Quality metrics for diversified similarity searching: What they stand for?*

Camila R. Lopes[1], **Daniel L. Jasbick**[2],
Marcos Bedo[3], **and Lúcio F. D. Santos**[1]

[1]Federal Institute of North of Minas Gerais (IFNMG)
R. Dois, 300 – Montes Claros/MG – Brazil

{camila,lucio.santos}@ifnmg.edu.br

[2]Institute of Computing – Fluminense Federal University (UFF)
Av. Gal. Milton Tavares de Souza, S/N – Niterói/RJ – Brazil

[3]Fluminense Northwest Institute – Fluminense Federal University (UFF)
Av. João Jasbick, S/N – St. A. Pádua/RJ – Brazil

{danieljasbick,marcosbedo}@id.uff.br

Abstract. *Diversity-oriented searches retrieve objects not only similar to a reference element but also related to the different types of collections within the queried dataset. While such characterization is flexible enough to include methods originally from information retrieval, data clustering, and similarity searching under the same umbrella, diversity metrics are expected to be much less paradigm-biased in order to discriminate which approaches are more suitable and when they should be applied. Accordingly, we extend and implement a broad set of quality metrics from those distinct realms and experimentally discuss their trends and limitations. In particular, we evaluate the suitability of data clustering indexes, and similarity-driven measures regarding their adherence to diversified similarity searching. Experiments in real-world datasets indicate such measures are capable of distinguishing diversity methods from different paradigms, but they heavily favor the approaches of the same group – especially cluster indexes. As an alternative, we argue diversity is better addressed by a set of measures rather than a single quality value. Therefore, we propose the Diversity Features Model (`DFM`) that combines the perspectives of the competing approaches into a multidimensional point whose features are calculated based on the distance distribution within both retrieved and queried datasets. Empirical evaluations showed `DFM` compares different diversity searching approaches by considering multiple criteria, whereas overall winners can be found by ranking aggregation or visualized through parallel coordinates maps.*

1. Introduction

The proverb *"A picture is worth a thousand words"* is no longer just a metaphor since images and videos produced in different application domains, such as social networks, biology, and astronomy, outweigh text content in orders of magnitude regarding data volume and size [Pouyanfar et al. 2018]. An efficient approach for querying such data is the

*This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), CNPq, and Research Support Foundation of Rio de Janeiro State.

Metric Spaces Model [Hetland 2009], where objects are mapped into a known domain so that elements become comparable by a distance function. While feature-learning can be used for mapping complex domains into simpler spaces, *e.g.*, multidimensional domains, *similarity criteria* are employed for the retrieval of objects according to their distances to a query object [Santos et al. 2013b, Aggarwal 2015]. Under that rationale, the farther the elements, the most dissimilar they are. The most common similarity criterion is that of neighborhood (k-NN) queries, which fetch the k nearest objects to a reference element.

Neighborhood searches are efficiently executed by index-and-query algorithms [Chen et al. 2017], but they present a semantic drawback in the querying of massive datasets. For instance, suppose a composer runs a search for the *five most similar tunes to the "Beatles Yellow Submarine"* in a social network repository that returns versions and parodies of the aforementioned song in different languages. Although the result set can be correct from a k-NN query standpoint, the answer is likely unfruitful. A *diversified* answer could consider not only the closest songs but also those of distinct musical styles, *i.e.*, different *collections* within the queried dataset.

Diversified similarity searching methods cover that semantic query aspect according to three main paradigms: *(i)* distance-based, *(ii)* novelty-based, and *(iii)* coverage-based [Drosou et al. 2017]. Methods from the first paradigm aim at maximizing a single objective function regarding the distances between the elements within the result set [Zheng et al. 2017], whereas approaches of the second paradigm rely on a two-phase execution in which an enlarged subset of candidates is chosen and then filtered to ensure diversification. Such a two-phase strategy can be reduced to the problem of solving a bi-criteria objective function where similarity and diversity compete linearly following a user-defined parameter [Vieira et al. 2011]. Finally, coverage-based methods separate candidates on-the-fly following a similarity threshold, which creates *dynamic* clusters in the search space [Santos et al. 2013b].

While a plethora of diversity-driven methods can be found in the literature [Drosou et al. 2017], quality metrics for assessing their efficacy are rather scarce [Smyth and McClave 2001, Santos et al. 2013a]. In fact, most studies borrow and adapt quality metrics originally designed for information retrieval and similarity searching tasks [Vieira et al. 2011, Zheng et al. 2017]. Examples of those metrics include *information retrieval* measures NDCG-IA (*Intent-Aware Normalized Discounted Cumulative Gain*) [Agrawal et al. 2009], and similarity-oriented approaches RB (*Relative Benefit*) and OEM (*Overlap Evaluation Method*) [Smyth and McClave 2001, Santos et al. 2013a].

Aiming at investigating the biases of those metrics towards diversity algorithms, we *(i)* adapt a set of distinct measures, and *(ii)* extend clustering indexes, such as Silhouette [Aggarwal 2015], for the evaluation of different diversified similarity searching methods. Results indicate distinct quality measures may favor groups of diversity algorithms, being that separation clearer for scores from cluster-oriented metrics. Accordingly, we propose a new multidimensional evaluation measure for diversity, named *Diversity Features Model* (`DFM`), by combining previous relevant metrics from data clustering and similarity-driven statistics. Our argument is diversified similarity searching is fairer addressed by a multidimensional viewpoint whose entries are calculated based on the distance distributions within both retrieved and queried sets, rather than single quality indexes. Moreover, we claim `DFM` outputs can be interpreted as multiple lists of preferences so that overall winners can be found by ranking aggregation methods [Fagin et al. 2003]. We evaluated DFM in real-world datasets, and results indicated our approach is flexible enough for comparing different diversity methods with multiple criteria, whereas DFM entries can also be visualized and interpreted through parallel coordinates maps. The main contributions of this study are summarized as follows:

1. Extension of data clustering indexes and similarity-driven measures for the evaluation of diversified similarity searches,
2. Experimental shreds of evidence that metrics may favor groups of algorithms from the same paradigm,
3. A new multidimensional measure for assessing the quality of diversified similarity searching outputs, which complies with ranking aggregation principles.

