DiVE: Diversifying View Recommendation for Visual Data Exploration

Rischan Mafrur
The University of Queensland
Queensland, Australia
r.mafrur@uq.edu.au

Mohamed A. Sharaf
The University of Queensland
Queensland, Australia
m.sharaf@uq.edu.au

Hina A. Khan
The University of Queensland
Queensland, Australia
h.khan3@uq.edu.au

ABSTRACT

To support effective data exploration, there has been a growing interest in developing solutions that can automatically recommend data visualizations that reveal important data-driven insights. In such solutions, a large number of possible data visualization views are generated and ranked according to some metric of importance, then the top-k most important views are recommended. However, one drawback of that approach is that it often recommends similar views, leaving the data analyst with a limited amount of gained insights. To address that limitation, in this work we posit that employing diversification techniques in the process of view recommendation allows eliminating that redundancy and provides a concise coverage of the possible insights to be discovered. To that end, we propose a hybrid objective utility function, which captures both the importance, as well as the diversity of the insights revealed by the recommended views. While in principle, traditional diversification methods provide plausible solutions under our proposed utility function, they suffer from a significantly high query processing cost. In particular, directly applying such methods leads to a "process-first-diversify-next" approach, in which all possible data visualization are generated first via executing a large number of aggregate queries. To address that challenge, we propose the *DiVE* scheme, which efficiently selects the top-k recommended view based on our hybrid utility function. DiVE leverages the properties of both the importance and diversity metrics to prune a large number of query executions without compromising the quality of recommendations. Our experimental evaluation on real datasets shows the performance gains provided by DiVE.

CCS CONCEPTS

• **Information systems** → **Database management system engines**; **Database query processing**;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '18, October 22–26, 2018, Torino, Italy
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271744

KEYWORDS

Data Exploration; Visual Analytics; Data Diversification

ACM Reference Format:
Rischan Mafrur, Mohamed A. Sharaf, and Hina A. Khan. 2018. DiVE: Diversifying View Recommendation for Visual Data Exploration. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271744

1 INTRODUCTION

The need for effective visual data exploration is gaining wider recognition, where automated solutions are provided to support users from professional data analysts in industry and science to data enthusiasts who lack formal training in data analytics. The goal is to enable data-driven discoveries, wherein interesting insights are unearthed from large volumes of collected data. As such, recent years have seen the introduction of many visual analytic tools (e.g., Tableau, Qlik, and Spotfire). These tools aim to provide aesthetically high-quality visualizations in terms of charts, which are essentially aggregated views of the underlying data (e.g., bar charts). For instance, the commercial Tableau visualization tool presents users with aggregate charts, with the expectation that some of those charts would reveal insights that a user finds interesting. Clearly, however, manually looking for insights in each visualization is labor-intensive and time-consuming.

Such challenge motivated multiple research efforts that focused on automatic recommendation of visualizations based on some metrics that capture the utility of a recommended visualizations (e.g., [4, 10, 13, 16–18, 21–23]). For instance, recent case studies have shown that a *deviation-based* formulation of that utility is able to provide analysts with interesting visualizations that highlight some of the particular trends of the analyzed datasets [4, 5, 21, 22]. Differently from Tableau's user-driven approach, in that deviation-based data-driven approach, certain views of a query result (i.e., *target view*) are recommended if they deviate significantly from those exhibited by a reference dataset (i.e., *reference view*). The intuition is that a view with high deviation is expected to reveal some important insights that are very particular to the data subset under analysis.

For instance, consider a data analyst trying to gain some insights into the *Cleveland heart disease* dataset[1]. Naturally, a first step in that exploratory analysis is to conduct some comparison between patients with heart disease and those

