig. 1 shows us that *DFP* commonly finds a sub-optimal solution. Since the process of selecting exemplar documents plays a fundamental role for implicit SRD, the effectiveness of *DFP* is therefore greatly impacted, which is shown in Sections 4.3.3, 4.3.4 and 4.3.5.

4.3.2 Analysis of Initial Runs

Since the diversification models take the initially retrieved documents by either *DLM* or *BM25* as input, a thorough exploration of the results by *DLM* and *BM25* is necessary in order to understand the effectiveness of each diversification model. Table 1 shows the performance in terms of nERR-IA@20 and α-nDCG@20, where the superscript $*$ indicates statistically significant differences when compared to the best result based on the Wilcoxon signed-rank test with $p < 0.05$.

Table 1: Performance of the initial retrieval. For each measure, the best result is indicted in bold.

Initial retrieval model	nERR-IA@20	α-nDCG@20
DLM	0.1596*	0.2235*
BM25	**0.2168**	**0.2784**

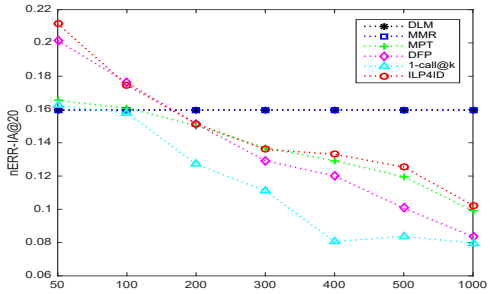
From Table 1, we can observe that *BM25* has significantly better performance than *DLM*. To examine how many relevant documents there are in each initial run, we can look at Fig. 2, which shows the averaged number of documents that provide information relevant to at least one subtopic in the initial run. The x-axis denotes the cutoff values (i.e., the top-m documents to be used).

Fig. 2 demonstrates that the results by *BM25* provide more relevant documents than that of *DLM*. At the same time, Fig. 2 also indicates to what extent the noisy documents will be mixed when we increase the number of used documents.

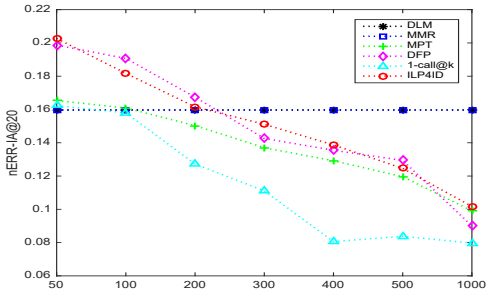
In the following experiments, the results of *DLM* and *BM25* are also used as naive baselines without diversification, which helps to show the positive/negative effects of deploying a diversification model. Using different ad-hoc weighting models, we can investigate the effect of an initial run. In particular, the experiments over the retrieval with *BM25* will allow to study the effect of using a high-quality initial run, while the ones with *DLM* will let us analyze the effect of providing a poor quality initial retrieval.

4.3.3 Implicit SRD Performance

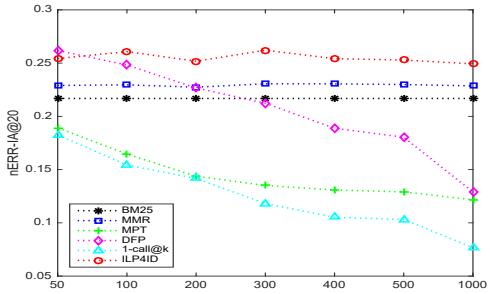
In this section, we examine how the diversification models vary when we change the initial runs (i.e., *DLM* and *BM25*), the number of input documents (i.e., $m \in \{50, 100, 200, 300,$



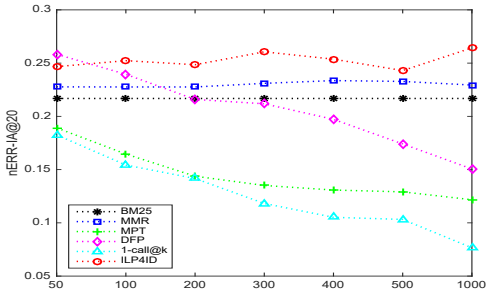
(a) Initial run: *DLM*; document similarity: COS.