The remainder of the paper is organized as follows. Section 2 provides background concepts and discuss diversity searching. Section 3 presents the extensions for diversity metrics and introduces DFM. Section 4 provides the experimental evaluations and comparisons, while Section 5 concludes the study.

2. Preliminaries

2.1. Diversified similarity searching

A *metric space* is a pair $\langle \mathbb{S}, \delta \rangle$ with a given domain $\mathbb{S}$ and a distance function δ that comply with properties of *(i)* symmetry, $\delta(s_i, s_j) = \delta(s_j, s_i)$; *(ii)* non-negativity, $\delta(s_i, s_j) \geq 0$; and *(iii)* triangle inequality, $\delta(s_i, s_g) + \delta(s_g, s_j) \geq \delta(s_i, s_j)$, for any objects $s_g, s_i, s_j \in \mathbb{S}$. Examples of distance functions include the Minkowski family L_p, being the L_2 function the well-known Euclidean distance.

Given a particular dataset $\mathcal{S} \subseteq \mathbb{S}$, a *range query* ($Rq$) retrieves every element in $\mathcal{S}$ that is far from a query element $s_q \in \mathbb{S}$ at most a given threshold $\xi \in \mathbb{R}_+$ so that $Rq(s_q, \xi, \mathcal{S}, \delta) = \{s_i \mid s_i \in \mathcal{S}, \delta(s_i, s_q) \leq \xi\}$. Analogously, a *neighborhood query* (k-NN) retrieves $k, k \in \mathbb{N}$, elements in $\mathcal{S}$ whose distances to query element $s_q \in \mathbb{S}$ are the smallest, *i.e.*, a neighborhood query is a range search[1] with an initially unknown radius ξ so that $|Rq| = k$ [Hetland 2009, Chen et al. 2017].

Range and k-NN queries may struggle in the searching of high-density datasets since distances among elements may be very close to each other. In this scenario, small variations in the radius ξ produce large variations in the result set cardinality. As a consequence, k-NN queries may become unstable since their results are likely non-unique[1] [Pestov 2013]. Density may also reduce the utility of similarity searching in data exploration as retrieved elements are expected to be very similar among themselves. *Diversity* is employed in that context for enhancing neighborhood queries.

Roughly speaking, given a diversity metric div and a value $k \leq |\mathcal{S}|, \mathcal{S} \subseteq \mathbb{S}$, a diversified result set $\mathcal{R} \subseteq \mathcal{S}$ complies with $\mathcal{R} = \text{argmax}_{\mathcal{R}' \subseteq \mathcal{S}, |\mathcal{R}'|=k} div(\mathcal{R}')$. Approaches for solving such an optimization problem are categorized into three groups, namely *(i)* distance-based, *(ii)* novelty-based, and *(iii)* coverage-based. Distance-based approaches, however, examine the entire queried set $\mathcal{S}$ rather than considering the query element perspective [Jain et al. 2004, Zheng et al. 2017].

[1] Ties at the k^{th} position are broken arbitrarily.

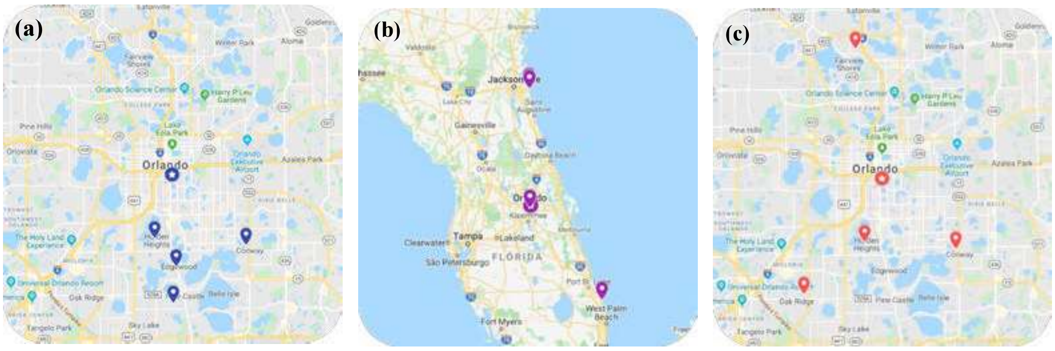

Figure 1. (a) k-NN, (b) GMC, and (c) BRID results for query "Find the 04 closest and diverse cities to Orlando/FL" with $\delta_{sim} = \delta_{div} = \delta_{div_2} = L_2$, and $\lambda = 0.5$.

Novelty-based methods model the optimization problem as a *dual-criteria* function in which *similarity* and *diversity* compete among themselves ruled by a linear λ parameter defined in the $[0, 1]$ interval. The setting of parameter $\lambda = 0$ (weighting diversity as irrelevant) turns the optimization problem in a neighborhood query, whereas increasing λ values push retrieved elements away from the queried object. The finding of the optimal λ value is an NP-hard problem [Drosou et al. 2017], and practical solutions rely on meta-heuristics for producing suitable outputs [Vieira et al. 2011].