[1] http://archive.ics.uci.edu/ml/datasets/heart+Disease

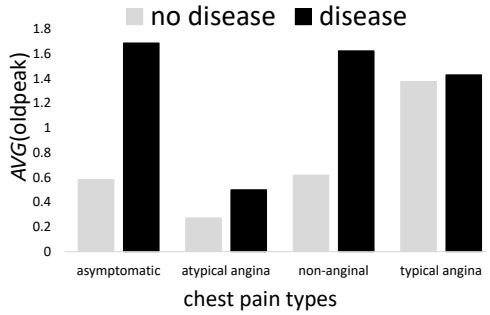


Figure 1: AVG(oldpeak) vs. chest pain types

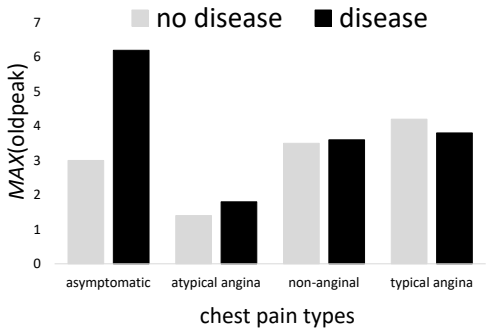


Figure 2: MAX(oldpeak) vs. chest pain types

without heart disease. Hence, the analyst writes an SQL query that selects patients with heart disease (i.e., `disease`) as the target data subset for analysis, and the remaining patients are selected as the reference data subset (i.e., `no-disease`).

Since the analyzed data contains different dimensions (e.g., chest pain types, sex, etc.) and different measures (oldpeak, age, etc.), it is a challenging task for the analyst to manually select the combinations of dimensions and measures that reveal interesting insights. Hence, to automatically recommend interesting bar chart visualizations, different SQL aggregate functions are applied on the views generated from all possible pairwise combinations of dimensions and measures, then the most *important* views are presented to the analyst. For this example, Figure 1 shows the top-1 recommended view according to the deviation-based metric [21, 22]. The figure shows that an aggregate view (i.e., bar chart) based on *average oldpeak* (i.e., pressure of the ST segment, where ST segment is an isoelectric section of the ECG) vs. *chest pain types* exhibits a large deviation between the target view (`disease`) and reference view (`no-disease`). That is, patients with heart disease often suffer more from asymptomatic and non-angina chest pains, in comparison to those without heart disease.

While recommending views based on their importance has been shown to reveal some interesting insights [4, 21, 22], such approach still suffers from a "tunnel vision", where it often recommends similar and redundant views. For instance, Figure 2 shows the second top recommended view for the analysis described above. Comparing Figures 2 and 1, it is easy to see that both views are based on the same dimension (i.e., *chest pain types*) and the same measure (i.e., *oldpeak*), and the only difference between them is the aggregate function (i.e., `MAX` vs `AVG`). Despite that similarity between the two views, they are still both recommended to the analyst due to the high deviation between target and reference views.

To address that limitation, in this work we posit that employing *diversification* techniques in the process of view recommendation allows eliminating that redundancy and provides full coverage of the possible insights to be discovered. In fact, diversity is rapidly becoming one of the fundamental features for maximizing information gain in web search and recommendation engines (e.g., [2, 15, 26, 27]). Similarly, it is highly desirable to recommend views that reveal interesting insights, while at the same time provide the analyst with a broad scope of those insights.

To that end, we propose a hybrid objective utility function, which captures both the importance (i.e., a deviation-based metric), as well as the diversity of insights revealed by the recommended views. The main goal is to select and recommend top-k views that balance the tradeoff between importance and diversity based on the hybrid objective function.

In principle, traditional data diversification methods that consider both relevance and diversity can be directly applied in the context of our problem to maximize the overall objective function (e.g., [15, 26, 27]). However, differently from assessing relevance, evaluating the importance of a view is a computationally expensive operation, which requires the execution of rather data-intensive queries. As such, directly applying those methods leads to a "process-first-diversify-next" approach [14], in which all possible data visualizations are generated first via executing a large number of aggregate queries. To address that challenge and minimize the incurred query processing cost, we propose an integrated scheme called *DiVE*, which leverages the properties of both the importance and diversity to prune a large number of low-utility views without compromising the quality of recommendations. The main contributions of this paper are summarized as follows:

- We formulate the problem of recommending views that are both important and diverse based on a hybrid objective function, which balances the tradeoff between the *content* and the *context* of recommended view (**Section 3**).
- We propose the novel *DiVE* schemes, which employ several algorithms to select the recommended visualizations based on our hybrid ranking/objective function (**Section 4**).
- We propose novel optimization techniques that leverage the salient characteristics of our objective function to minimize the query processing cost incurred in view recommendation while maximizing the quality of recommendation (**Section 5**).
- We conduct an extensive experimental evaluation on real datasets, which compare the performance of various algorithms and illustrate the benefits achieved by *DiVE* (**Sections 6**).

