Sparse Spatial Selection for Novelty-based Search Result Diversification

Veronica Gil-Costa[1], Rodrygo L.T. Santos[2],
Craig Macdonald[2], and Iadh Ounis[2]

[1] Universidad Complutense de Madrid, Spain and Yahoo! Research Latin America
gvcosta@yahoo-inc.com
[2] University of Glasgow, UK
{rodrygo,craigm,ounis}@dcs.gla.ac.uk

Abstract. Novelty-based diversification approaches aim to produce a diverse ranking by directly comparing the retrieved documents. However, since such approaches are typically greedy, they require $O(n^2)$ document-document comparisons in order to diversify a ranking of n documents. In this work, we propose to model novelty-based diversification as a similarity search in a sparse metric space. In particular, we exploit the triangle inequality property of metric spaces in order to drastically reduce the number of required document-document comparisons. Thorough experiments using three TREC test collections show that our approach is at least as effective as existing novelty-based diversification approaches, while improving their efficiency by an order of magnitude.

1 Introduction

Search result diversification has emerged as an effective approach for tackling ambiguous queries. In particular, a diverse ranking aims to satisfy as many *aspects* of an ambiguous query as possible, and as early as possible. By satisfying multiple query aspects, a high *coverage* of these aspects is achieved. By having different aspects satisfied as early as possible, a high *novelty* is also attained [10].

Promoting coverage is typically more efficient than promoting novelty: while coverage can be estimated for different documents independently, the same is not true for novelty. In particular, the notion of novelty entails a dependence between the relevance of different documents—i.e., a novel document is one that covers aspects not covered by the other documents. As a result, novelty-based diversification becomes essentially the problem of finding a set of documents that together cover most of the aspects of a query at a given rank cutoff. In this general formulation, this is an NP-hard problem [1]. Most of the approaches proposed in the literature for this problem deploy a greedy approximation algorithm: at each iteration, the algorithm selects a document that covers the most aspects not yet covered by the documents selected in the previous iterations. In a typical case, after the system retrieves n documents to be diversified, this greedy algorithm performs $O(n^2)$ document-document comparisons—i.e., $O(n)$ similarity searches across n iterations [7]—which can severely impact the efficiency of these approaches.

In this paper, we propose to reduce the number of required similarity computations in novelty-based diversification approaches, by modelling novelty in a metric space [7]. Metric spaces have been used for similarity search in many modern database applications [2]. Similarity search in metric spaces focuses on retrieving objects which are similar to a query point, with a metric distance function measuring the objects' similarity. By representing the retrieved documents as m-dimensional vectors in a metric space, we can exploit the triangle inequality property of such spaces to dramatically reduce the number of similarity computations required to diversify these documents.

A number of metric space search algorithms have been proposed in the literature (e.g., [3, 16–18]). In this paper, we show that effective and efficient diversification can be obtained using a sparse spatial selection approach [3], which selects pivot documents from the result set at running time, in order to reduce the number of required similarity computations. Although metric spaces have been used for image search diversification [14], to the best of our knowledge, our approach is the first attempt to leverage the properties of such spaces for diversifying textual documents. Moreover, while the data structure used in [14] has a $O(n^2)$ construction time [15], the one used in this paper can be built in linear time.

The contributions of this paper are two-fold: (1) we propose to model novelty-based diversification as a sparse spatial selection over a metric space; (2) we thoroughly investigate the effectiveness and efficiency of our proposed approach, using three standard TREC test collections for diversity evaluation. Our experimental results attest both the effectiveness and the efficiency of our approach compared to existing novelty-based diversification approaches.

In Section 2, we review existing approaches for search result diversification and similarity search in metric spaces. Section 3 shows how novelty-based diversification can be modelled in a metric space. Sections 4 and 5 detail our experimental setup and evaluation, respectively. Conclusions follow in Section 6.

2 Background and Related Work

In this section, we review existing approaches to search result diversification (Section 2.1) and similarity search in metric spaces (Section 2.2).

2.1 Search Result Diversification

Diversification approaches can be broadly classified as *implicit* or *explicit*. Implicit approaches assume that different documents will cover different query aspects. As a result, these approaches promote novel documents as a means to indirectly cover multiple aspects. The definition of a 'novel' document is precisely what distinguishes the approaches in this family. For instance, Carbonell and Goldstein [4] proposed to compare documents based on their cosine similarity. Zhai et al. [23] proposed an extension of this idea, by comparing documents with respect to the divergence of their language models. More recently, Wang and Zhu [22] proposed to use the correlation of documents' relevance scores.

