# Efficient Diversification of Top-k Queries over Bounded Regions*

Piero Fraternali, Davide Martinenghi, and Marco Tagliasacchi

Dipartimento di Elettronica e Informazione – Politecnico di Milano
Piazza Leonardo da Vinci, 32 – 20133 Milano, Italy
{piero.fraternali,davide.martinenghi,marco.tagliasacchi}@polimi.it

**Abstract.** This paper reports on recent findings regarding diversity queries over objects embedded in a low-dimensional vector space. Among the many contexts of interest, we mention spatial Web objects, which are abundant in location-based services that let users attach content to places. Typical queries aim at retrieving the best set of relevant objects that are well distributed over a region of interest. Existing methods for answering diversified top-$k$ queries are too costly, as they evaluate diversity by accessing and scanning all relevant objects, even if only a small subset thereof is needed. Our proposal, named SPP, is an algorithm that, while finding exactly the same result as MMR (one of the most popular diversification algorithms), does not require retrieving all the relevant objects and, indeed, minimizes the number of accessed objects. Experiments confirm that SPP saves a significant amount of accesses while incurring a very low computational overhead.

## 1 Introduction

Geo-referenced data are becoming more and more available on the Web, especially after the advent of location-based services, whereby users can create content attached to places. Web spatial objects are also found in real estate directories, local news aggregators, image sharing sites, and travel services. Queries that require the uniform coverage of a region through spatial scattering of results are very common. An example is that of a user moving to a new city who wants an overview of real estate offers that meet some relevance criterion (e.g., price) and cover most neighborhoods.

In this paper, we address the problem of answering *top-$k$ diversity queries* over online data sources *covering a region of interest* by presenting a synthesis of the results described in [6]. We assume that objects are represented in a vector space and can be fetched through interfaces, common for Web data sources, granting *sorted* access either by relevance (e.g., an object property or the degree of match with the query) or by distance from a given point. Our objective is to improve performance of diversified query processing by accessing only a small number of objects that guarantee to find the best result set in terms of both relevance and diversity. This is in contrast with classical diversification techniques, which access all the objects first, and then choose the best subset of diversified objects. As an example, consider a real estate query: a sample search in a commercial service for flats in London between £200,000 and £300,000 returned 60,000+ results; if the user wants to browse just a few dozens of them in diverse neighborhoods, we should access and present a number of objects proportional

* Extended Abstract

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

to the user's wishes, scattered throughout the London region, without accessing all the 60,000+ relevant flats. In addition, we rely solely on the presence of sorted access methods based on relevance and distance, without requiring the knowledge of the specific index structures being used, as these typically reside on remote third-party services.

Top-$k$ diversity queries over a vector space require a mix of techniques from top-$k$ query processing and result diversification. As in top-$k$ query processing [9, 11], the cost model of access methods requires minimizing the number of fetched objects. To this end, a threshold (upper bound) on the value of an objective function that quantifies both score and diversity is maintained. As more candidate objects are accessed via *probe queries*, the upper bound decreases until the guarantee is reached that no unseen object can lead to a value of the objective function better than the one determined by the already retrieved objects. Unlike in top-$k$ query processing, the objective function of diversity queries cannot be computed on individual objects, as it uses a diversity measure (e.g., spatial scatter) that requires comparing the next object with the previously ranked ones. Existing diversification methods [2, 7] solve the issue by comparing *all* the relevant objects (i.e., the results of the user query) with the objects that have been top-ranked so far, thus materializing and scanning all of them several times [2].

