The DisC Diversity Model*

Marina Drosou
Computer Science & Engineering Dept.
University of Ioannina, Greece
mdrosou@cs.uoi.gr

Evaggelia Pitoura
Computer Science & Engineering Dept.
University of Ioannina, Greece
pitoura@cs.uoi.gr

ABSTRACT

In this paper, we summarize our work on diversification based on *dissimilarity* and *coverage* (*DisC* diversity) by presenting our main theoretical results and contributions.

1. DISC DIVERSITY

Diversification has attracted considerable attention, often as a means of enhancing the quality of the query results presented to users [3]. Most diversification approaches rely on assigning a diversity score to each data *item* and then selecting as diverse either the k items with the largest score for a given k (e.g., [1]), or the items with score larger than some predefined threshold (e.g., [9]).

In our work [4, 5], we address diversity through a different perspective and aim at selecting a representative subset that contains items that are *both* dissimilar with each other *and* cover the whole result set.

Let $\mathcal{P}$ be a set of items. We define similarity between two items using a distance metric d. For a real number r, $r \geq 0$, we use $N_r(p_i)$ to denote the set of *neighbors* (or, the *neighborhood*) of an item $p_i \in \mathcal{P}$, i.e., the items lying at distance at most r from p_i:

$$N_r(p_i) = \{p_j \mid p_i \neq p_j \wedge d(p_i, p_j) \leq r\}$$

We use $N_r^+(p_i)$ to denote the set $N_r(p_i) \cup \{p_i\}$. Items in the neighborhood of p_i are considered similar to p_i, while items outside its neighborhood are considered dissimilar to p_i. We define an r-DisC diverse subset as follows:

Definition 1. (r-DisC Diverse Subset) *Let $\mathcal{P}$ be a set of items and r, $r \geq 0$, a real number. A subset S of $\mathcal{P}$ is an r-Dissimilar-and-Covering diverse subset, or r-DisC diverse subset, of $\mathcal{P}$, if the following two conditions hold: (i) (coverage condition) $\forall p_i \in \mathcal{P}$, $\exists\ p_j \in N_r^+(p_i)$, such that $p_j \in S$ and (ii) (dissimilarity condition) $\forall\ p_i, p_j \in S$ with $p_i \neq p_j$, it holds that $d(p_i, p_j) > r$.*

*This work was supported by "Epirus on Android" a research project co-financed by the European Union (European Regional Development Fund-ERDF) and Greek national funds through the Operational Program "THESSALY-MAINLAND GREECE AND EPIRUS-2007-2013" of the National Strategic Reference Framework (NSRF 2007-2013)

The first condition ensures that all items in $\mathcal{P}$ are represented by at least one similar item in S and the second condition that the items in S are dissimilar to each other. We call every item $p_i \in S$ an *r-DisC diverse item* and r the *radius* of S. Instead of specifying a required size k of the diverse set or a threshold, our tuning parameter r explicitly expresses the degree of diversification and determines the size of the diverse set. Increasing r results in a smaller, more diverse subset, while decreasing r results in a larger, less diverse subset.

There may be more than one dissimilar and covering diverse subsets for the same set of items $\mathcal{P}$. Since we want a concise representation of $\mathcal{P}$, we select the smallest one:

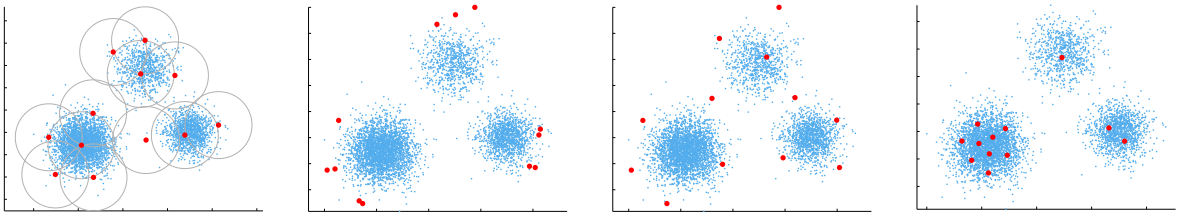
Definition 2. (Minimum r-DisC Diverse Subset Problem) *Given a set $\mathcal{P}$ of items and a radius r, $r \geq 0$, find an r-DisC diverse subset S^* of $\mathcal{P}$, such that, for every r-DisC diverse subset S of $\mathcal{P}$, it holds that $|S^*| \leq |S|$.*

It has been shown that any r-DisC diverse subset S of $\mathcal{P}$ is at most B times larger than any minimum r-DisC diverse subset S^*, where B is the maximum number of independent (i.e., dissimilar to each other) neighbors of any item in P [4]. B depends on the distance metric used and on the dimensionality of the data space. In many cases, B is a constant, e.g., for the 2D Euclidean plane, $B = 5$.