2 PRELIMINARIES AND RELATED WORK

Several recent research efforts have been directed to the challenging task of recommending aggregate views that reveal interesting data-driven insights (e.g., [4, 21, 22]). As in previous works, we assume a similar model, in which a visual data exploration session starts with an analyst submitting a query Q on a multi-dimensional database D_B. Essentially, Q selects a subset D_Q from D_B by specifying a query predicate T. Hence, Q is defined as: Q: `SELECT * FROM` D_B `WHERE` T;

Ideally, the analyst would like to generate some aggregate views (e.g., bar charts or scatter plots) that unearth some valuable insights from the selected data subset D_Q. However, achieving that goal is only possible if the analyst knows exactly what to look for! That is, if she knows the parameters, which specify some aggregate views that lead to those valuable insights (e.g., aggregate functions, grouping attributes, etc.). Hence, the goal of existing works, such as [4, 13, 21–23], is to *automatically* recommend such aggregate views.

To specify and recommend such views, as in previous works, we consider a multi-dimensional database D_B, which consists of a set of dimensional attributes $\mathbb{A}$ and a set of measure attributes $\mathbb{M}$. Also, let $\mathbb{F}$ be a set of possible aggregate functions over measure attributes. Hence, specifying different combinations of dimension and measure attributes along with various aggregate functions, generates a set of possible views $\mathbb{V}$ over the selected dataset D_Q. For instance, a possible aggregate view V_i is specified by a tuple $<A_i, M_i, F_i>$, where $A_i \in \mathbb{A}$, $M_i \in \mathbb{M}$, and $F_i \in \mathbb{F}$, and it can be formally defined as: V_i: `SELECT` A_i, F_i `(`M_i`) FROM` D_B `WHERE` T `GROUP BY` A_i;

Manually looking for insights in each view $V_i \in \mathbb{V}$ is a labor-intensive and time-consuming process. Particularly, the number of views to explore is equal to: $|\mathbb{V}| = |\mathbb{A}| \times |\mathbb{M}| \times |\mathbb{F}|$, where $|\mathbb{F}|$ is the number of SQL aggregate functions, $|\mathbb{A}|$ and $|\mathbb{M}|$ are the number of attributes and measures. Such challenge motivated multiple research efforts that focused on automatic recommendation of views based on some metrics that capture the utility of a recommended view [4, 10, 13, 16–18, 21–23].

Those approaches can be broadly classified as *user-driven* or *data-driven*. User-driven solutions focus on recommending views that facilitate a particular user intent or task. For example, VizDeck [13] utilizes user feedback as a basis for view recommendation, whereas Profiler [10] detects anomalies and recommends views based on mutual information metric. Similarly, Rank-by-Feature Framework [18] enables users to select their criterion for ranking histograms and scatter-plots.

Meanwhile, *data-driven* focus on enabling the discovery of interesting insights from large volumes of data without requiring much prior knowledge of the data. Towards that end, data-driven metrics are employed to capture the *interestingness* or *importance* of a recommended view. Recent case studies have shown that a *deviation-based* metric is effective in providing analysts with *important* views that highlight some of the particular trends of the analyzed datasets [4, 5, 21, 22].