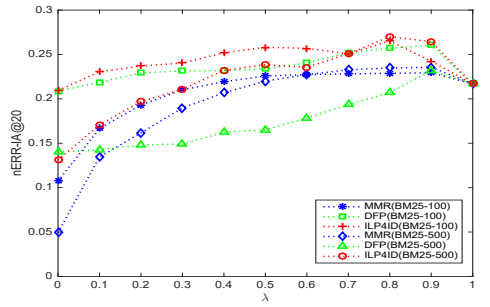
(b) Initial run: *DLM*; document similarity: JSD.



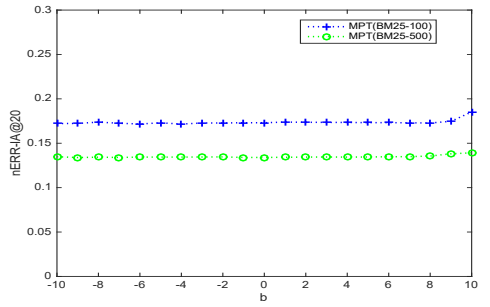
(c) Initial run: *BM25*; document similarity: COS.



(d) Initial run: *BM25*; document similarity: JSD.



(e) Comparison with a uniform λ setting.



(f) Performance variation of *MPT* w.r.t. b.

Figure 3: Cross-validation performance for implicit SRD (Figs 3(a)-3(d)), where the x-axis indicates the number of used documents. Per-λ comparison (Fig. 3(e)). Per-b performance of *MPT* (Fig. 3(f)).

Table 2: Performance of different models w.r.t. faceted and ambiguous queries. The best result of each setting is indicted in bold. The superscript † indicates statistically significant difference when compared to the best result based on the Wilcoxon signed-rank test with $p < 0.05$.

Data	Model	Type	$nERR-IA@20$			$\alpha-nDCG@20$		
			top-100	top-300	top-1000	top-100	top-300	top-1000
Faceted: 141 Ambiguous: 57	BM25	Faceted	0.2515	0.2515	0.2515	0.316	0.316†	0.316†
		Ambiguous	0.131†	0.131†	0.131†	0.1852†	0.1852†	0.1852†
	MMR	Faceted	0.2622	0.269	0.2659	0.3294	0.337	0.3337
		Ambiguous	0.1421	0.137†	0.1389†	0.2009	0.2005†	0.1981†
	MPT	Faceted	0.1898†	0.1578†	0.151†	0.2302†	0.1704†	0.1496†
		Ambiguous	0.1024†	0.0789†	0.0492†	0.1448†	0.1081†	0.0532†
	DFP	Faceted	0.2666†	0.2264†	0.1679†	0.3321	0.3007†	0.2383†
		Ambiguous	0.1726	0.1756†	0.1079†	**0.2254**	0.2238	0.1601†
	$1-call@k$	Faceted	0.1779†	0.1287†	0.0847†	0.2326†	0.1755†	0.1133†
		Ambiguous	0.0959†	0.0922†	0.0565†	0.1482†	0.1343†	0.0877†
	ILP4ID	Faceted	**0.2832**	**0.2804**	**0.2914**	**0.3455**	**0.349**	**0.358**
		Ambiguous	**0.176**	**0.2116**	**0.1971**	0.2194†	**0.2492**	**0.2423**

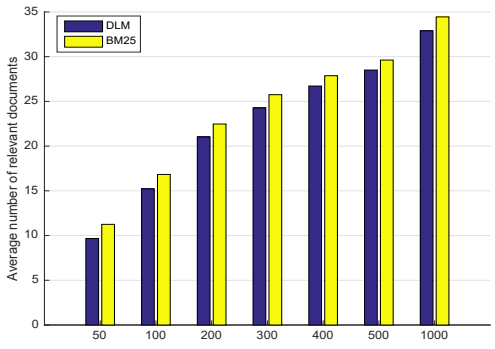


Figure 2: The statistics of the average number of relevant documents within the adopted initial runs.

$400, 500, 1000\}$ on the x-axis) and the ways for computing document similarity (i.e., *COS* and *JSD*).