We propose a novel approach, which integrates the notion of probe queries into the framework of result diversification, providing on-the-fly construction of top-$k$ result sets that are both relevant and diverse, by using only sorted access methods and without fetching all relevant objects. Our approach works as follows. The top-$k$ set is built incrementally, adding each time the object that maximizes an objective function based on both relevance and diversity. At each step, we probe the vector space by issuing distance-based queries at suitable points, called *probing locations*, that are likely to lie close to the best objects. We may alternatively use score-based access to retrieve objects with high relevance. Based on the retrieved objects, we maintain an upper bound on the value of the objective function that can be attained by using the unseen objects. The currently best object is added to the set when the value of the objective function determined by its inclusion is at least as high as the upper bound. Note that the choice of probing location and the alternation of score-based and distance-based access (akin to a pulling strategy [11]) exploits the geometry of the vector space. The proposed approach introduces efficiency without compromising the quality of diversification wrt. the best known general-purpose diversification algorithms.

## 2 Preliminaries

Consider a query $q$ selecting a finite set $\mathcal{O}$ of objects. The *relevance* of an object $o \in \mathcal{O}$ to $q$ is represented by a *score* $S_q(o) \in \mathbb{R}$. Let each object $o$ also be associated with a real-valued feature vector $\mathbf{x}(o) \in \mathbb{R}^d$, denoting characteristics of the object that can be used to compute diversity. Then, *diversity* of two objects is expressed by a measure $\delta : \mathcal{O} \times \mathcal{O} \to \mathbb{R}^+$, where the value 0 indicates maximum similarity.

Let $N = |\mathcal{O}|$ denote the cardinality of the set $\mathcal{O}$, and $\mathcal{O}_K \subseteq \mathcal{O}$ a subset of $K$ objects that are selected, e.g., to be presented to the user. We are interested in identifying a subset $\mathcal{O}_K$ that is both relevant and diverse. The diversification problem, i.e., computing the best subset $\mathcal{O}_K^*$ of $\mathcal{O}$, can be expressed as an optimization problem as follows [7]:

$$\mathcal{O}_K^* = \underset{\mathcal{O}_K \subseteq \mathcal{O}, |\mathcal{O}_K| = K}{\operatorname{argmax}} F(\mathcal{O}_K; S_q(\cdot), \delta(\cdot,\cdot)) \tag{1}$$

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

**Algorithm 1:** MMR algorithm

Input: *Set of objects $\mathcal{O}$; result size $K$*
Output: *Selection $\mathcal{O}_K$ from $\mathcal{O}$*
Parameters: *Initialization Strategy* IS
1. $\mathcal{O}_K := \{\texttt{IS.initialObject}()\}$;
2. **while** $(|\mathcal{O}_K| < K)$
3. $\quad o^* := \underset{o \in \mathcal{O} \setminus \mathcal{O}_K}{\operatorname{argmax}} \{(1-\lambda) S_q(o) + \lambda \min_{o' \in \mathcal{O}_K} \delta(o, o')\}$;
4. $\quad \mathcal{O}_K := \mathcal{O}_K \cup \{o^*\}$;
5. **return** $\mathcal{O}_K$;

where $F(\cdot)$ is an objective function that takes into account both relevance and diversity. Solving problem (1) is NP-hard [7] for various objective functions. Hence, the need for approximate greedy algorithms, among which Maximum Marginal Relevance (MMR) [2], is one of the most popular. MMR implicitly adopts the following $F(\cdot)$:

$$F(\mathcal{O}_K) = (1-\lambda) \sum_{o \in \mathcal{O}_K} S_q(o) + \lambda \min_{o_u, o_v \in \mathcal{O}_K} \delta(o_u, o_v) \tag{2}$$

where $\lambda$ is a parameter in $[0, 1]$ specifying the trade-off between relevance and diversity.

Algorithm 1 illustrates the details of MMR, which incrementally constructs $\mathcal{O}_K$ by adding, at each step, an object that is both relevant and distant from the already selected objects. The initial object is chosen according to some initialization strategy IS; typically, the object that maximizes $S_q(\cdot)$ is selected. The added object $o^*$ (line 3 of Algorithm 1) maximizes the *diversity-weighted score* $\sigma$, defined as:

$$\sigma(o; \mathcal{O}_K) = (1-\lambda) S_q(o) + \lambda \min_{o' \in \mathcal{O}_K} \delta(o, o') \tag{3}$$

thereby maximizing also $F(\mathcal{O}_K \cup \{o\})$.