The *Maximal Marginal Relevance* (MMR) [Carbonell and Goldstein 1998] method adapts that dual model for diversified neighborhood queries by assigning a score for each element s_i in the queried set $\mathcal{S}$ according to the function $MMR(s_i, s_q) = (1 - \lambda) \cdot \delta_{sim}(s_i, s_q) + 2 \cdot \lambda \cdot \sum_{s_j \in \mathcal{R}} \delta_{div}(s_i, s_j)$, where similarity and diversity are measured by distinct functions δ_{sim} and δ_{div}, respectively. The result set $\mathcal{R}$ (initially empty) is built in k incremental steps so that the element $s_i \in \mathcal{S} \setminus \mathcal{R}$ with highest value of $MMR(s_i, s_q)$ is chosen as the next nearest diversified neighbor at each step. The *Greedy with Marginal Contribution* (GMC) [Vieira et al. 2011] method uses a new objective function $MMC(s_i, s_q) = (1 - \lambda) \cdot \delta_{sim}(s_i, s_q) + 2 \cdot \lambda \cdot \sum_{s_j \in \mathcal{R}} \delta_{div}(s_i, s_j) + 2 \cdot \lambda \cdot \sum_{s_h \in \mathcal{R}'}^{\mathcal{R}' \subseteq \mathcal{S} \setminus \mathcal{R}, |\mathcal{R}'| = h - |\mathcal{R}|} \delta_{div_2}(s_i, s_h)$ for weighting the contribution of elements s_h *outside* the partial result set $\mathcal{R}$, whereas functions δ_{div} and δ_{div_2} can be different. Figures 1(a–b) show an example of a GMC result set in contrast to that of a k-NN query.

The *Swap* [Yu et al. 2009] novelty-based method uses similarity-driven permutations for avoiding result sets to being stuck at local maxima. First, Swap constructs a result set $\mathcal{R}$ with the closest k elements to the query object. Next, the remaining elements in $\mathcal{S} \setminus \mathcal{R}$ are sorted by similarity and individually swapped with $\mathcal{R}$ objects according to an objective function, such as MMC. If the swap provides higher scores, then the changes are consolidated in $\mathcal{R}$. The method stops when every $\mathcal{S}$ element is swapped at least once. Additionally, heuristic-driven approaches as GMC and Swap rely on data sampling for speeding-up their execution in practice. Therefore, a queried set $\mathcal{S}' \subseteq \mathcal{S}$ is usually employed as their inputs rather than the entire dataset $\mathcal{S}$ [Vieira et al. 2011].

Coverage-based methods follow a different premise since they rely on creating separations in the search space for modeling diversity. For instance, the *Motley* method [Jain et al. 2004] employs a user-provided separation distance $r \in \mathbb{R}_+$ for retrieving diversified nearest neighbors. Given a query object $s_q \in \mathbb{S}$, the *Motley* implementation sorts the elements $s_i \in \mathcal{S}$ regarding their distances to s_q and includes the nearest neighbor in the diversified result set. Next, it incrementally evaluates the sorted list of candidates until the remaining $k-1$ diversified neighbors are found, *i.e.*, a distance-sorted candidate s_i is included in the partial result set $\mathcal{R}'$ if $\delta(s_i, s_j) > r, \forall\ s_j \in \mathcal{R}'$.

The *Better Results with Influence Diversification* (BRID) [Santos et al. 2013b] method creates dynamic thresholds that eliminate the need for any extra user-provided parameters. BRID separations are based on *influence*, which expresses how much result set entries cover regions in the search space. Formally, the mutual influence between a pair of objects $s_i, s_j \in \mathbb{S}$ is calculated as $I(s_i, s_j) = 1/\delta(s_i, s_j)$ so that an element s_h is said to be *more influenced* by s_i than s_j whenever $I(s_h, s_i) \geq I(s_h, s_j)$. An influence-based relationship between a query object s_q and a result set entry s_i enables constructing *strong influence sets* that include only non-influenced elements. Accordingly, a strong influence set for $\langle s_q, s_i \rangle$ is $\ddot{I}_{s_q,s_i} = \{s_j \in \mathcal{S} \mid (I(s_i, s_j) \geq I(s_i, s_q)) \wedge (I(s_j, s_i) \geq I(s_j, s_q))\}$. BRID uses strong influence sets for retrieving the k closest elements in $\mathcal{S}$ to $s_q \in \mathbb{S}$ that are non-influenced by any result set entry so that a diversified k-NN query produces $\mathcal{R} = \{r_i \in \mathcal{S} \mid \forall\ s_j \in \mathcal{R} : r_i \notin \ddot{I}_{s_j,s_q} \wedge \forall\ s_i \in \mathcal{S} \setminus \mathcal{R} : \Big(\delta(r_i, s_q) \leq \delta(s_i, s_q) \vee \exists\ s_j \in \mathcal{R} : s_i \in \ddot{I}_{s_j,s_q}\Big) \wedge |\mathcal{R}| \leq k\}$. Figures 1(b–c) compare BRID and GMC result sets.

2.2. Quality metrics for diversified similarity searching

Metrics for quantifying results from diversified similarity searching are mainly "borrowed" from information retrieval and similarity searching [Santos et al. 2013a]. For instance, the metric *Intent-Aware Normalized Discounted Cumulative Gain* (NDCG-IA) extends measure NDCG [Agrawal et al. 2009] by considering the *categories* of retrieved elements in the assignment of scores regarding results that include different labels. The drawback of applying either NDCG or NDCG-IA metrics for diversity searching is they require the queried sets to be labeled.