We use 10-fold cross-validation to tune the trade-off parameters, namely b for *MPT* and λ for *MMR*, *DFP* and *ILP4ID*. Particularly, we explore the optimal results of *MMR*, *DFP* and *ILP4ID* by varying λ in the range $[0, 1]$ with a step of 0.1. We tune the b parameter of *MPT* with the range $[-10, 10]$, and a step of 1. The metric nERRIA@20 is used to determine the best results. Finally, the results are illustrated in Figs. 3(a)-3(d).

In Figs. 3(a) and 3(b), we note that the λ value of *MMR* determined via cross-validation is 1.0. Thus *MMR* fails to diversify the results (cf. Eq. 1). This is also why the performance curves of *MMR* overlap with those of *DLM* and *BM25*. The effect of tuning λ is detailed in Section 4.3.4.

At first glance, Figs. 3(a) and 3(b) based on *DLM* reveal that all the diversification models except *MMR* exhibit positive effectiveness when using the smaller number of documents (top-50 documents). We also see that *DFP* and *ILP4ID* which belong to the cluster-based diversification paradigm are more effective than other formulations, such as *MPT* and 1-*call*@k. This observation is consistent with the previous reports [40]. However, when we increase the initial number of retrieved documents, *MPT*, *DFP*, 1-*call*@k and *ILP4ID* consistently show decreased performance. In particular, when the number of used documents is quite large, these models can not even improve over the naive-baseline results with *DLM*. The plausible reason is that more *noisy documents* are included in larger document sets. This is actually supported by Fig. 2 which shows that relatively more non-relevant documents are included with the increase of the threshold of used documents.

A closer look at Figs. 3(a) and 3(b) reveals that the ways of computing document similarity also affects the performance of both *DFP* and *ILP4ID*, where the performance of *MPT* and 1-*call*@k can be used as a static reference since they do not rely on either *COS* or *JSD*. Sometimes, *DFP* can lead to better results than *ILP4ID*, e.g., using top-100/200 documents in Fig. 3(b). This may result from the second ranking procedure after the k exemplar documents have been selected. This is also where *ILP4ID* should be further improved.

When changing the initial run, i.e., using a better one such as *BM25*, Figs. 3(c)-3(d) demonstrate that the diversification models have quite different performances. Specifically, all the models tend to show better performance than the one based on the initial run with *DLM*. *MPT*, *DFP* and 1-*call*@k are characterized by the decreased performance when we increase the number of used documents. However, *MMR* and *ILP4ID* always demonstrate a positive diversification performance that does not degrade when increasing the number of documents. *ILP4ID* outperforms the other models in most reference comparisons.

Now we investigate the possible reasons for the above findings. Even though 1-*call*@k does not require 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*. The most possible explanation is that the top-m documents are directly used to train a latent subtopic model. As Fig. 2 shows, a large portion of documents are non-relevant, thus this method greatly suffers from the noisy information. Another awkward factor that may affect 1-*call*@k is that the topic number of the subtopic model has to be fine-tuned, otherwise the representation of each document as a subtopic vector would not be sufficiently precise.

Both *MMR* and *MPT* rely on the best first strategy, the advantage of which is that it is simple and computationally efficient (cf. Fig. 4). However, at a particular round, the document with the maximum gain via a specific heuristic criterion (i.e., Eq.1 of *MMR* and Eq.2 of *MPT*) may incur *error propagation*. For example, a long and relevant document may also include some noisy information. Once noisy information is included in the algorithm process, the diversity score of a document measured with respect to the previously selected documents would not be correct. This largely explains why both *MMR* and *MPT* underperform *DFP* and *ILP4ID* that globally select documents.

DFP can alleviate the aforesaid problem (i.e., error propagation) based on the swapping process as 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. In contrast, *ILP4ID* casts the process of selecting exemplar documents as an ILP problem. Thanks to this, *ILP4ID* is able to simultaneously consider all the candidate documents and to globally identify the optimal subset. The potential issue of error propagation is then avoided, making *ILP4ID* more robust to the noisy documents and letting it outperform the other models.

To summarize, *DFP* and *ILP4ID* which belong to the cluster-based diversification paradigm are more effective than *MMR*, *MPT* and 1-*call*@k. This echoes the findings in the previous work on cluster-based IR [40, 14, 23]. Benefiting from the advantage of obtaining the optimal solution, *ILP4ID* substantially outperforms the baseline methods in most reference comparisons. Furthermore, *for implicit SRD, the factors like different initial runs, the number of input documents, and the ways of computing document similarity greatly affect the performance of a specific model.*

4.3.4 Effects of Trade-off Parameters

To clearly show the effect of the trade-off parameters λ and b for balancing relevance and diversity, we investigate how *MMR*, *MPT*, *DFP* and *ILP4ID* vary per-λ or per-b.