Instead of assuming that different documents cover different aspects, explicit diversification approaches directly model these aspects as part of their strategy. For instance, Agrawal et al. [1] proposed a diversification approach based on an explicit representation of query aspects as taxonomy classes, in order to promote documents that cover classes also covered by the query. A similar approach was proposed by Carterette and Chandar [5], but with query aspects represented as topic models built from the top retrieved results for the query. Finally, Santos et al. [20] proposed to represent the aspects underlying a query as 'sub-queries'. In their approach, documents are promoted according to their estimated relevance to multiple sub-queries, as well as to the estimated importance of each sub-query.

Although having the same goal, these two families of approaches deploy rather distinct strategies. While implicit diversification approaches are driven by novelty, explicit ones usually target coverage. Since coverage can be estimated independently for different documents, explicit approaches are generally more efficient than implicit ones. In this paper, we propose to reduce the overhead incurred by document-document comparisons in novelty-based diversification approaches. In particular, we model novelty seeking within a metric space, and exploit the properties of this space to efficiently identify novel documents.

2.2 Search in Metric Spaces

Metric spaces are useful to represent complex data objects, such as documents or images, in a searchable collection. Search queries are represented as objects of the same type as the objects in the collection wherein, for example, one is interested in retrieving the most similar objects to the query. Formally, a *metric space* $(\mathcal{U}, \delta)$ comprises a universe of objects $\mathcal{U}$ and a *distance function* $\delta : \mathcal{U} \times \mathcal{U} \rightarrow \mathcal{R}^+$, which determines the similarity between any pair of objects [7]. The definition of the distance function depends on the type of the objects being compared. In an m-dimensional vector space—a particular case of metric spaces in which every object is represented by a vector of m real coordinates—δ could be a distance function of the family $L_s(x, y) = (\sum_{1 \leq i \leq m} |x_i - y_i|^s)^{\frac{1}{s}}$. For example, $s = 2$ yields the Euclidean distance. For any $x, y, z \in \mathcal{U}$, the function δ holds several properties: non-negativity ($\delta(x, y) \geq 0$), reflexivity ($\delta(x, y) = 0$ iff $x = y$), symmetry ($\delta(x, y) = \delta(y, x)$), and the triangle inequality ($\delta(x, z) \leq \delta(x, y) + \delta(y, z)$). The latter property is of particular interest, as it can be used to improve efficiency by avoiding unnecessary similarity computations, as will be shown in Section 3.

The finite subset $\mathcal{X} \subseteq \mathcal{U}$, with $n = |\mathcal{X}|$, denotes the working set of objects where searches are performed (e.g., the top-n documents retrieved for a query). A type of similarity search of particular interest to this work involves *range queries* [15]. In this search type, the goal is to retrieve all objects within distance r to a query object q, where r denotes the *search range*. Fig. 1 shows how the triangle inequality property of metric spaces can be exploited to avoid unnecessary similarity computations for range queries. In particular, given the distance $\delta(x, y)$ between the objects x and y and the distance $\delta(q, x)$ between x and a query object q with search range r, we can avoid computing $\delta(q, y)$.

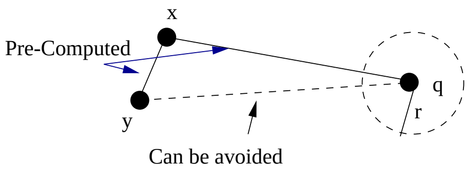


Fig. 1: The triangle inequality property. Once the distances $\delta(x, y)$ and $\delta(q, x)$ are computed, computing $\delta(q, y)$ for a query q with range r is unnecessary.

Most algorithms for similarity search in metric spaces fall into one of two categories: clustering and pivoting. Clustering techniques divide the working set of objects into groups (called clusters), such that similar objects fall into the same group [7]. Pivoting techniques select some objects as pivots, calculate the distance between every other object and each pivot, and apply the triangle inequality to avoid unnecessary similarity computations between objects. A key challenge for pivoting techniques is to determine the number of pivots needed to cover all objects in the working set. Moreover, the number of pivots tends to increase with the size of the working set. In the next section, we show how to adapt a pivot technique that avoids these problems in order to effectively and efficiently promote novel search results in the ranking.

3 Sparse Spatial Selection Diversification

Let $\mathcal{D}$ contain the documents initially retrieved for a query q. Existing diversification approaches typically re-score a document $d_i \in \mathcal{D}$ in light of the query q and the documents in $\mathcal{S} \subseteq \mathcal{D} \setminus \{d_i\}$, according to the following abstract model [21]:

$$\text{score}(q, d_i) = (1 - \lambda)\,\text{rel}(q, d_i) + \lambda\,\text{div}(q, d_i, \mathcal{S}), \tag{1}$$

where relevance $(\text{rel}(q, d_i))$ and diversity $(\text{div}(q, d_i, \mathcal{S}))$, as estimated by a given diversification approach, are traded off through the interpolation parameter λ.