On the other hand, similarity searching metrics evaluate only the distances among result set elements. For instance, the *Relative Benefit* (RB) [Smyth and McClave 2001] tackles the *trade-off* between similarity and diversity by adopting k-NN queries as a baseline, *i.e.*, RB assumes the results of a k-NN query represent similarity only so that the relative benefit of a diversity algorithm can be calculated by counting the differences between its result and that of a k-NN. While such a measure highlights how much methods diverge from k-NN, it does not address the quality of diversity itself.

Quality metric *Overlap Evaluation Method* (OEM) [Santos et al. 2013a] uses the distances between elements in the result set $\mathcal{R}$ and the query object for constructing strong influence sets for every entry in $\mathcal{R}$. OEM calculates the overlapping among those strong sets by counting the number of elements lying in the intersection of both sets as an ω-score $\omega(\mathcal{R}, s_q) = 1/2 \cdot \sum_{s_i \in \mathcal{R}} \sum_{s_j \in \mathcal{R}} \Big(1 - (|\ddot{I}_{s_i,s_q} \cap \ddot{I}_{s_j,s_q}| / |\ddot{I}_{s_i,s_q} \cup \ddot{I}_{s_j,s_q}|)\Big), s_i \neq s_j$. The higher the ω-score, the better the algorithm according to the bias in which result set entries shall not influence themselves.

Finally, the metric *Dissimilarity Feature Method* [Santos et al. 2013a] introduces the idea of using multiple measures for addressing diversity. It uses six statistics calculated as the minimum, maximum, mean, and standard deviation of the distance distribution within $\mathcal{R}$. Such entries can be used for finding the most suitable diversity method through a weighted sum [Santos et al. 2013a].

3. Material and Methods

3.1. Cluster metrics extension for diversified similarity searching

Relative measures for validating clusters quantify the difference between two data partitions, which may also be suitable for comparing how representative the retrieved elements of a diversity query are. Let the elements of a set $\mathcal{S}$ be divided into k clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}, \cup_{i=1}^{k} C_i = \mathcal{S}$, and each subgroup C_i be represented by an object $s_i \in \mathcal{S}$, relative metrics target a dual optimization problem in which *(i)* inner cluster distances must be minimal, and *(ii)* distances between clusters' representatives must be maximal. The *Silhouette index* ($Sil(\mathcal{C})$) models that optimization problem by using the average distance within clusters as a normalization factor. We extend $Sil(\mathcal{C})$ into $Sil^*(\mathcal{R}, \mathcal{S}', \mathcal{C})$ for measuring the quality of a result set $\mathcal{R}$ regarding the queried set $\mathcal{S}' \subseteq \mathcal{S}$, as in Eq. (1).

$$Sil^*(\mathcal{R}, \mathcal{S}', \mathcal{C}) = \frac{\Psi(\mathcal{C})}{|\mathcal{C}|}; \Psi(\mathcal{C}) = \frac{\sum_{C_i \in \mathcal{C}} \delta_{div}(s_i, s_h)}{\psi(C_i)}; \psi(C_i) = \frac{\sum_{s_j \in C_i} \delta_{div}(s_j, s_i)}{|C_i|} \tag{1}$$

where $s_h \in S'$ is the representative element of C_h that is the closest cluster to C_i, i.e., $\delta(s_i, s_h) \leq \delta(s_i, s_j)$ for any pair $\langle C_j, s_j \rangle, C_j \in \mathcal{C}$. Another relative-based cluster measure is the *Dunn index* ($Dunn(\mathcal{C})$), which models inner and outer cluster separation according to their diameters. We extend the Dunn index into $Dunn^*(\mathcal{R}, \mathcal{S}', \mathcal{C})$ by using result set entries $s_i \in \mathcal{S}'$ and their distances to the query object $s_q \in \mathbb{S}$. Accordingly, the largest inner cluster distance is normalized by the diversity to the query element, as in Eq. (2).

$$Dunn^*(\mathcal{R}, \mathcal{S}', \mathcal{C}) = \min_{C_i, C_j \in \mathcal{C}, C_i \neq C_j} \left(\delta_{div}(s_i, s_j) / \max_{C_m \in \mathcal{C}} \left(\max_{s_h \in C_m} \delta_{div}(s_h, s_q) \right) \right) \tag{2}$$

Finally, the *Davies-Bouldin index* models clusters' quality by using the distances among them for normalization. We extend that metric into $DB^*(\mathcal{R}, \mathcal{S}', \mathcal{C})$ as in Eq. (3).

$$DB^*(\mathcal{R}, \mathcal{S}', \mathcal{C}) = \frac{1}{|\mathcal{C}|} \cdot \sum_{C_i, C_j \in \mathcal{C}, C_i \neq C_j} \left(\frac{diam(C_i) + diam(C_j)}{\delta_{div}(s_i, s_j)} \right) \tag{3}$$

where $diam(C_i)$ is the largest distance between any pair of elements in C_i.

3.2. Clusters uncovered by matching the results to the queried set

Although diversified similarity searching methods do not generate clusters on the queried set, retrieved elements can be seen as cluster *representatives* found by the querying algorithm. Under that premise, we employ the returned elements in $\mathcal{R}$ as "*medoids*" for the construction of clusters over the queried set $\mathcal{S}' \subseteq \mathcal{S}$. Such a rationale, coupled with novelty-based diversity algorithms, generates partitions similar to those of the k-Medoids clustering method, in which $\mathcal{S}'$ elements are assigned to the closest object in $\mathcal{R}$ rather than a *de facto* medoid. Figures 2(a–d) illustrate the approach where $\mathcal{S}'$ delimits a search region in $\mathcal{S}$, and queried objects are clustered around $\mathcal{R}$ entries.