The algorithms proposed for optimizing the MMR objective function assume that all the $N$ objects relevant to the query are retrieved and re-ranked so as to select the top $K$ diversified elements. Therefore, all such algorithms exhibit a complexity that depends on $N$. For example, the overall time complexity of MMR is $O(K^2 N)$.

## 3 Bounded Diversification with Sorted Access Methods

The diversification problem addressed in this paper assumes that the feature vectors of the objects are contained in a finite boundary region. We consider two kinds of sorted access methods for fetching the objects:

– *Score-based access*. The set $\mathcal{O}$ is accessed sequentially in decreasing order of $S_q(\cdot)$, i.e., of relevance to the query.

– *Distance-based access*. The set $\mathcal{O}$ is accessed sequentially in increasing order of $\delta(\cdot, \mathbf{v})$, where $\mathbf{v}$ is an arbitrary vector in $\mathbb{R}^d$ called *probing location*. For example, objects are retrieved by geographical distance wrt. a given point.

Bounded diversification problems are diversification problems where the objects lie in a bounded region and can be accessed only by score or distance. We restrict our attention to the class of MMR-*correct* algorithms, defined as those deterministic algorithms

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

**Algorithm 2:** PBMMR$(K, \mathcal{U})$

Input: *result size $K$; bounding region $\mathcal{U}$*
Output: *Selection $\mathcal{O}_K$ from the objects enclosed in $\mathcal{U}$*
Main vars: *set $P$ of retrieved objects; discarded region $\mathcal{D}$; last score $S_q^{last}$; upper bound $\tau$; top diversity-weighted score $\sigma^*$; top object $o^*$*
Parameters: *Init. Strategy* IS*; Pulling Strategy* PS*; Bounding Scheme* BS

1. $\mathcal{O}_K := \{$IS.initialObject()$\}$;
2. $\mathcal{D} := \emptyset$; $S_q^{\text{last}} := 1$; $P := \mathcal{O}_K$;
3. **while** ($|\mathcal{O}_K| < K$ )
4. $\quad \tau := \infty$; $\sigma^* = -\infty$;
5. $\quad$ **if** ($P \setminus \mathcal{O}_K \neq \emptyset$) **then** $o^* = \underset{o \in P \setminus \mathcal{O}_K}{\operatorname{argmax}} \sigma(o; \mathcal{O}_K)$; $\sigma^* = \sigma(o^*; \mathcal{O}_K)$;
6. $\quad$ **while** ($\sigma^* < \tau$ **and** $\mathcal{D} \subset \mathcal{U}$ )
7. $\quad\quad m :=$ PS.chooseAccessMethod();
8. $\quad\quad o := m$.getNextObject();
9. $\quad\quad P := P \cup \{o\}$;
10. $\quad\quad$ **if** ($\sigma(o; \mathcal{O}_K) > \sigma^*$) **then** $\sigma^* = \sigma(o; \mathcal{O}_K)$; $o^* := o$;
11. $\quad\quad (\mathcal{D}, S_q^{\text{last}}) := m$.updateDiscardedRegionAndScore$(\mathcal{D}, S_q^{\text{last}})$;
12. $\quad\quad \tau :=$ BS.updateBound$(P, \mathcal{D}, S_q^{\text{last}})$;
13. $\quad \mathcal{O}_K := \mathcal{O}_K \cup \{o^*\}$;
14. **return** $\mathcal{O}_K$;

that select at each step exactly the same object as MMR, and thus output the same result $\mathcal{O}_K^{\text{MMR}}$. A family of algorithms solving this problem is shown in Algorithm 2, which we call *Pull/Bound* MMR (PBMMR), adapted from the *Pull/Bound Rank Join* template originally introduced for the rank join problem in [11]. The algorithm selects one object per outermost iteration. The selection is made by keeping track of an upper bound $\tau$ (computed via a *bounding scheme* BS) on the best diversity-weighted score attainable by visiting unseen objects, based on the region $\mathcal{D}$ of space already explored, the best score possible $S_q^{\text{last}}$, and the visited objects $P$. At each step of the exploration, the chooseAccessMethod function of a given *pulling strategy* PS decides the access method $m$ to use for retrieving the next object (score-based or distance-based), in the latter case also deciding which probing location to use, i.e., from which point in the vector space to start returning objects in increasing order of distance.