In Equation (1), the relevance component $\text{rel}(q, d_i)$ can be estimated using any standard retrieval model. In a novelty-based diversification approach, the diversity component $\text{div}(q, d_i, \mathcal{S})$ is typically estimated in a greedy, iterative fashion. In particular, at any given iteration, every document $d_i \in \mathcal{D} \setminus \mathcal{S}$ is compared to every document $d_j \in \mathcal{S}$, where $\mathcal{S}$ comprises the documents selected in the previous iterations. This way, the document d_i that differs most from the already selected documents in $\mathcal{S}$ is itself included in $\mathcal{S}$. Such document-document comparisons are usually performed as distance computations in an m-dimensional term-frequency space, where m is the number of unique terms in the underlying document collection. As discussed in Section 2.1, these approaches differ mainly in their choice of a distance function (e.g., cosine [4], divergence [23], or correlation [22]). Regardless of the chosen distance function, these approaches require $O(n^2)$ distance computations to diversify a list of n documents. In particular, they perform an $O(n)$ similarity search across n iterations.

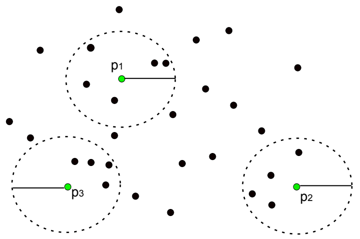


Fig. 2: Objects in the range of pivots p_1, p_2, and p_3 are considered redundant.

```
SSSD1$[q, \mathcal{D} = \{d_1, \ldots, d_n\}, \delta, M, \phi]$
1 $\mathcal{P} \Leftarrow \{d_1\}$
2 for all $d_i \in \mathcal{D} \setminus \{d_1\}$ do
3     if $\delta(d_i, p_j) \geq \phi M \ \forall p_j \in \mathcal{P}$ then
4         $\mathcal{P} \Leftarrow \mathcal{P} \cup \{d_i\}$
5     end if
6 end for
7 $\mathcal{P} \Leftarrow \mathcal{P} \cup (\mathcal{D} \setminus \mathcal{P})$
```

Alg. 1: Single-Step Sparse Spatial Selection Diversification (SSSD1).

In order to reduce the quadratic number of distance computations incurred by the existing greedy novelty-based diversification approaches, we propose to exploit a key property of metric spaces, namely, the triangle inequality. Our approach is based on an efficient pivoting similarity search algorithm. As illustrated in Fig. 2, the Sparse Spatial Selection (SSS) algorithm [3] identifies a set of k "pivots" among the n objects in the search space. By pre-computing the distances between the k pivots and the n objects, the number of subsequent distance computations can be drastically reduced. For instance, suppose we want to find all objects similar to an object x within a range r. If, for some pivot p, it holds that $|\delta(x, p) - \delta(y, p)| > r$, then we know, from the triangle inequality, that $\delta(x, y) > r$. Therefore, we do not need to explicitly evaluate $\delta(x, y)$.

In this paper, we propose a novelty-based diversification approach inspired by the SSS pivoting search algorithm. Our novel Sparse Spatial Selection Diversification (SSSD) approach incorporates the notion of pivots to reduce the number of distance computations required to diversify a set of documents. In particular, we develop two variants of SSSD. Our first variant, SSSD1, builds upon the SSS algorithm to skip redundant documents in the ranking. As described in Alg. 1, SSSD1 takes as input a query q, an initial set of documents $\mathcal{D}$ retrieved for this query, a distance function δ with upper-bound M, and the search radius ϕ, with $0 \leq \phi \leq 1$, such that $r = \phi M$ determines the search range of each pivot.

The core of SSSD1 is the selection of pivots (lines 1-6 in Alg. 1). To this end, let $(\mathcal{U}, \delta)$ be a metric space, with $\mathcal{D} \subseteq \mathcal{U}$ comprising the documents retrieved for the query q. The pivot set $\mathcal{P}$ is initialised with the first retrieved document, i.e., $d_1 \in \mathcal{D}$. For each remaining document $d_i \in \mathcal{D} \setminus \{d_1\}$, d_i is chosen as a new pivot if its distance to every pivot in $\mathcal{P}$ is greater than or equal to the search range ϕM. Hence, a retrieved document becomes a new pivot if and only if it is located outside the search range of all current pivots. Moreover, documents within the search range of an already selected pivot are considered redundant and skipped, and are later added to the bottom of the ranking (line 7), in the same order as they were originally retrieved in the initial ranking $\mathcal{D}$.