Coverage-based Motley and BRID algorithms return diversified elements that are separated according to distance thresholds. Therefore, we consider the queried set $\mathcal{S}'$ as the union of the closed balls in the search space centered at elements in $\mathcal{R}$ with coverage radii equal to the distance separation r. Figures 2(e–f) show examples of clusters produced by Motley in a particular diversified neighborhood search as well as the limits of the query set $\mathcal{S}'$. Analogously, we model BRID queried set of elements $\mathcal{S}'$ according to the dynamic thresholds that define the *influences* from the entries in $\mathcal{R}$ and the query object s_q. The separation distance r for every cluster in BRID is defined as the distance between the representative element in $\mathcal{R}$ and s_q. Accordingly, clusters are also closed balls in the search centered at result set entries with coverage radii as large as their influences over the queried object. Figures 2(g–h) illustrate such cluster formation in which the limits of the query set $\mathcal{S}'$ are expected to increase with k.

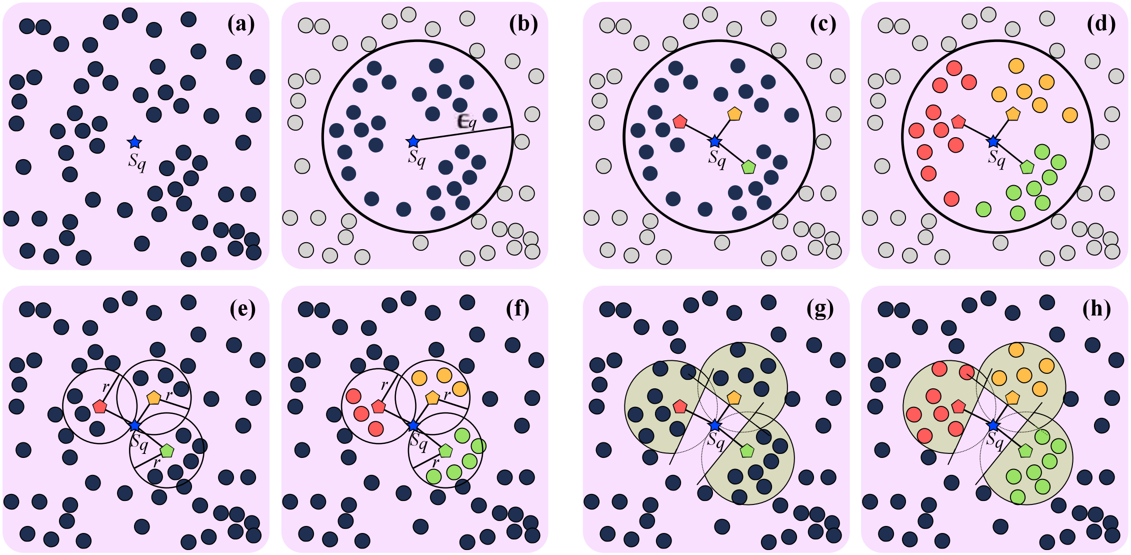

Figure 2. Clusters for diversified results. (a) Entire dataset $\mathcal{S}$, (b) closed query ball delimiting the queried set $\mathcal{S}'$, (c) $\mathcal{S}'$ in novelty-based methods, (d) elements are assigned to clusters defined by the k retrieved elements, (e) $\mathcal{S}'$ in Motley, (f) clusters are formed according to distance separation r, (g) $\mathcal{S}'$ in BRID, and (h) clusters follow dynamic *influences* to retrieved elements.

3.3. The Diversity Features Model – `DFM`

While cluster-oriented metrics measure the cohesion and separation generated by diversity searching on top of queried sets, they may also favor approaches that produce closed balls in the search space, *i.e.*, coverage-based. Such a claim is reinforced by the empirical observations reported in Section 4.1, in which cluster indexes indicated either Motley or BRID as the most suitable searching routine in 9 out of 12 comparisons.

We argue *diversity* is fairer addressed by a set of measures rather than a single quality value as an alternative for softening such separation-based bias. Accordingly, we introduce the *Diversity Features Model* (`DFM`), a multidimensional model that combines the perspectives of both cluster and similarity-oriented metrics by using the distance distributions within both retrieved and queried sets. DFM produces a seven-dimensional score for every compared search approach, where each DFM entry can be seen as a list of individual preferences so that overall winners are found by ranking aggregation. Moreover, DFM scores can be visualized through parallel coordinates maps in which a baseline line is drawn for a k-NN query. DFM $(\mathcal{R}, \mathcal{S}', \mathcal{C})$ entries are calculated as in Eq. (4).

$$
\begin{aligned}
\texttt{DFM}(\mathcal{R}, \mathcal{S}', \mathcal{C}) &= \langle Sil^*, Dunn^*, RB^*, \mu_{div}, \sigma_{div}, \mu_{sim}, \sigma_{sim} \rangle \\
\mu_{div} &= 1/\left(2 \cdot (k^2 - k)\right) \cdot \sum_{s_i \in \mathcal{R}} \sum_{s_j \in \mathcal{R}, s_j \neq s_i} \delta_{div}(s_i, s_j) \\
\sigma_{div} &= \sqrt{\frac{1}{2 \cdot (k^2 - k)} \cdot \sum_{s_i \in \mathcal{R}} \sum_{s_j \in \mathcal{R}, s_j \neq s_i} \left(\delta_{div}(s_i, s_j) - \mu_{div}\right)^2} \\
\mu_{sim} &= 1/k \cdot \sum_{s_i \in \mathcal{R}} \delta_{sim}(s_i, s_q) \\
\sigma_{sim} &= \sqrt{\frac{1}{k-1} \cdot \sum_{s_i \in \mathcal{R}} \left(\delta_{sim}(s_i, s_q) - \mu_{sim}\right)^2}
\end{aligned}
\tag{4}
$$

4. Experiments

This section provides an empirical evaluation of quality metrics for diversified similarity searches regarding four real-world datasets. Table 4 describes the queried datasets, including their cardinality $|\mathcal{S}|$, dimensionality $\mathbb{R}^d$, and associated distance function δ.