Specifically, the top-100, 500 documents of the initial run with $BM25$ are used, respectively. All the 198 queries are tested. λ is set in the range $[0, 1]$ with a step of 0.1, and b is set in the range $[-10, 10]$ with a step of 1. In particular, for MMR, DFP and $ILP4ID$, $\lambda \in (0, 1)$ implies that the ranking process relies on both the relevance part and diversity part. The closer λ is to 1, the less effect the diversity part has. With $\lambda = 1$, MMR, DFP and $ILP4ID$ simply rely only on the relevance of documents, hence, they have the same performance as the initial run. With $\lambda = 0$, the performance of a model merely depends on the ability of selecting the representative documents. Regarding the effect of b on MPT (cf. Eq. 2), a positive b indicates that MPT performs a risk-aversion ranking, namely an unreliably-estimated document (with high variance) should be ranked at lower positions. The smaller b is, the less risk-averse the ranking.

In terms of ERR-IA@20, Fig. 3(e) shows how MMR, DFP and $ILP4ID$ vary with a uniform λ setting, and Fig. 3(f) demonstrates how MPT varies per-b.

From Fig. 3(e), we see that tuning λ has a large effect in the performance. This indicates that λ needs to be fine-tuned to achieve an optimal performance. The performance of MPT is slightly enhanced when b is close to 10. When b is set using smaller values, the effect is not quite obvious. Moreover, a closer look at Figs. 3(e)-3(f) reveals that $ILP4ID$ outperforms the baseline methods across most λ settings (and b for MPT), even though different numbers of documents of the initial run are used. This again clearly attests the potential merits of the proposed method for implicit SRD.

4.3.5 *Effectiveness w.r.t. Query Types*

The adopted dataset contains 141 faceted queries and 57 ambiguous queries. We now investigate the effectiveness of the different methods with resect to their type, either *faceted* or *ambiguous*. In particular, the comparison is conducted based on the initial retrieval with $BM25$ by using the top-100, 300 and 1,000 documents, separately. Table 2 shows the results obtained for MMR, MPT, DFP, 1-$call@k$ and $ILP4ID$ on faceted and ambiguous queries, respectively.

At first glance, Table 2 shows that all models perform worse in terms of both $nERR\text{-}IA@20$ and $\alpha\text{-}nDCG@20$ on ambiguous queries than they do on faceted queries. This reveals that it is relatively harder to select diverse relevant documents for ambiguous queries. Such situation mainly results from the intrinsic difference between faceted queries and ambiguous queries. The TREC assumption [5] goes like this: For an ambiguous query that has diverse interpretations, users are assumed to be interested in only one of these interpretations. For a faceted query that reflects an under-specified subtopic of interest, the users are assumed to be interested in one subtopic, but they may still be interested in others as well. That is, heterogeneous documents providing more divergently relevant information are required for ambiguous queries. We examined the distribution of relevant documents based on the ground-truth files. For each query type, we computed the average number of relevant documents and the average number of relevant documents that are relevant to at least 2 subtopics (termed *multi-relevant documents*). For faceted queries, these numbers are 112.42 and 47.27 whereas for ambiguous queries they are 109.35 and 19.6, respectively. These results, especially the average number of multi-relevant documents, demonstrate that it is relatively easy to retrieve some relevant documents to satisfy the subtopics of faceted queries, thus higher $nERR\text{-}IA@20$ and $\alpha\text{-}nDCG@20$ scores are observed in Table 2.

In terms of both $nERR\text{-}IA@20$ and $\alpha\text{-}nDCG@20$, $ILP4ID$ outperforms the baseline methods in most reference comparisons for both the types of queries. Quite a few of the improvements are statistically significant.