**Theorem 1.** PBMMR *is* MMR*-correct.*

## 4 Space Partitioning and Probing

### 4.1 Probing Locations

We start the illustration of the SPP algorithm by discussing the policy for determining the probing locations, i.e., the starting points used for distance-based access. Ideally, each time a distance-based access is made, one should explore the region of space that grants the highest chances to retrieve the object with the best diversity-weighted score. To this end, at each of the $K$ iterations of the algorithm (line 3), we fix the probing locations at the most promising points of the unexplored space. Then, we use these probing locations in the iterations of the inner loop (line 6), possibly querying the same location multiple times with an increased search radius.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

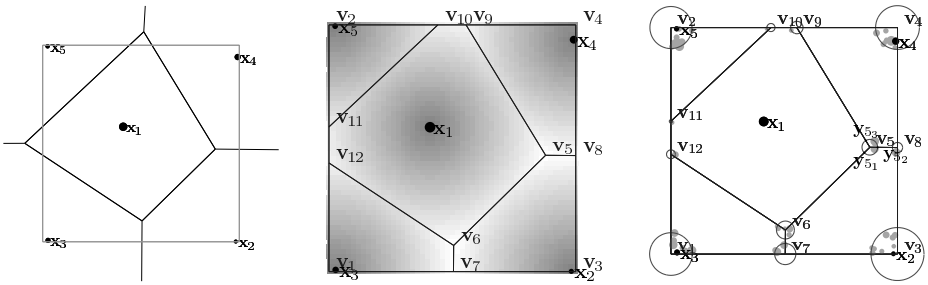


**Fig. 1.** Voronoi diagrams for Example 1.

At each of the $K$ main iterations, the most promising probing locations are points that lie within the bounding region $\mathcal{U}$ and are as far as possible from *all* the objects of the current selection $\mathcal{O}_\ell$. Let $\mathcal{X} = \{\mathbf{x}(o) \in \mathbb{R}^d | o \in \mathcal{O}_\ell\}$ denote the set of points corresponding to the current selection $\mathcal{O}_\ell$. Then, the probing locations can be defined as the local maxima of the function $f$ that expresses the distance of a point $\mathbf{x} \in \mathcal{U}$ from the closest object in the current selection: $f(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$.

An effective procedure for determining probing locations when $\mathcal{U}$ is a bounded polyhedron is provided by Theorem 2, which ensures that the local maxima of $f(\mathbf{x})$ lie in a subset of the vertices of the *bounded Voronoi diagram* $\text{Vor}(\mathcal{X},\mathcal{U})$ [3] of the points $\mathcal{X}$ corresponding to the current selection $\mathcal{O}_\ell$, obtained by restricting the conventional Voronoi diagram $\text{Vor}(\mathcal{X})$ to the region $\mathcal{U}$.

**Theorem 2.** *If $\mathbf{x}^* \in \mathcal{U}$ is a local maximum of $f(\mathbf{x})$ then, $\mathbf{x}^*$ is a vertex of $\text{Vor}(\mathcal{X},\mathcal{U})$.*

Theorem 2 allows us to efficiently find a superset of the local maxima by constructing $\text{Vor}(\mathcal{X},\mathcal{U})$ and enumerating its vertices $\mathbf{v}_u$, $u = 1, \ldots, V$. Vertices that are not local maxima can be disregarded.