Table 1. List of queried datasets.

Name	$\|\mathcal{S}\|$	$\mathbb{R}^d$	$\delta_{sim} = \delta_{div}$	Description
US_CITIES	25,375	2	L_2	Geographic entries of U.S. cities.
NASA	40,150	20	L_2	Features from NASA/SISAP images.
PHOTO_F	300	256	L_2	Features from photos of human faces.
FACES	1,016	761	L_1	Characteristics from face images.

Datasets were split according to a *holdout* rule in every evaluation where 100 elements were employed as query objects, and the remaining entries were used as the queried set. Additionally, since novelty-based algorithms examine a factorial-based number of combinations on data cardinality, we restrict the queried set $\mathcal{S}'$ within the original dataset $\mathcal{S}$ so that experiments could finish within a reasonable time (weeks). In particular, we set $\mathcal{S}'$ by using an average radius that covers at least $10\times$ the maximum value of queried neighbors. Under that rationale, we bound $|\mathcal{S}'|$ to $300, 300, 200,$ and 500 in datasets US_CITIES, NASA, PHOTO_F, and FACES, respectively. Motley distance-separation threshold r was determined empirically for every queried set $\mathcal{S}'$ so that exactly k neighbors are returned for each diversified neighborhood search. The approaches were implemented by using Python version 3.6.8 running in a local machine GNU/Linux Mint 19.2 with an Intel Core $i7$ processor, 16GB RAM, and a 1TB SATA disk.

4.1. Individual quality metrics for diversity searching

The reported values represent the average measures obtained for the execution of 100 diversity searches for $k = \{5, 7, 9, 11, 13, 15, 17\}$. Motley's parameter was set to $r = 0.3, 0.3, 1.10^4$, and 25 in datasets US_CITIES, NASA, PHOTO_F, and FACES, respectively. The objective function metric was defined as the same scoring function of method MMR. In all experiments, the higher the individual metric value, the better.

In the first evaluation, we evaluate the impact of parameter λ in novelty-based algorithms. Due to space limitations, we report only a representative analysis for varying λ values ($\lambda = \{0.0, 0.3, 0.5, 0.7, 1\}$) and a fixed number of neighbors $k = 9$ regarding the queried set US_CITIES, which summarizes the dominant behavior we found by searching other datasets with the same λ setup. Figure 3 shows the quality metrics for novelty-based methods Swap, MMR, and GMC. Cluster-oriented measures Sil^*, $Dunn^*$, and DB^* increased with λ, but similarity-driven metrics peaked with distinct and intermediate λ values in each dataset. The average performance was reached by value $\lambda = 0.5$, which we use as the query setup in the next experiments.

Figure 4 juxtaposes both novelty and coverage-based methods, and results show cluster-oriented metrics Sil^*, $Dunn^*$ and DB^* separated those two approaches by assigning better scores to coverage-based algorithms. Silhouette endorsed BRID as the most suitable choice, *Davies-Bouldin* chose Motley, and *Dunn* exposed the choice between BRID and Motley may depend on the neighborhood value k. Notice, although different cluster-oriented metrics produced different outputs, all of them picked coverage-based methods. Finally, experimental results pinpoint cluster-oriented metrics may be unsuitable for measuring diversity regarding high-dimensional datasets as the ratio between similarity-only k-NN and coverage-based diversity methods reduced expressively for larger values of k in high-dimensional sets PHOTO_F and FACES, respectively.

On the other hand, the metric *Objective function* assigned the highest scores to algorithm MMR (which attempts to maximize the scoring function itself) but no overall significant differences were found between the quality of MMR and either novelty-based (*e.g.*, Swap) or coverage-based approaches (*e.g.*, Motley). Analogously, the influence-based OEM metric assigned higher scores to BRID (which uses influences for finding diversity), but it was unable to promote other coverage-based methods. Altogether, OEM-

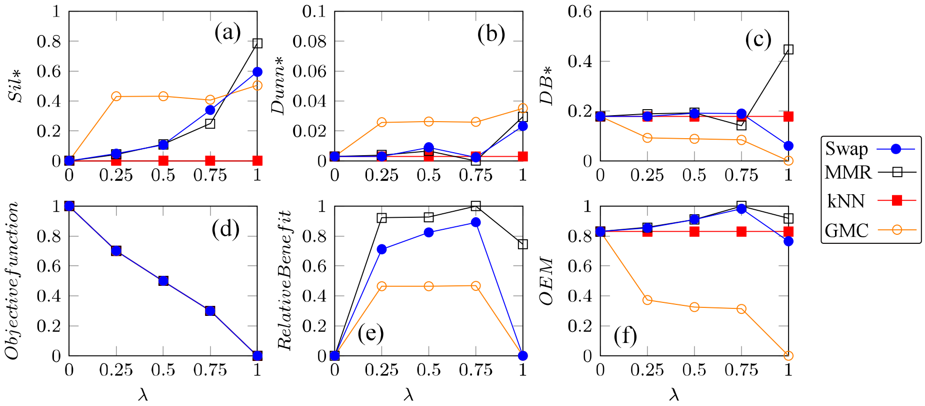

Figure 3. Individual quality scores for different values of λ.