Comparison with Other Models. Let us now compare DisC with two widely used diversification models, namely MaxMin and MaxSum, that aim at selecting a subset S of $\mathcal{P}$ so as the minimum or the average pairwise distance of the selected items is maximized (e.g., [7, 8, 2]). We also compare DisC with k-medoids, a widespread clustering algorithm. In this case, the located medoids constitute the representative subset S. Input in all the above approaches is the size k of the diverse subset S. Figure 1 shows the corresponding sets attained by first locating an r-DisC diverse subset for a given r and then using the size of the produced diverse subset as the input k of the other approaches. Here, $r = 0.15$ and $k = 12$.

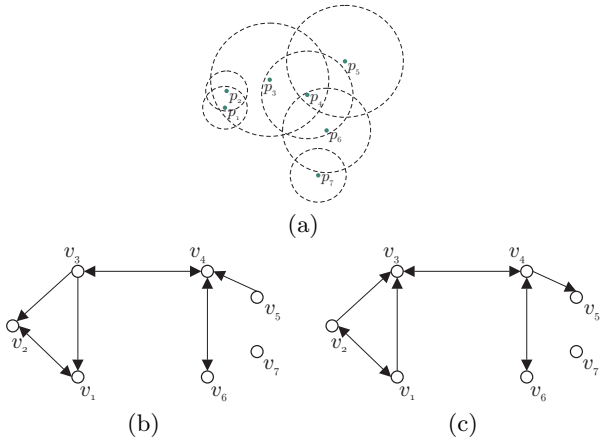
MaxSum and k-medoids fail to cover all areas of the dataset; MaxSum focuses on the outskirts of the dataset, whereas k-medoids reports only central items, ignoring items that are further away. MaxMin performs better in this aspect. However, since MaxMin seeks to retrieve items that are as far apart as possible, it fails to retrieve items from dense areas. DisC avoids most of these problems.

Multiple Radii. There may be cases in which we want different parts of the data space to be represented with more



(a) r-DisC. (b) MaxSum. (c) MaxMin. (d) k-medoids.

Figure 1: Diverse subsets of size $k = 12$ produced by different diversification methods for a clustered dataset. Selected items are shown as (red) solid circles. Circles around items of the DisC solution denote the radius r of the selected items.



(a) (b) (c)

Figure 2: (a) A set of items associated with different radii and their graph representation for the (b) Covering and (c) CoveredBy problems. A directed edge from v_i to v_j indicates that $d(p_i, p_j) \leq r(p_i)$ and $d(p_i, p_j) \leq r(p_j)$ respectively.

or less items. Thus, we consider the more general case where each item p_i is associated with a different radius $r(p_i)$.

The problem now loses its symmetry, since an item p_i may be in the neighborhood of an item p_j, while p_j is not in the neighborhood of p_i. This gives rise to two different interpretations of radius. One interpretation is that p_i can *represent* all items in its neighborhood. The other interpretation is that p_i can *be represented* by all items its neighborhood. We call the first problem *Covering DisC diverse subset problem* and the second one *CoveredBy DisC diverse subset problem*.

Definition 3. (Covering (resp. CoveredBy) DisC Diverse Subset) *Let $\mathcal{P}$ be a set of items and $r : \mathcal{P} \rightarrow \mathbb{R}^+$ be a function determining the radius of each item in $\mathcal{P}$. A subset S of $\mathcal{P}$ is a Covering (resp. CoveredBy) Dissimilar-and-Covering diverse subset, or Covering (resp. CoveredBy) DisC diverse subset, of $\mathcal{P}$, if the following two conditions hold: (i) (coverage condition) $\forall p_i \in \mathcal{P}$, $\exists\, p_j$ with $d(p_i, p_j) \leq r(p_j)$ (resp. $d(p_i, p_j) \leq r(p_i)$), such that $p_j \in S$ and (ii) (dissimilarity condition) $\forall\, p_i, p_j \in S$ with $p_i \neq p_j$, it holds that $d(p_i, p_j) > \max\{r(p_i), r(p_j)\}$.*

Figure 3 presents a qualitative view of various options of assigning radii to items. We present three different scenaria. The first one corresponds to the case where some parts of the dataset are considered more important than others and we want them to be represented with more items. In Figure 3a, items in each of the four quadrants are assigned increasing radii as we move clockwise. The second scenario corresponds to the case in which we want to take into account density, so that dense areas are not under-represented in the diverse subset. In this case, we assign smaller radii to items in denser areas (Figure 3b). The third scenario corresponds to the case in which we want to relate representation with relevance. For example, for the CoveredBy problem, we assign smaller radii to items wither larger relevance (Figure 3c and Figure 3d). This ensures that each item can be covered only by items that have a larger relevance than it.