*Example 1.* The left graph of Figure 1 illustrates an example when $d = 2$ and $\ell = 5$ objects have already been determined within a bounding rectangle $\mathcal{U}$. The corresponding points $\mathbf{x}_1, \ldots, \mathbf{x}_5$ define the Voronoi diagram $\text{Vor}(\mathcal{X})$. Note that $\mathcal{C}_1$ of $\mathbf{x}_1$ is bounded, whereas all the other cells are unbounded. The corresponding bounded Voronoi diagram $\text{Vor}(\mathcal{X},\mathcal{U})$ is represented in the middle graph of Figure 1. Vertices $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ and $\mathbf{v}_4$ correspond to the original vertices of $\mathcal{U}$. Out of the four vertices of $\text{Vor}(\mathcal{X})$, only two are retained (i.e., $\mathbf{v}_5$ and $\mathbf{v}_6$), as the other ones are outside $\mathcal{U}$. The remaining vertices ($\mathbf{v}_7$ to $\mathbf{v}_{12}$) are due to intersections between $\text{Vor}(\mathcal{X})$ and the edges of $\mathcal{U}$. The shading indicates the distance from the closest point in $\mathcal{X}$, where brighter indicates larger distance. Such a distance is maximized at the vertices, as confirmed by Theorem 2.

### 4.2 Bounding scheme

We now turn to the computation of the upper bound $\tau$ in a given running state. To exemplify a running state, the right graph of Figure 1 shows the discarded region $\mathcal{D}$ as a set of hyperspheres (in red) enclosing the previously accessed objects (shown as light red

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

discs with sizes proportional to the scores). Note that Vor($\mathcal{X}$,$\mathcal{U}$) and the corresponding probing locations are updated each time a new selected object is added by PBMMR.

The unseen objects retrievable with the next distance-based access belong to the set $\mathcal{Z} = \mathcal{U} \setminus \mathcal{D}$, which leaves out each explored hypersphere $\Sigma_u$ centered in $\mathbf{v}_u$, $u = 1, \ldots, V$. Indeed, after an object at a distance $r_u$ is extracted from $\mathbf{v}_u$, no new object can lie closer than that to $\mathbf{v}_u$. A tight upper bound can be found as follows

$$\tau = (1 - \lambda) S_q^{\text{last}} + \lambda \max_{\mathbf{x} \in \mathcal{Z}} \min_{\mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\| \tag{4}$$

Theorem 3 provides an effective computation procedure for (4).

**Theorem 3.** *The point $\mathbf{x}^* \in \mathcal{Z}$ that maximizes the minimum distance from all the points in $\mathcal{X}$ is a vertex of the convex hull of $\mathcal{P}_i \setminus \mathcal{D}$, where $\mathcal{P}_i$ is one of the cells of Vor($\mathcal{X}$,$\mathcal{U}$).*

Thanks to Theorem 3, $\tau$ as of (4) can be computed by enumerating the cells of Vor($\mathcal{X}$,$\mathcal{U}$) and, for each cell diminished by $\mathcal{D}$, the vertices of its convex hull. Equivalently, in 2D, we can enumerate each vertex $\mathbf{v}_u$ of Vor($\mathcal{X}$,$\mathcal{U}$), and find the intersections of the circumference of $\Sigma_u$ with the edges or other circumferences.

*Example 2.* With reference to Figure 1, let us consider $\mathbf{v}_5$. We have $p_5 = 3$ vertices (namely, $\mathbf{v}_6$, $\mathbf{v}_8$, and $\mathbf{v}_9$) connected to $\mathbf{v}_5$ through edges. The circumference $\Sigma_5$ centered in $\mathbf{v}_5$ intersects such edges in three points: $\mathbf{y}_{5_1}$, $\mathbf{y}_{5_2}$, and $\mathbf{y}_{5_3}$.

The appropriateness of the bounding scheme stems from tightness of the upper bound, in the sense that the value of the bound can be achieved in some hypothetical continuation of the instance being explored, i.e., for some assignment of admissible location and score to the unseen objects.