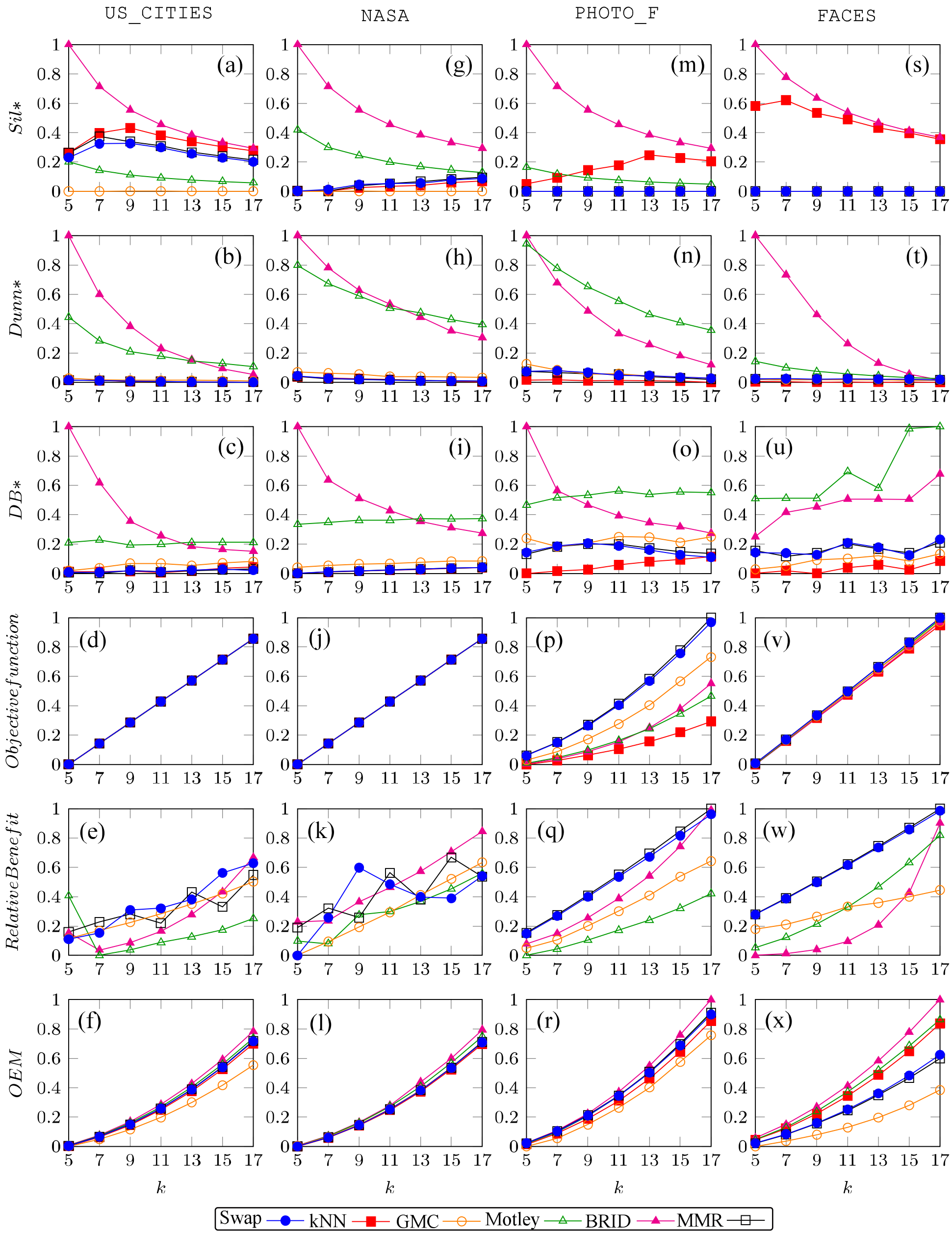

Figure 4. Quality of searching methods according to six metrics.

oriented results indicate influences were also found by novelty-based approaches as in the MMR and Swap performances over US_CITIES. Similarity-based RB metric was also unable to separate novelty and coverage-based methods, as no consistent differences were observed between optimization methods over competitors BRID and Motley.

Such findings indicate cluster-oriented metrics favored coverage-based diversity approaches, which can be explained by the separation principle found within clustering outputs. On the other hand, similarity-driven metrics were unable to assert fair winners, since they promote algorithms that follow a predictable result set construction rule, *e.g.*, OEM favored BRID, and Objective function endorsed MMR. In the next experiments, we examine how those biases can be softened by using a multidimensional metric model.

4.2. Multidimensional metrics for diversity searching

We evaluate DFM outputs regarding two distinct perspectives: *(i)* a visual quality assessment by using parallel coordinates maps, and *(ii)* a consolidation of compared diversity methods through a ranking aggregation approach. In both cases, we juxtapose the DFM results regarding diversity algorithms in comparison to those produced by a k-NN query. Figures 5(a–d) show DFM outputs for value $k = 9$ and the search setup of experiments in Figures 4, in which multidimensional entries are normalized.