In particular, the deviation-based metric measures the distance between target view $V_i(D_Q)$ and reference view $V_i(D_R)$. That is, it measures the deviation between the aggregate view

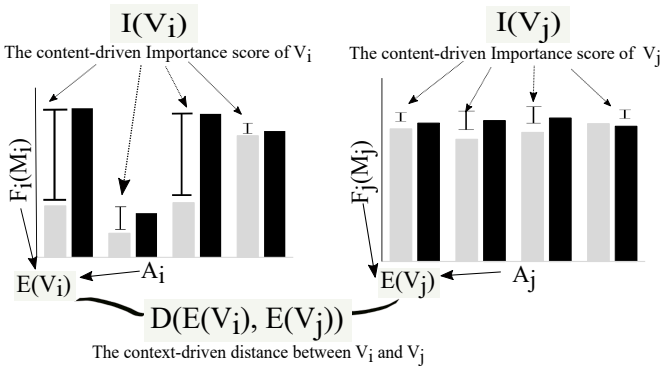


Figure 3: Content vs. Context of views.

V_i generated from the subset data D_Q vs. that generated from a reference dataset D_R, where $V_i(D_Q)$ is denoted as *target* view, whereas $V_i(D_R)$ is denoted as *reference* view. That reference dataset could be the whole database (i.e., $D_R = D_B$) or a selected subset of the database (e.g., $D_R =$ `no-disease`, as described in Sec. 1). The premise underlying the deviation-based metric is that a view V_i that results in a high deviation is expected to reveal some important insights that are very particular to the subset D_Q and distinguish it from the patterns in D_R. In case, $D_R = D_B$, then the patterns extracted from D_Q are fundamentally different from the generals ones manifested in the entire database D_B.

While recommending views based on their importance has been shown to reveal some interesting insight, it also suffers from the drawback of recommending similar and redundant views, which leaves the data analyst with a limited scope of possible insights. As illustrated in Sec. 1, Figures 1 and 2 show two recommended views that basically reveal the same insight. To address that limitation, in this work we posit that employing *diversification* techniques [2, 3, 8, 14, 15, 24, 27] in the process of view recommendation allows eliminating that redundancy and provides a good and concise coverage of the possible insights to be discovered. In the next section, we discuss in details the formulation of both importance and diversity, and their impact on the view recommendation process.

3 DIVERSIFYING RECOMMENDED VIEWS

Towards formulating our hybrid objective for view recommendation, in this section we describe the *content-based* deviation metric for assessing the importance of each view (Sec. 3.1), together with our *context-based* measure of the (dis)similarity between different views (Sec. 3.2). Those two metrics provide the foundations for our hybrid objective that aims to balance the tradeoff between importance and diversity in view recommendation (Sec. 3.3).

3.1 Content-Driven Importance

As described in the previous section, we adopt a deviation-based metric to quantify the importance of a view [21, 22].

Essentially, the deviation-based metric compares an aggregate view generated from the selected subset dataset D_Q (i.e., target view $V_i(D_Q)$) to the same view if generated from a reference dataset D_R (i.e., reference view $V_i(D_R)$).

Clearly, the deviation between a target and a reference view is a *data-driven* metric. That is, it measures the deviation between the *result* of $V_i(D_Q)$ and that of $V_i(D_R)$. Consequently, from a visualization point of view, that deviation is a *content-based* metric that captures the difference between the content of the visualization generated by $V_i(D_Q)$ vs. the visual content of $V_i(D_R)$. In the next, we formally describe the standard computation of that data-driven content-based metric, whereas the discussion of its counterpart context-driven metric is deferred to the next section.

To calculate the content-based deviation, each target view $V_i(D_Q)$ is normalized into a *probability distribution* $P[V_i(D_Q)]$ and similarly, each reference view into $P[V_i(D_R)]$. In particular, consider an aggregate view $V_i =< A_i, M_i, F_i >$. The result of that view can be represented as the set of tuples: $<(a_1, g_1), (a_j, g_j), ..., (a_t, g_t)>$, where t is the number of distinct values (i.e., groups) in attribute A_i, a_j is the j-th group in attribute A_i, and g_j is the aggregated value $F_i(M_i)$ for the group a_j [4, 22]. Hence, V_i is normalized by the sum of aggregate values $G = \sum_{j=1}^{t} g_j$, resulting in the probability distribution $P[V_i] =< \frac{g_1}{G}, \frac{g_2}{G}, ..., \frac{g_t}{G} >$.