**Theorem 4.** *The bounding scheme* (4) *is tight.*

### 4.3 Pulling strategy

The pulling strategy determines how to fetch the next object, alternating between distance-based and score-based access (when available). The pulling strategy can be as simple as a *round-robin* (RR) scheduling, whereby distance- and score-based access are alternated, and probing locations are uniformly explored. Tightness of the bounding scheme and a RR strategy are sufficient to guarantee a form of instance optimality. Let $\mathcal{A}^{\text{Vor}}$ be the class of MMR-correct, deterministic bounded diversification algorithms that can discover objects by both score-based and distance-based access using the vertices of Vor($\mathcal{X}$,$\mathcal{U}$) as probing locations.

**Theorem 5.** *Algorithm 2 with tight bounding scheme* (4) *and a RR pulling strategy is instance-optimal wrt. $\mathcal{A}^{Vor}$.*

In order to further decrease $sumDepths$ (i.e., the overall amount of accessed objects), a *potential adaptive* (PA) pulling strategy can be devised so as to lower the upper bound $\tau$ more quickly. PA works as follows: when a distance-based access is to be made, the probing location $\mathbf{v}_{u^*}$ is selected with $u^* = \operatorname{argmax}_{u=1,\ldots,V} \tau_u$. Conventionally, ties are broken in favor of the probing location with the least depth, then the one with the least index in $1, \ldots, V$.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

**Theorem 6.** *Let $A^{RR}$ and $A^{PA}$ be algorithms in $\mathcal{A}^{Vor}$ using tight bounding scheme* (4) *with the RR and the PA pulling strategies, respectively. Then $sumDepths(A^{PA}, I) \leq sumDepths(A^{RR}, I)$ for all bounded diversification problems $I$.*

When both distance-based and score-based access are available, we seek the one that reduces $\tau$ at a faster rate. To this end, we compute the partial derivatives of $\tau$ wrt. the number of fetched objects. Let $n^S$ denote the number of objects retrieved by means of score-based access, and $n_u$, $u = 1, \ldots, V$, the number of objects retrieved by distance-based access from probing location $\mathbf{v}_u$. Score-based access is preferred if $\left|\frac{\partial \tau}{\partial n^S}\right| > \left|\frac{\partial \tau}{\partial n_{u^*}}\right|$. These partial derivatives can be either computed exactly if the distribution is known, or approximated by using, e.g., linear predictors.

We define SPP as the instance of Algorithm 2 that uses the tight bounding scheme (4) and the PA pulling strategy.

## 5 Related work

**Result Set Diversification.** A general formulation of diversification is introduced in [7]. Existing approaches (surveyed in [5]) rerank relevant results to introduce diversity. Unlike SPP, these approaches scan all the $n$ candidate results. Diversification in multiple dimensions is addressed in [4], where the problem is reduced to MMR by collapsing diversity dimensions in one composite similarity function. The recent work [1] examines diversity-aware search under the angle of performance. Unlike [1], our work addresses diversification in a different scenario, where objects are embedded in a vector space, and exploits the geometry in order to limit the number of accessed objects.

We have focused on algorithms that can extract, on the fly, the top $k$ relevant and diversified objects from data sources providing sorted access methods (by score or distance). This class of algorithms is very general and applies to all those cases where extracting *all* the relevant objects and then scanning them for diversification is impractical (e.g., mobile queries). Many general-purpose diversification algorithms exist [15]. For some of these (e.g., Motley), the same formal apparatus as PBMMR is clearly applicable. However, other algorithms described in [15] (e.g., GMC) fall outside the PBMMR paradigm, in that they require knowing all the objects beforehand.

**Spatial Diversification.** Spatial diversification was originally introduced by [14]. The scattered ranking approach exploits the geometry of the metric space to reduce the number of operations for creating the ranking; however, the proposed algorithms access all the $N$ relevant points. The experiment with Mechanical Turk in [13] shows that users prefer spatially diversified rankings over undiversified ones.