Importantly, during the selection of pivots, it is not necessary that all documents in $\mathcal{D}$ be compared against all pivots. When a document d_i is compared against a pivot p_j and does not satisfy the condition $\delta(d_i, p_j) \geq \phi M$, this document is discarded and no additional comparisons are required. In the best case scenario, documents are compared only with the first pivot when they are within the range of this pivot. Assuming that an unseen document d_i has a constant probability $\nu = f(\mathcal{U}, \delta, M, \phi)$ of lying outside the range of all pivots $p_j \in \mathcal{P}$ given the metric space $(\mathcal{U}, \delta)$ and the search range ϕM, it can easily be shown that Alg. 1 requires $\sum_{i=1}^{n-1} \nu^{i-1}(n-i)$ document-pivot comparisons to diversify n documents. In the worst case, when $\nu = 1$ (i.e., all documents are outside the range of all pivots and become themselves pivots), this algorithm exhibits the same quadratic complexity as the greedy novelty seeking approach. However, in practical deployments, $\nu < 1$, which results in a drastic reduction in the number of required document-pivot comparisons, as we will show in Section 5.

SSSD1 promotes novelty in an iterative fashion, by prospecting new pivots from the ranking $\mathcal{D}$. Although respecting the order that the documents were originally retrieved for the query q, this variant does not perform any explicit re-scoring of these documents, and in fact treats the non-pivot documents indistinctly. To investigate whether a more fine-grained re-ranking could be beneficial, we propose a second variant of our approach. In particular, the SSSD2 variant extends SSSD1, by performing a second step over the retrieved documents, in order to assign each document d_i a score with respect to the query, in light of the abstract diversification model defined in Equation (1). This variant is described in Alg. 2. SSSD2 is essentially equivalent to SSSD1, except for the introduced scoring step (lines 7-9 in Alg. 2), and the additional parameter β. In particular, this introduced step scores a document d_i as a linear combination of its estimated relevance to the query and its estimated novelty, as given by the inverse of the distance between d_i and its most similar pivot $p_j \in \mathcal{P}$. A balance between relevance and novelty is achieved through an appropriate setting of β. In a naive implementation, this step takes additional $O(kn)$ document-document comparisons, which is already substantially more efficient than existing novelty-based diversification approaches, if $k \ll n$. However, even when this is not the case—particularly for high-dimensional spaces—the second step can reuse all the comparisons performed for the pivot selection in the first step. Therefore, as we will show in Section 5, in contrast to SSSD1, SSSD2 can deploy a traditional diversification scoring scheme, at the cost of typically only a few additional comparisons.

```
SSSD2[$q, \mathcal{D} = \{d_1, \ldots, d_n\}, \delta, M, \phi, \beta$]
1 $\mathcal{P} \Leftarrow \{d_1\}$
2 for all $d_i \in \mathcal{D} \setminus \{d_1\}$ do
3     if $\delta(d_i, p_j) \geq \phi M \; \forall p_j \in \mathcal{P}$ then
4         $\mathcal{P} \Leftarrow \mathcal{P} \cup \{d_i\}$
5     end if
6 end for
7 for all $d_i \in \mathcal{D}$ do
8     $\text{score}(q, d_i) = (1 - \beta)\,\text{rel}(q, d_i) + \beta\left[1 - \max_{p_j \in \mathcal{P}} \delta(d_i, p_j)\right]$
9 end for
```

Alg. 2: Two-Step Sparse Spatial Selection Diversification (SSSD2).

4 Experimental Setup

Our investigation aims to answer two major research questions:

1. How do SSSD and existing approaches compare in terms of effectiveness?
2. How do SSSD and existing approaches compare in terms of efficiency?

To evaluate our approach in different metric spaces, we experiment with three test collections for diversity evaluation, comprising both web and newswire documents. The first two are from the diversity task of the TREC 2009 and 2010 Web tracks [8, 9]—henceforth WT09 and WT10, respectively. WT09 includes 50 topics, while WT10 comprises 48 topics. Our third collection includes 20 topics from the Interactive track of TREC-6, TREC-7, and TREC-8 [12]—henceforth IT678. For WT09 and WT10, we index the TREC ClueWeb09 (cat. B) corpus, with 50 million web documents. For IT678, we index the Financial Times portion of TREC Disks 4&5, with 210,000 newswire documents. Both corpora are indexed using Terrier [19], with Porter's stemmer and standard stopword removal.