Finally, the importance score of V_i is measured in terms of the distance between $P[V_i(D_Q)]$ and $P[V_i(D_R)]$ (as illustrated in Figure 3), and is simply defined as:

$$I(V_i) = dist\left(P\left[V_i(D_Q)\right], P\left[V_i(D_R)\right]\right) \quad (1)$$

where $I(V_i)$ is the importance score of V_i and *dist* is a distance function. Similar to existing work [21, 22], we adopt a Euclidian distance, but other distance measures are also applicable (Earth Mover's distance, K-L divergence, etc.).

In current approaches for view recommendation, the importance value $I(V_i)$ of each possible view V_i is computed, and the k views with the highest deviation are recommended. However, in this work, our goal is to ensure that recommended views provide a good coverage of possible insights, which is described next.

3.2 Context-Driven Similarity

As mentioned above, recommending views based only on their data content often leads to a set of similar views. In order to provide full coverage of all possible interesting insights, in this work, we posit that achieving *diversity* within the set of recommended views is an essential quality measure. Before discussing the details of diversity computation in Sec. 3.3, it is important to notice that central to that computation is some notion of distance measure between data objects. Existing work provides multiple metrics for measuring that distance between traditional data objects, such as web documents (e.g., [2, 15, 27]), database tuples (e.g, [20]), etc. However, our work in this paper is the first to consider diversity in the context of aggregate data visualizations. As such, a metric is needed to quantify the (dis)similarity between the distinct features of different visualizations. Towards this, we re-emphasize that each visualization is merely a data view generated by an aggregate query. Thus, such metric naturally lends itself to considering the query underlying each view (i.e., the query executed to create the view). In turn, the distance between two views is measured based on the distance between their underlying queries. Hence, in addition to our data-driven content-based deviation, we also introduce a query-driven *context-based* deviation metric. Figure 3 illustrates and summarizes our proposed metrics.

Towards measuring the context-based deviation, we extend on existing work in the area of query recommendation and refinement (e.g., [1, 11, 12, 20]). In that work, the distance between two range queries q_1 and q_2 is mapped to that of measuring the edit distance needed to transform q_1 into q_2. In the context of our work, however, views are generated from aggregate queries without range predicates. In particular, a view is fully defined in terms of a combination of attribute, measure and an aggregate function. Hence, in addition to the content of a view V_i which is described by its probability distribution (i.e., $P(V_i)$) as defined in Sec. 3.1), we also consider the context of the view $E(V_i)$, which is defined in terms of the query underlying V_i as: $E(V_i) = \{A_i, M_i, F_i\}$.

Such definition of view context leads to a special case of the existing work on query recommendation [11, 12, 20], in which the normalized distance between two queries is simply measured using the *Jaccard* similarity measure. Hence, the Jaccard similarity between two aggregate views V_i and V_j is measured as: $J\left(V_i, V_j\right) = \frac{|E(V_i) \cap E(V_j)|}{|E(V_i) \cup E(V_j)|}$

We note that the jaccard similarity assigns equal weights to each of the element in a set. Accordingly, when applied to aggregate views, then two views with the same attribute and different measure and aggregate function will have the same similarity score as any other pair of views with same measure but different attribute and aggregate function. However, an analyst may consider two views with the same attribute A_i more similar than two views with same measure attribute M_i. To allow the analyst to specify such preference, each contextual component of a view is associated with a weight that specifies its impact on determining the (dis)similarity between views. Specifically, for any view V_i, let $w(e)$ be the weight assigned to the e^{th} context component of $E(V_i)$. Since, $E(V_i)$ is a set of three components $\{A_i, M_i, F_i\}$, then $\sum_{e=1}^{3} w(e) = 1$. Accordingly, the similarity between any two views V_i and V_j is measured as:

$$J(V_i, V_j) = \frac{\sum_{e \in E(V_i) \cap E(V_j)} w(e)}{\sum_{e \in E(V_i) \cup E(V_j)} w(e)}$$

Consequently, the context-based deviation between V_i and V_j is calculated as:

$$D\left(V_i, V_j\right) = 1 - J\left(V_i, V_j\right) \quad (2)$$

3.3 Problem Definition

In this section, we formally define our problem for recommending diversified interesting aggregate views. Towards this, we first define the metrics to measure the performance of our proposed visualization recommendation system.