**Top-$k$ Query Processing.** The main design dimensions and tools for top-$k$ queries are surveyed in [8]. In [10], rank join is extended to objects in a $d$-dimensional space, with the aim of finding the best combinations of objects with high score that are close to a given point and to each other. The technique in [10] also uses geometry-driven bounds, but for a different problem (rank join) and geometry than ours.

## 6 Conclusions and Future Work

We have addressed the problem of efficiently diversifying the results of top-$k$ queries over spatial objects contained in a bounded region when only sorted access methods

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 

based on distance and/or score are permitted. Our work on top-$k$ has included the following contributions: *i) Bounded diversification with sorted access methods* is introduced for the first time and defined formally. *ii)* The *Pull/Bound Maximum Marginal Relevance* (PBMMR) family of algorithms is illustrated, which exploits spatial probing locations and the adaptive alternation of score-based and distance-based access to reduce the number of fetched objects. *iii)* An instance of PBMMR, called *Space Partitioning and Probing* (SPP), is presented, whose pulling strategy uses a tight upper bound. *iv)* SPP is shown to attain the same diversification quality and exactly the same output as MMR, the most popular result diversification algorithm, but accessing only a fraction of the objects. *v)* Experiments, omitted here for brevity, show that, with a negligible computational overhead, SPP accesses in typical conditions less than 20% of the objects (less than 2% in best conditions), a substantial gain over past work on spatial scatter queries [14], which accesses all objects. Future work includes extending SPP to arbitrary joins of data sources. Also, we plan to tackle the possible presence of uncertainty in the data as was done in [12] for top-$k$ queries.

**Acknowledgements** The authors acknowledge support from *i)* the EC's FP7 "CUbRIK" project, *ii)* the ERC "Search Computing" project, *iii)* the Italian (MIUR) "EASE" project.

## References

1. A. Angel and N. Koudas. Efficient diversity-aware search. In *SIGMOD Conference*, pages 781–792, 2011.
2. J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98*, pages 335–336, 1998.
3. M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, 2008.
4. Z. Dou et al. Multi-dimensional search result diversification. In *WSDM '11*, pages 475–484, 2011.
5. M. Drosou and E. Pitoura. Search result diversification. *SIGMOD Rec.*, 39(1):41–47, 2010.
6. P. Fraternali, D. Martinenghi, and M. Tagliasacchi. Top-k bounded diversification. In *SIGMOD Conference*, 2012.
7. S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW '09*, pages 381–390, 2009.
8. I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-$k$ query processing techniques in relational database systems. *ACM Comput. Surv.*, 40(4), 2008.
9. A. Marian, N. Bruno, and L. Gravano. Evaluating top- queries over web-accessible databases. *ACM Trans. Database Syst.*, 29(2):319–362, 2004.
10. D. Martinenghi and M. Tagliasacchi. Proximity rank join. *PVLDB*, 3(1):352–363, 2010.
11. K. Schnaitter and N. Polyzotis. Evaluating rank joins with optimal cost. In *PODS*, pages 43–52, 2008.
12. M. A. Soliman, I. F. Ilyas, D. Martinenghi, and M. Tagliasacchi. Ranking with uncertain scoring functions: semantics and sensitivity measures. In *SIGMOD Conference*, pages 805–816, 2011.
13. J. Tang and M. Sanderson. Evaluation and user preference study on spatial diversity. In *ECIR*, pages 179–190, 2010.
14. M. J. van Kreveld et al. Multi-dimensional scattered ranking methods for geographic information retrieval. *GeoInformatica*, 9(1):61–84, 2005.
15. M. R. Vieira, H. L. Razente, M. C. N. Barioni, M. Hadjieleftheriou, D. Srivastava, C. T. Jr., and V. J. Tsotras. On query result diversification. In *ICDE*, pages 1163–1174, 2011.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6, 