To retrieve an initial pool of documents to be diversified, we apply either BM25 or the Divergence from Randomness DPH model, as implemented in Terrier. On top of these adhoc retrieval baselines, we deploy two well-known novelty-based diversification approaches as diversification baselines: Maximal Marginal Relevance (MMR [4]) and Mean-Variance Analysis (MVA [22]). As these approaches compute novelty based on cosine or correlation estimations, respectively, we deploy both variants of our approach using both cosine and Pearson's correlation as instantiations of the distance function δ. To cope with the quadratic complexity of MMR and MVA while keeping a uniform setting across all approaches, both of these baselines as well as our SSSD variants are applied to diversify the top 100 documents retrieved by BM25 or DPH.

Effectiveness is assessed using the primary metrics in the diversity task of the TREC 2010 Web track, namely, ERR-IA [6] and α-nDCG [10]. To train the parameters of our approach (ϕ for SSSD1 and both ϕ and β for SSSD2), as well as the parameters for MMR (λ [4]) and MVA (σ and b [22]), we perform a simulated annealing [13] through a 5-fold cross validation. In particular, we train the parameters of all approaches to maximise α-nDCG@100 on the training folds, and report the results as an average across the corresponding separate test folds. As for efficiency, we report the number of document-document comparisons performed, as well as the time spent in performing such comparisons.

5 Experimental Results

In this section, we investigate whether novelty-based diversification approaches can be made efficient without compromising their effectiveness. Before investigating the efficiency of SSSD, we evaluate its effectiveness compared to MMR [4] and MVA [22] as baselines. Table 1 shows the diversification performance of both SSSD variants as well as these two baselines across the WT09, WT10, and IT678 settings. As the distance function δ, we consider both cosine (denoted c) and Pearson's correlation (denoted ρ). All approaches are applied on top of both

Table 1: Diversification performance across the WT09, WT10, and IT678 topics.

	WT09		WT10		IT678	
	ERR-IA @20	α-nDCG @20	ERR-IA @20	α-nDCG @20	ERR-IA @20	α-nDCG @20
BM25	0.1304	0.2290	0.1628	0.2349	0.1541	0.4703
+MMR	0.1341	0.2366	0.1652	0.2379	0.1573	**0.4806**
+MVA	0.1336	0.2369	0.1654	0.2343	0.1547	0.4708
+SSSD1(c)	**0.1429**$^{\triangle}$	**0.2526**$^{\triangle}$	0.1688$^{\triangle}$	**0.2447**	**0.1600**	0.4764
+SSSD1(ρ)	0.1242	0.2234	0.1585	0.2324	0.1500	0.4577
+SSSD2(c)	0.1237	0.2178	0.1628	0.2356	0.1483	0.4481
+SSSD2(ρ)	0.1279	0.2248	**0.1695**	0.2402	0.1532	0.4662
DPH	0.1430	0.2426	0.1952	0.2977	0.1658	0.4833
+MMR	0.1378	0.2363	0.1963	0.2889	0.1652	**0.4842**
+MVA	0.1314	0.2203	0.1908	0.2841	0.1636	0.4674
+SSSD1(c)	0.1474$^{\blacktriangle}$	0.2608$^{\blacktriangle}$	0.1952	**0.2981**	0.1620	0.4689
+SSSD1(ρ)	0.1333	0.2266	**0.1973**	0.2977	**0.1678**	0.4831
+SSSD2(c)	0.1344	0.2367	0.1944	0.2945	0.1639	0.4807
+SSSD2(ρ)	**0.1637**	**0.2646**$^{\triangle}$	0.1847	0.2796	0.1518	0.4692

BM25 and DPH. Significance between both SSSD1 and SSSD2 and the best between MMR and MVA is verified with the Wilcoxon matched-pairs test. In particular, the symbols $\triangle$ and $\triangledown$ denote a significant increase or decrease with $p < 0.05$, while $\blacktriangle$ and $\blacktriangledown$ denote significant increases or decreases with $p < 0.01$.

From Table 1, we note that both SSSD1 and SSSD2 can improve over MMR and MVA across several settings. Such improvements are significant for SSSD1(c) using BM25 (for WT09 and WT10) and DPH (for WT09), and for SSSD2(ρ) using DPH (for WT09). In all other cases, there is no significant difference between these approaches. This answers our first question, by showing that SSSD performs *at least as effectively* as existing novelty-based approaches. As for distance functions, when the initial ranking is given by BM25, cosine gives superior results for SSSD1, while Pearson's correlation is the most effective function for SSSD2. When DPH provides the initial ranking, there is no consistently best choice of function. As for the two SSSD variants themselves, SSSD1 performs generally better than SSSD2 (except for IT678 using DPH) when cosine is fixed as the distance function. For Pearson's correlation, SSSD2 is generally the best of the two variants for BM25, while SSSD1 is generally best for DPH. Overall, these results show that the choice of an SSSD variant depends on the considered metric space, as determined by the target test collection and the chosen distance function.