Graph Representation and NP-hardness. Besides the geographical interpretation of DisC diversity, there is also a corresponding graph representation. We define next the corresponding graph models for both the single and the multiple radii cases.

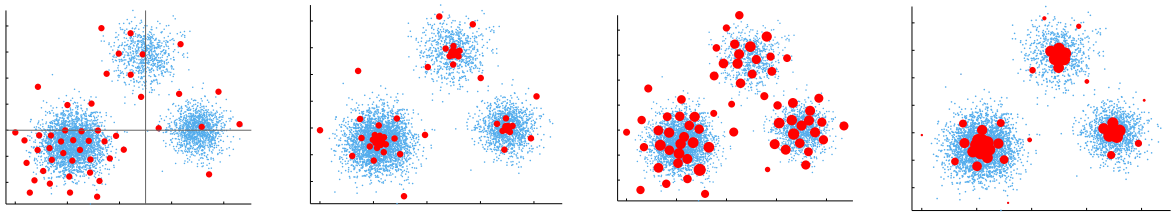
For a single radius r, let $G_{\mathcal{P},r} = (V, E)$ be an undirected graph such that there is a vertex $v_i \in V$ for each item $p_i \in \mathcal{P}$ and an edge $(v_i, v_j) \in E$, if and only if, $d(p_i, p_j) \leq r$ for the corresponding items p_i, p_j. Considering multiple radii, let $G_{\mathcal{P},r(.)} = (V, E)$ be a directed graph such that there is a vertex $v_i \in V$ for each item $p_i \in \mathcal{P}$ and a (directed) edge $(v_i, v_j) \in E$, if and only if, for the corresponding items p_i, p_j, it holds that $d(p_i, p_j) \leq r(p_i)$ (Covering problem) or $d(p_i, p_j) \leq r(p_j)$ (CoveredBy problem). An example is shown in Figure 2.

It turns out that DisC diverse subsets correspond to independent and dominating sets of the corresponding graphs. A *dominating* set D for a graph G is a subset of vertices of G such that every vertex of G not in D is joined to at least one vertex in D by some edge when G is undirected and by an incoming edge when G is directed. An *independent* set I for a graph G is a set of vertices of G such that for every two vertices in I, there is no edge connecting them. Intuitively, a dominating set of $G_{\mathcal{P},r}$ satisfies the covering condition of the DisC diverse subset, whereas an independent set of $G_{\mathcal{P},r}$ satisfies the dissimilarity condition of the DisC diverse subset.

Lemma 1. *Finding a DisC diverse subset for a set $\mathcal{P}$ is equivalent to finding an independent dominating set of the corresponding graph G.*

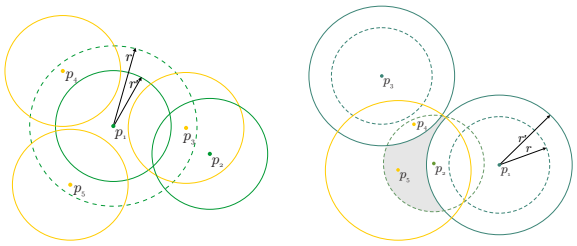
Finding a minimum independent dominating set of a graph has been proven to be NP-hard (e.g., [6]).

Computing DisC Diverse Subsets. Next, we present a general algorithm for locating DisC diverse subsets (Algorithm 1). For presentation convenience, let us call *black* the items of $\mathcal{P}$ that are in the diverse subset S, *grey* the



(a) Areas. (b) Density. (c) Uniform relevance. (d) Clustered relevance.

Figure 3: Using multiple radii. Selected items are shown as solid circles.



(a) (b)

Figure 4: Zooming (a) in and (b) out. Solid (resp. dashed) circles around items denote the previous (resp. new) radius.

Algorithm 1 Locating DisC diverse subsets.