4.4 Efficiency

Common formulations of search result diversification (say, MPT, DFP and $ILP4ID$) are NP-hard (cf. [8, 29] for detailed analysis), thus approximate methods are generally adopted to find the solution. Although solving arbitrary ILPs is also an NP-hard problem, various efficient branch-and-bound algorithms have been developed. In fact, modern ILP solvers (e.g., GLPK, CPLEX and Gurobi) can find the optimal solution for moderately large optimization problems in a reasonable amount of time. In this paper we use the Gurobi solver.

In our study, we have also evaluated the overhead of MMR, MPT, DFP, 1-$call@k$ and $ILP4ID$ by measuring the average run-time per query when generating the diversified results. All the experiments are conducted using Java (JRE 1.8.0_31) on an IMac (Intel Core i7, 4GHz, 32 GB of RAM). Based on the initial run by $BM25$, Fig. 4 plots the run-time of each model (i.e., y-axis) versus the number of input documents (i.e., x-axis).

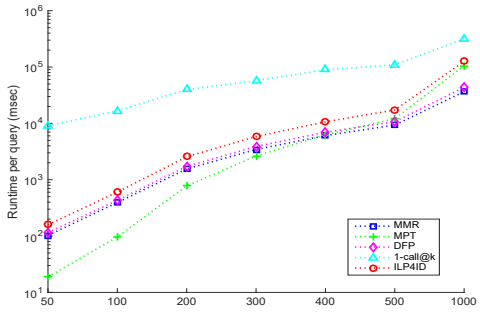


Figure 4: Average runtime per-query (msec).

From Fig. 4, we see that although both MMR and MPT rank documents sequentially, MPT requires less time when dealing with a small number of documents (say less than 400 documents). However, when the amount of documents increases, MPT requires more time than MMR and DFP. The main overhead is incurred by the calculation of relevance variance based on term occurrences (the time complexity is $\mathcal{O}(m^2 \cdot |W|)$, where W denotes the number of unique terms within the top-m documents). Although the formulations of DFP and $ILP4ID$ are similar, $ILP4ID$ has a higher computational cost. This is not surprising given the deployment of a branch-and-bound algorithm in order to obtain the optimal solution. Moreover, 1-$call@k$ is the most computationally expensive. In fact, the time overhead is mostly caused by training the LDA subtopic model. We note that these results should be considered as indicative only as it is possible to optimize the codes of each method, which is beyond the scope of this paper (for example, using the highly-efficient algorithm [36] for topic modeling, distributed algorithms for solving ILP problems[2], etc.)

[2]http://www.gurobi.com/products/distributed-

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel model to solve the problem of implicit SRD. The key idea is to formulate implicit SRD as a process of selecting and ranking k exemplar documents from the top-m documents of an initial retrieval. By relying on the ILP formulation, we are able to obtain the optimal solution of the target formulation. We have shown that the proposed method *ILP4ID* leads to substantially improved performance when compared to state-of-the-art baseline methods, which helps to demonstrate the impact of optimization strategy on implicit SRD. Since problems analogous to implicit SRD arise in a variety of applications, e.g., recommender systems, we believe that our method provides a new perspective for addressing problems of this kind.

Even though we addressed the problem of obtaining the optimal set of exemplar documents for implicit SRD, the following practical issues have not been explored well in this work. First, the optimal k of *ILP4ID* essentially differs from query to query. The effect of tuning k on *ILP4ID* is then worthy to be investigated in the future. Second, given the optimal set of exemplar documents, the documents have been essentially clustered. However, the following information, such as the internal correlations among documents within the same cluster and the external correlations among clusters, has not been well utilized. For future work, we also plan to study how to induce a high-quality ranking of documents by taking into account the aforesaid information.

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the 2nd WSDM*, pages 5–14, 2009.

[2] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd ICML*, pages 89–96, 2005.

[3] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th ICML*, pages 129–136, 2007.

[4] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st SIGIR*, pages 335–336, 1998.

[5] C. L. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2009 web track. In *TREC*, 2009.

[6] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st SIGIR*, pages 659–666, 2008.

[7] V. Dang and W. B. Croft. Diversity by proportionality: an election-based approach to search result diversification. In *Proceedings of the 35th SIGIR*, pages 65–74, 2012.

[8] T. Deng and W. Fan. On the complexity of query result diversification. *ACM Transactions on Database Systems*, 39(15):15:1–15:46, 2014.

[9] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

optimization

[10] M. J. Garnett, E. J. Edelman, S. J. Heidorn, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, pages 570–575, 2012.