To answer our second question, we investigate how the pivot selection impacts the efficiency of our approach. In particular, the number of selected pivots is a function of both the dimensionality of the search space and the search radius ϕ. Hence, we analyse the efficiency of SSSD1 and SSSD2 over a range of ϕ values, as well as over the search spaces of the three considered test collections. For the WT09, WT10, and IT678 collections, Fig. 3 shows how the number of selected pivots (Figs. 3(a)-(c)), the number of document-document comparisons

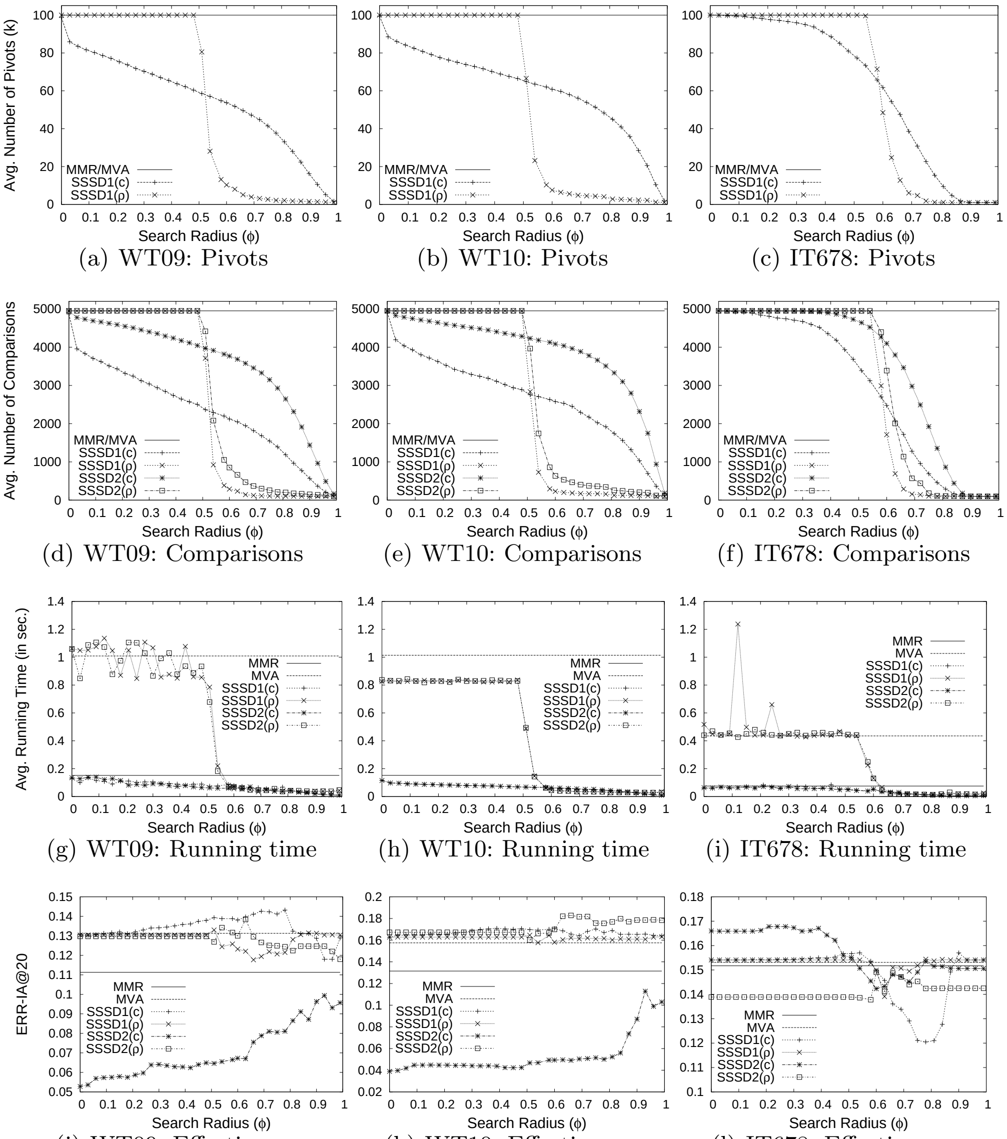
(Figs. 3(d)-(f)), and the running time[3] (Figs. 3(g)-(i)) of our approach are affected by the parameter ϕ. Additionally, to enable the analysis of efficiency in context, Figs. 3(j)-(l) show how ϕ impacts the effectiveness of our approach.