3.3.1 Hybrid Objective Function. Given set of all possible views $\mathbb{V}$, our goal is to recommend set $S \subseteq \mathbb{V}$, where $|S| = k$. Our hybrid objective function is designed to consider both the importance and diversity of the recommended views. Particularly, it integrates two components: 1) the total importance score of set S, and 2) the diversity score of S.

The importance score of S is calculated as the average value of the importance measure of each view in S, as given in Eq. 1. Hence, the total importance score of S is defined as:

$$I(S) = \sum_{i=1}^{k} \frac{I(V_i)}{I_u}, V_i \in S,$$

where I_u is the upper bound on the importance score for an individual view, which is achieved when for each group a_i, the corresponding value $\frac{g_i}{G}$ in $P[V_i(D_R)]$ or $P[V_i(D_Q)]$ is zero. Thus, $I_u = \sqrt{2}$, and is used to normalize the average importance score for set S.

In order to measure the diversity of a set of objects, several diversity functions have been employed in the literature [2, 24]. Among those, previous research has mostly focused on measuring diversity based on either the average or the minimum of the pairwise distances between the elements of a set [25]. In this work, we focus on the first of those variants (i.e., average), as it maximizes the coverage of S. Hence, given a distance metric $D\left(V_i, V_j\right)$, as given in Eq. 2, the diversity of a set S can be simply measured as follows:

$$f(S, D) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j>i}^{k} D\left(V_i, V_j\right), V_i, V_j \in S$$

Since the maximum context-based deviation between any two views in Eq. 2 is 1.0, then dividing the sum of distances by $k(k-1)$ ensures that the diversity score of set S is normalized and bounded by 1.0.

Putting it together, for a set of views $S \subseteq V$, our hybrid objective function is formulated as the linear weighted combination of the importance score, $I(S)$ and diversity score $f(S, D)$, and is defined as:

$$F(S) = (1 - \lambda) \times I(S) + \lambda \times f(S, D) \quad (3)$$

where $0 \leq \lambda \leq 1$ is employed to control the preference given to the importance and diversity components. For instance, a higher value of λ results in a set of more diverse views, whereas a lower value of λ generates a set of important views, which might exhibit some redundancy.

Hence, our goal is to find an optimal set of views S^*, which maximizes the objective function $F(S)$, and is formally defined as follows:

Definition 1. **Recommending diversified important views:** *Given a target subset D_Q and a reference subset D_R, the goal is to recommend a set $S \subseteq \mathbb{V}$, where $|S| = k$, and $\mathbb{V}$ is the set of all possible target views, such that the overall hybrid objective $F(S)$ is maximized.*

3.3.2 Cost of Visualization Recommendation. Exisiting research has shown that recommending aggregate data visualizations based on data-driven content-based deviation is a computationally expensive task [4, 21, 22]. Moreover, integrating diversification into the view recommendation problem, as described above, further increases that computational cost. In particular, the incurred processing cost includes the following two components: 1) Query processing cost C_Q: measured in terms of the time needed to execute and compare all the queries underlying the set of target views as well as their corresponding reference views (i.e., content-based deviation), and 2) View diversification cost C_D: measured in terms of the time needed to compute all the pairwise distances between each pair of target views (i.e., context-based deviation). Consequently, the total cost C_T for recommending a set of views is simply defined as: $C_T = C_Q + C_D$.

In principle, traditional data diversification methods that consider both relevance and diversity can be directly applied in the context of our problem to maximize the objective function Eq. 3. For instance, in the context of recommending web search, such methods are designed to recommend a set of diversified objects (e.g., web documents) that are relevant to the user needs (e.g., [2, 15, 27]), database tuples (e.g., [20]), etc. However, in that setting, the relevance of an object is either given or simply computed. To the contrary, in our setting for view recommendation, the importance of a view is a computational expensive operation, which requires the execution of a target and reference view. As such, directly applying those methods leads to a "process-first-diversify-next" approach [14], in which all possible data visualization are generated first via executing a large number of aggregate queries. To address that challenge and minimize the incurred query processing cost, next we propose our *DiVE* scheme, which leverages the properties of both the importance and diversity to prune a large number of low-utility views, without compromising the quality of recommendations.