[11] I. E. Givoni and B. J. Frey. A binary variable model for affinity propagation. *Neural Computation*, 21(6):1589–1600, 2009.

[12] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *Proceedings of the 18th WWW*, pages 381–390, 2009.

[13] S. Guo and S. Sanner. Probabilistic latent maximal marginal relevance. In *Proceedings of the 33rd SIGIR*, pages 833–834, 2010.

[14] J. He, E. Meij, and M. de Rijke. Result diversification based on Query-specific cluster ranking. *JASIST*, 62(3):550–571, 2011.

[15] O. Kurland. Re-ranking search results using language models of query-specific clusters. *Journal of Information Retrieval*, 12(4):437–460, 2009.

[16] O. Kurland. The cluster hypothesis in information retrieval. In *SIGIR2013 tutorial*, 2013.

[17] O. Kurland and L. Lee. Clusters, language models, and ad hoc information retrieval. *ACM Transactions on Information Systems*, 27(3):13:1–13:39, 2009.

[18] J. H. Lee. Analyses of multiple evidence combination. In *Proceedings of the 20th SIGIR*, pages 267–276, 1997.

[19] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of the 27th SIGIR*, pages 186–193, 2004.

[20] X. Liu and W. B. Croft. Evaluating text representations for retrieval of the best group of documents. In *Proceedings of the 30th ECIR*, pages 454–462, 2008.

[21] A. F. T. Martins, N. A. Smith, and E. P. Xing. Concise integer linear programming formulations for dependency parsing. In *Proceedings of the 47th ACL*, pages 342–350, 2009.

[22] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th ICML*, pages 784–791, 2008.

[23] F. Raiber and O. Kurland. Ranking document clusters using markov random fields. In *Proceedings of the 36th SIGIR*, pages 333–342, 2013.

[24] C. J. V. Rijsbergen. *Information Retrieval*. 2nd edition, 1979.

[25] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Proceedings of TREC*, 1994.

[26] D. Roth and W. Yih. Integer linear programming inference for conditional random fields. In *Proceedings of the 22nd ICML*, pages 736–743, 2005.

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

[28] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th WWW*, pages 881–890, 2010.

[29] R. L. T. Santos, C. Macdonald, and I. Ounis. Search result diversification. *Foundations and Trends in*

Information Retrieval, 9(1):1–90, 2015.

[30] J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd SIGIR*, pages 115–122, 2009.

[31] Y. Wang and L. Chen. K-MEAP: multiple exemplars affinity propagation with specified K clusters. *IEEE Transactions on Neural Networks and Learning Systems*, PP(99):1–13, 2015.

[32] K. Woodsend and M. Lapata. Multiple aspect summarization using integer linear programming. In *EMNLP-CoNLL2012*, pages 233–243, 2012.

[33] L. Xia, J. Xu, Y. Lan, J. Guo, and X. Cheng. Learning maximal marginal relevance model via directly optimizing diversity evaluation measures. In *Proceedings of the 38th SIGIR*, pages 113–122, 2015.

[34] J. Xiao, J. Wang, P. Tan, and L. Quan. Joint affinity propagation for multiple view segmentation. In *Proceedings of the 11th ICCV*, pages 1–7, 2007.

[35] H. Yu and F. Ren. Search result diversification via filling up multiple knapsacks. In *Proceedings of the 23rd CIKM*, pages 609–618, 2014.

[36] J. Yuan, F. Gao, Q. Ho, W. Dai, J. Wei, X. Zheng, E. P. Xing, T. Liu, and W. Ma. LightLDA: big topic models on modest computer clusters. In *Proceedings of the 24th WWW*, pages 1351–1361, 2015.

[37] Y. Yue and T. Joachims. Predicting diverse subsets using structural SVMs. In *Proceedings of the 25th ICML*, pages 1224–1231, 2008.

[38] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

[39] G. Zuccon and L. Azzopardi. Using the quantum probability ranking principle to rank interdependent documents. In *Proceedings of the 32nd ECIR*, pages 357–369, 2010.

[40] G. Zuccon, L. Azzopardi, D. Zhang, and J. Wang. Top-k retrieval using facility location analysis. In *Proceedings of the 34th ECIR*, pages 305–316, 2012.