From Figs. 3(a)-(c), we first observe, as expected, that the number of pivots selected by SSSD1[4] decreases as ϕ increases, since the area covered by each pivot increases. However, while the number of selected pivots decreases smoothly for SSSD1(c), a more abrupt drop is observed for SSSD1(ρ), with an inflection around $\phi = 0.5$ for WT09 and WT10, and $\phi = 0.6$ for IT678. This suggests that correlation is more sensitive than cosine as a distance function. In particular, in such sparse spaces as those considered here, documents which share only a few but highly informative terms can exhibit negligible correlations, while still having a noticeable cosine. Next, we assess how ϕ (and consequently, the number pivots) impacts the number of comparisons and the running time of our approach.

Contrasting Figs. 3(a)-(c) and (d)-(f), we observe a similar shape between the number of selected pivots and that of performed comparisons. Indeed, there is an almost perfect linear correlation between the number of comparisons and of selected pivots (WT09: 0.993 for SSSD1(c), 0.998 for SSSD1(ρ); WT10: 0.993 for SSSD1(c), 0.999 for SSSD1(ρ); IT678: 0.992 for SSSD1(c), 0.997 for SSSD1(ρ)). This provides empirical evidence that SSSD1 has an average-case complexity of $O(k)$, where $k = O(n)$ is the number of selected pivots. Moreover, when SSSD2 is considered, only a constant number of additional comparisons is performed, hence leaving the asymptotic cost unchanged. In practice, this shows that our approach is *an order of magnitude faster* compared to the quadratic number of comparisons performed by both MMR and MVA (precisely, $n(n-1)/2$ comparisons), hence answering our second research question. This observation is further confirmed by Figs. 3(g)-(i), which show the running time of our approach, compared to both MMR and MVA, for a range of ϕ values, and averaged across the WT09, WT10, and IT678 topics, respectively. Although dominated by the number of comparisons, these figures exemplify another facet of the time complexity of all novelty-based approaches, namely, the unitary cost of a comparison. Indeed, computing the cosine between two documents is cheaper than computing their correlation, even though both are optimised to exploit the sparsity of the considered spaces. Nonetheless, the variants of SSSD using these distance functions are faster than MMR (which uses cosine) and MVA (which uses correlation), respectively, across the entire range of ϕ values, and for the three considered collections. To further test these approaches over a representative query stream, we select the first 1,000 queries from the MSN 2006 query log [11], after removing empty queries and queries with no results in the ClueWeb09 corpus. Figs. 4(a) and (b) show the results of this investigation in terms of number of comparisons and running time, respectively. These results closely match those shown in Figs. 3(d)-(f) and (g)-(i), hence further attesting the efficiency of our approach.

[3] Running times are based on a Linux Quad-Core Intel Xeon 2.4GHz 8GB, and denote the time spent to compare documents, as the cost to retrieve the initial documents and represent these documents in a vector space is the same for all approaches.

[4] SSSD2 uses the same pivot selection as SSSD1, and is hence omitted from the figures.



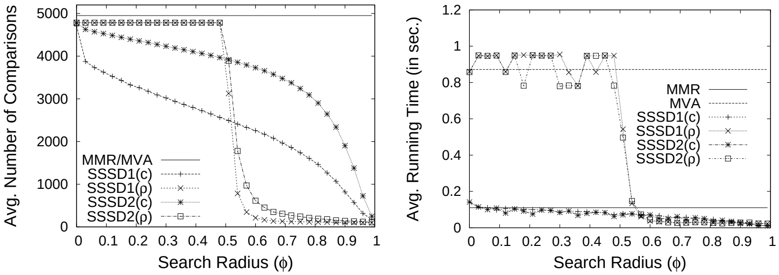
(a) WT09: Pivots (b) WT10: Pivots (c) IT678: Pivots

(d) WT09: Comparisons (e) WT10: Comparisons (f) IT678: Comparisons

(g) WT09: Running time (h) WT10: Running time (i) IT678: Running time

(j) WT09: Effectiveness (k) WT10: Effectiveness (l) IT678: Effectiveness

Fig. 3: Number of pivots, number of document-document comparisons, running time, and diversification performance for the WT09, WT10, and IT678 test collections (left, middle, and right columns, respectively), across a range of ϕ values. All figures are averages across the topics of the corresponding collection (50, 48, and 20, respectively).



(a) MSN06: Comparisons (b) MSN06: Running time

Fig. 4: Number of document-document comparisons and running time across a range of ϕ values. All figures are averages over 1000 queries from the MSN 2006 query log.