4 THE DiVE SCHEMES

In this section, we first discuss two simple baseline solutions for view recommendation (Sec. 4.1). Then, we present our DiVE schemes for recommending diversified top views, as captured by Eq. 3. Towards this, we first expand on the well-known *Greedy* heuristic and propose our DiVE-Greedy scheme (Sec. 4.2), whereas our *Swap*-based DiVE scheme is introduced in Sec. 4.3.

4.1 Baseline Solutions

As baseline solutions to compare the performance of our proposed DiVE schemes, we simply incorporate methods from existing work that optimize either for importance or diversity. In terms of diversity, we employ the classical *Greedy Construction* algorithm [19], which has been shown to maximize diversity within reasonable bounds compared to the optimal solution [24, 26]. In this work, we refer to that baseline as *Greedy-Diversity*. Similarly, in terms of importance, we adopt the work on `SeeDB` for recommending the top-k views with

Algorithm 1: DiVE-Greedy

Input: Set of views $\mathbb{V}$ and result set size k
Output: Result set $S \subseteq \mathbb{V}$, |S| = k

$S \leftarrow \left[V_i, V_j\right]$ get two most distant views
$X \leftarrow [\mathbb{V} \backslash S]$
$i \leftarrow len\,(S)$
while |S| < k **do**
  $X_i \leftarrow argmax((1-\lambda) \times I\,(X_i) + \lambda \times setDist\,(X_i, S))$
  $S.add\,(X_i)$
  $X.remove\,(X_i)$
end
return S

the highest deviation [21, 22]. Particularly, in that method, all possible target and reference views are generated by executing their underlying queries, then the list of views is linearly scanned to recommend the top-k for which the target view shows high deviation from its corresponding reference view (denoted as *Linear-Importance* in this work).

Clearly, those two methods are "oblivious" to our hybrid objective function (i.e., Eq.3). Moreover, as expected and shown in our experimental evaluation (Sec. 6), Greedy-Diversity provides its best performance in terms of effectiveness when λ = 1.0 (i.e., all preference is given to diversity), whereas Linear-Importance is the winner when λ = 0.0 (i.e., all preference is given to importance). Next, we present our DiVE schemes which are able to provide the best performance, irrespective of the value of λ.

4.2 The DiVE-Greedy Scheme

In this section, we discuss our first DiVE scheme (*DiVE-Greedy*), which simply extends the basic Greedy Construction algorithm to work under our hybrid objective function (i.e., Eq. 3). Such extension is straightforward and is described in Algorithm 1. Similar to the classical Greedy Construction, DiVE-Greedy initializes the set S with the two most distant views, where the distance between any two views is calculated using our context-based function, as given in Eq.2. Then, DiVE-Greedy iteratively selects new views to be added to S. Particularly, in each iteration a view is selected from the set of remaining views X and is added to S. To make that selection, DiVE-Greedy assigns a score to each view in X, which is based on the hybrid objective function $F\,(S)$, as defined in Eq. 3. Specifically, the *utility score* assigned to a view $X_i \in X$ is computed as:

$$U\,(X_i) = (1-\lambda) \times I\,(X_i) + \lambda \times setDist\,(X_i, S) \qquad (4)$$

where $setDist\,(X_i, S) = \frac{1}{|S|} \sum_{\substack{j=1 \\ V_j \in S}}^{|S|} D\left(V_i, V_j\right)$ Thus, in each iteration, the view with highest utility score is selected and added to S, until $|S| = k$, as shown in Algorithm 1.