Input: A set of items $\mathcal{P}$, a radius function $r(.)$ and a selection criterion $\mathcal{C}(.)$.
Output: A DisC diverse subset S of $\mathcal{P}$.

```
1: S ← ∅
2: for all p_i ∈ P do
3:     color p_i white
4: end for
5: while there exist white items do
6:     select the white item p_i with the largest value of C(p_i)
7:     S = S ∪ {p_i}
8:     color p_i black
9:     for all p_j ∈ N^W_{r(p_i)}(p_i) (Covering) or p_j s.t. p_i ∈
       N_{r(p_j)}(p_j) (CoveredBy) do
10:        color p_j grey
11:    end for
12: end while
13: return S
```

items covered by some item in S and *white* the items that are neither black nor grey. $N_r^W(p_i)$ denotes the set of white neighbors of p_i. Initially, S is empty and all items are white. Items are selected for inclusion in S in rounds based on some selection criterion C.

For the single radius case, selecting at each round any white item will result in a DisC diverse subset. In addition, the greedy algorithm that selects at each round the white item p_i with the largest white neighborhood $N_r^W(p_i)$ results in DisC diverse subsets with size close to the minimum one [4]. For the multiple radii case, to attain DisC diverse items, we need to select white items in decreasing order of their radius for the Covering problem and in increasing order of their radius for the CoveredBy problem.

Zooming. We also consider a *zooming* operation where, after being presented with an initial set of results for some radius r, a user asks to see either more or less results by correspondingly decreasing or increasing the radius. For simplicity, we shall focus on zooming in the case of a single radius. Formally, given a set of items $\mathcal{P}$ and an r-DisC diverse subset S of $\mathcal{P}$ for some specific radius, we want to compute an r'-DisC diverse subset S' of $\mathcal{P}$. There are two cases: (i) $r' < r$ (*zooming-in*) and (ii) $r' > r$ (*zooming-out*). Ideally, $S' \supseteq S$, for $r' < r$ and $S' \subseteq S$, for $r' > r$ (Figure 4).

To study the relationship between S and S', for two radii r_1, r_2, $r_2 \geq r_1$, we define the set $N^I_{r_1,r_2}(p_i)$, as the set of items at distance at most r_2 from p_i which are at distance at least r_1 from each other. $|N^I_{r_1,r_2}(p_i)|$ can be bounded for specific distance metrics and dimensionality [4].

When zooming-in, we construct diverse sets that are supersets of S by adding items to S. It holds that:

LEMMA 2. *For zooming-in: (i)* $S \subseteq S'$ *and (ii)* $|S'| \leq |S| + \sum_{p_i \in S} |N^I_{r',r}(p_i)|$

When zooming-out, it may not be possible to construct a DisC diverse subset S' that is a subset of S. Thus, we proceed in two passes. In the first pass, we examine all items of S in some order and remove their diverse neighbors that are now covered by them. At the second pass, items from any uncovered areas are added to S'. It holds that:

LEMMA 3. *For zooming-out: (i) There are at most* $\sum_{p_i \in S} |N^I_{r,r'}(p_i)|$ *items in* $S \backslash S'$*, (ii) For each item of* S *not included in* S'*, at most* $B - 1$ *items are added to* S'*.*

2. SUMMARY AND FUTURE WORK

In a nutshell, we introduced a new, intuitive definition of diversity based on using a radius r rather than a size limit k. We presented both a geometrical and an equivalent graph-based interpretation of our model. We introduced incremental diversification through zooming-in and zooming-out, showed that locating DisC diverse subsets is an NP-hard problem and provided efficient algorithms for their computation. Directions for future work include extending our approach to the budgeted r-DisC problem, that is, computing DisC subsets of a specific size that maximize coverage and also studying different variations of our zooming operations.

3. REFERENCES

[1] A. Angel and N. Koudas. Efficient diversity-aware search. In *SIGMOD Conference*, 2011.

[2] A. Borodin, H. C. Lee, and Y. Ye. Max-sum diversification, monotone submodular functions and dynamic updates. In *PODS*, pages 155–166, 2012.

[3] M. Drosou and E. Pitoura. Search result diversification. *SIGMOD Record*, 39(1):41–47, 2010.

[4] M. Drosou and E. Pitoura. Disc diversity: result diversification based on dissimilarity and coverage. *PVLDB*, 6(1):13–24, 2012.

[5] M. Drosou and E. Pitoura. Poikilo: A tool for evaluating the results of diversification models and algorithms. *PVLDB*, 6(12):1246–1249, 2013.

[6] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[7] S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW*, pages 381–390, 2009.

[8] 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.

[9] C. Yu, L. V. S. Lakshmanan, and S. Amer-Yahia. It takes variety to make a world: diversification in recommender systems. In *EDBT*, pages 368–378, 2009.