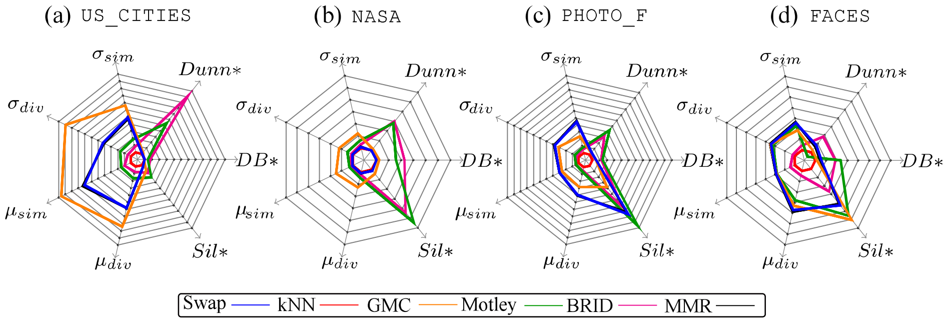
(a) US_CITIES (b) NASA (c) PHOTO_F (d) FACES

Figure 5. Visualization of DFM outputs by parallel coordinates maps.

The average scores for the k-NN result sets form the smallest circles inside the parallel coordinates and distortions onto the map borders indicate better performances. Statistic-driven measures attracted novelty-based methods, whereas cluster-oriented metrics drew coverage-based approaches. GMC and Swap covered the largest area for set US_CITIES, whereas methods BRID and Motley showed the largest distortions in comparison to k-NN in the DFM map of NASA set. Finally, methods GMC and MMR Swap covered the largest area in the parallel coordinates of PHOTO_F and FACES sets.

Table 2. Top-3 diversity methods according to DFM entries and MedianRank.

Top-k	US_CITIES	NASA	PHOTO_F	FACES
#1	GMC	GMC	MMR	MMR
#2	Swap	Motley	Swap	Swap
#3	MMR	BRID	GMC	Motley
#4	Motley	MMR	Motley	GMC
#5	BRID	Swap	BRID	BRID
#6	k-NN	k-NN	k-NN	k-NN

We employed the MedianRank algorithm [Fagin et al. 2003] for ranking those visual observations and determine which were the most suitable diversity methods. Accordingly, we consider each DFM dimension as one individual and weightless ranking whose positions are determined by the scores reached by every compared quality metric – Table 2. Results indicate distinct algorithms may be more efficient for querying different sets regardless of their paradigm. For instance, the top-3 outcomes hint GMC and Swap algorithms were the most suitable methods for querying set US_CITIES, while GMC and Motley approaches were the most appropriate for searching NASA. Such findings are slightly different from the visual area analysis since MedianRank fetches the median position of algorithms in individual rankings. The results pinpoint DFM enables softening individual biases from different metrics by combining those measures through either data visualization, which indicates how much a diversity search output diverges from a k-NN query, or ranking aggregation, which filters top-k performances.

5. Conclusions

This paper has discussed diversity searching algorithms and quality metrics for measuring their performances. Since individual indexes tend to favor particular groups of algorithms, we proposed a multidimensional metric model, coined `DFM`, for softening individual biases. Experimental evaluations indicated `DFM` outputs are visualized by parallel coordinates maps, whereas overall winners can be spotted through ranking aggregation.

References

Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. *ACM WSDM*, 1(1):5–14.

Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. *ACM SIGIR*, 1(1):335–336.

Chen, L., Gao, Y., Zheng, B., Jensen, C. S., Yang, H., and Yang, K. (2017). Pivot-based metric indexing. *PVLDB*, 10(10).

Drosou, M., Jagadish, H., Pitoura, E., and Stoyanovich, J. (2017). Diversity in big data: A review. *Big data*, 5(2):73–84.

Fagin, R., Kumar, R., and Sivakumar, D. (2003). Efficient similarity search and classification via rank aggregation. In *ACM SIGMOD*, pages 301–312.

Hetland, M. (2009). The Basic Principles of Metric Indexing. In *Swarm Intell. for Multi-objective Problems in Data Mining*, pages 199–232. Springer.

Jain, A., Sarda, P., and Haritsa, J. R. (2004). Providing diversity in k-nearest neighbor query results. In *CKDM*, pages 404–413. Springer.

Pestov, V. (2013). Is the k-nn classifier in high dimensions affected by the curse of dimensionality? *Computers & Mathematics with Applications*, 65(10):1427–1437.

Pouyanfar, S., Yang, Y., Chen, S.-C., Shyu, M.-L., and Iyengar, S. (2018). Multimedia big data analytics: A survey. *ACM CSUR*, 51(1):1–34.

Santos, L., Oliveira, W., Ferreira, M., Cordeiro, R., Traina, A., and Traina Jr, C. (2013a). Evaluating the diversification of similarity query results. *JIDM*, 4(3):188–188.

Santos, L., Oliveira, W., Ferreira, M., Traina, A., and Traina Jr, C. (2013b). Parameter-free and domain-independent similarity search with diversity. In *SSDBM*, pages 1–12.

Smyth, B. and McClave, P. (2001). Similarity vs. diversity. *PICCR*, 1(1):347–361.

Vieira, M., Razente, H., Barioni, M., Hadjieleftheriou, M., Srivastava, D., Traina Jr., C., and Tsotras, V. (2011). On query result diversification. In *ICDE*, pages 1163–1174.

Yu, C., Lakshmanan, L. V., and Amer-Yahia, S. (2009). Recommendation diversification using explanations. In *ICDE*, pages 1299–1302. IEEE.

Zheng, K., Wang, H., Qi, Z., Li, J., and Gao, H. (2017). A survey of query result diversification. *Knowledge and Information Sys.*, 51(1):1–36.