Lastly, Figs. 3(j)-(l) bridge our two research questions, by showing the impact of increasing the search radius ϕ on the effectiveness of both variants of SSSD, in terms of ERR-IA@20. In general, we observe two distinct behaviours. Firstly, a steady improvement is observed for SSSD1(c) for WT09 (up to $\phi \approx 0.8$) and SSSD2(c) for both WT09 and WT10. With a higher search radius ϕ, these variants perform a more aggressive diversification, by creating fewer pivots and considering more documents as redundant. Secondly, a dual impact is observed for the SSSD2(ρ) variant for $\phi > 0.5$, which coincides with the inflection point in Figs. 3(a)-(c). In particular, while the effectiveness of SSSD2(ρ) decreases after this point for WT09, it increases for WT10. Likewise, a region of instability is observed for other variants after the inflection point (i.e., $0.5 \leq \phi \leq 0.9$. This is the case for SSSD1(ρ) for WT09 and IT678, and for SSSD1(c), SSSD2(c), and SSSD2(ρ) for IT678. Overall, these results show that even documents of a similar nature (e.g., web pages) can result in rather different spaces. Hence, carefully choosing a search radius for the test collection at hand is key for attaining a suitable trade-off between an effective and efficient diversification.

6 Conclusions

We have introduced a new approach for novelty-based search result diversification, by exploiting the properties of metric spaces. Our Sparse Spatial Selection Diversification (SSSD) approach selects a set of pivots from the space of documents retrieved for a query, and leverages the triangle inequality property of metric spaces to regard documents covered by a pivot as redundant. As an extended variant, we further score the retrieved documents with respect to their distance to the selected pivots, in order to perform a more fine-grained re-ranking.

In a thorough investigation across three standard TREC test collections for diversity evaluation, we have shown that both variants of our approach (SSSD1 and SSSD2) perform at least as effectively as well-known novelty-based diversification approaches in the literature, while improving their efficiency by an order of magnitude. Moreover, by evaluating our approach across metric spaces induced by different document collections and distance functions, we have shown that a careful selection of pivots is paramount for appropriately trading-off effectiveness and efficient in novelty-based search result diversification.

Acknowledgements The work has been performed under the HPC-EUROPA2 project (project number: 228398) with the support of the European Commission–Capacities Area–Research Infrastructures.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM. pp. 5–14 (2009)
2. Barrios, J.M., Diaz-Espinoza, D., Bustos, B.: Text-based and content-based image retrieval on Flickr: DEMO. In: SISAP. pp. 156–157 (2009)
3. Brisaboa, N.R., Farina, A., Pedreira, O., Reyes, N.: Similarity search using sparse pivots for efficient multimedia information retrieval. In: ISM. pp. 881–888 (2006)
4. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. pp. 335–336 (1998)
5. Carterette, B., Chandar, P.: Probabilistic models of ranking novel documents for faceted topic retrieval. In: CIKM. pp. 1287–1296 (2009)
6. Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: CIKM. pp. 621–630 (2009)
7. Chávez, E., Navarro, G., Baeza-Yates, R., Marroquín, J.L.: Searching in metric spaces. ACM Comput. Surv. 33(3), 273–321 (2001)
8. Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 Web track. In: TREC (2009)
9. Clarke, C.L.A., Craswell, N., Soboroff, I., Cormack, G.V.: Preliminary overview of the TREC 2010 Web track. In: TREC (2010)
10. Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR. pp. 659–666 (2008)
11. Craswell, N., Jones, R., Dupret, G., Viegas, E. (eds.): Proceedings of the 2009 Workshop on Web Search Click Data (2009)
12. Hersh, W., Over, P.: TREC-8 Interactive track report. In: TREC (2000)
13. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science 220(4598), 671–680 (1983)
14. van Leuken, R.H., Garcia, L., Olivares, X., van Zwol, R.: Visual diversification of image search results. In: WWW. pp. 341–350 (2009)
15. Mamede, M., Barbosa, F.: Range queries in natural language dictionaries with recursive lists of clusters. In: ISCIS (2007)
16. Micó, L., Oncina, J., Carrasco, R.C.: A fast branch & bound nearest neighbour classifier in metric spaces. Pattern Recogn. Lett. 17(7), 731–739 (1996)
17. Navarro, G., Reyes, N.: Fully dynamic spatial approximation trees. In: SPIRE. pp. 254–270 (2002)
18. Navarro, G., Reyes, N.: Dynamic spatial approximation trees for massive data. In: SISAP. pp. 81–88 (2009)
19. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: a high performance and scalable information retrieval platform. In: OSIR (2006)
20. Santos, R.L.T., Macdonald, C., Ounis, I.: Exploiting query reformulations for Web search result diversification. In: WWW. pp. 881–890 (2010)
21. Santos, R.L.T., Macdonald, C., Ounis, I.: Selectively diversifying Web search results. In: CIKM (2010)
22. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: SIGIR. pp. 115–122 (2009)
23. Zhai, C., Cohen, W.W., Lafferty, J.: Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In: SIGIR. pp. 10–17 (2003)