DiVE-Greedy Cost: Notice that the only difference between DiVE-Greedy and our baseline Greedy-Diversity (i.e., the classical Greedy algorithm) is in the utility score assigned to each view (i.e., $U(X_i)$ in Eq.4). In fact, in the special case where λ = 1.0, Eq. 4 boils down to $U\,(X_i) = setDist\,(X_i, S)$, which is the same score used by Greedy-Diversity for maximizing diversification. However, that simple change in the utility score leads to executing the query underlying each view X_i in order to compute the $(1-\lambda) \times I\,(X_i)$ component of its score. Hence, the overall cost of DiVE-Greedy is $C_T = C_Q + C_D$, as opposed to the cost of Greedy-Diversity, which is only $C_T = C_D$, where C_Q is the query processing cost (i.e., data-driven), and C_D is the cost for computing Jaccard distances (i.e., query-driven), as described in Sec. 3. Clearly, C_Q is equal to the number of possible views and is $O(n)$, where n is the number of possible views, whereas C_D is $O(kn)$, where k is the number of recommended views.

Algorithm 2: DiVE-Swap

Input: Set of views $\mathbb{V}$ and result set size k
Output: Result set $S \subseteq \mathbb{V}$, |S| = k

$S \leftarrow$ set of maximum importance or maximum diversity
$X \leftarrow [\mathbb{V} \backslash S]$
$F_{current} \leftarrow 0$
$improve \leftarrow True$
while $improve = True$ **do**
  for X_i *in set* X **do**
    $S' \leftarrow S$
    for S_j *in set* S **do**
      if $F\left(S'\right) < F\left(S \backslash S_j \cup X_i\right)$ **then**
        $S' \leftarrow S \backslash S_j \cup X_i$
      end
    end
    if $F\left(S'\right) > F\,(S)$ **then**
      $S \leftarrow S'$
    end
  end
  if $F\,(S) > F_{current}$ **then**
    $F_{current} \leftarrow F\,(S)$
    $improve \leftarrow True$
  else
    $improve \leftarrow False$
  end
end
return S

4.3 The DiVE-Swap Scheme

The DiVE-Greedy algorithm presented in the previous section is of the constructive type. That is, it starts with an empty set of views and incrementally constructs it by adding one view at a time. To the contrary, our DiVE-Swap presented in this section falls under the *local search* type of algorithms. In general, a local search algorithm starts out with a complete initial solution and then attempts to find a better solution in the neighborhood of that initial one. Like constructive algorithms, local search algorithms are also widely used in solving optimization problems including diversification. For instance, the Swap local search method has been utilized to maximize diversity [3, 8, 24], and in this paper, we further expand it to our DiVE schemes.

The basic idea underlying DiVE-Swap is to start with an initial set S of size k and then iteratively modify the set S in order to improve the value of the objective function $F(S)$. One of the main design criteria in local search algorithms is the choice of the initial solution. In DiVE-Swap, we consider two natural variants: 1) *DiVE-iSwap*, and 2) *DiVE-dSwap*. In DiVE-iSwap, S is initialized with the k views that maximize importance and can be easily obtained using our baseline Linear-Importance (Sec. 4.1). Alternatively, in DiVE-dSwap, S is initialized with the k views that maximize diversity using Greedy-Diversity (Sec. 4.1).

Apart from the initialization approach, both variants work similarly. Particularly, in each iteration, each unselected view $X_i \in X$ is interchanged with all views in S (Algorithm 2 line 9). That is, the overall hybrid objective function is computed as $F(S \backslash S_j \cup X_i)$. Then the one interchange that leads to the highest new value for F is applied and S is updated accordingly (Algorithm 2 line 14). Such iterations are repeated until no more views can be swapped between X and S, which is reached when no further improvement is achieved in the value of F (Algorithm 2 line 17).

In comparison to DiVE-Greedy, DiVE-Swap incurs the same query processing cost C_Q. Furthermore, it incurs even higher C_D cost for computing diversity, which can reach up to $O\left(kn^2\right)$. However, DiVE-Swap offers a valuable opportunity for maximizing the number of pruned views, and in turn reducing the query processing cost C_Q, as described in the next section.

5 THE DiVE SCHEMES WITH PRUNING

As described above, both DiVE-Greedy and DiVE-Swap execute all the underlying queries for each view X_i in X. However, only a small fraction of those views is actually included in the final top-k recommended set. Consequently, a significant amount of query processing cost is incurred for generating low-utility views. Thus, in this section, we propose efficient techniques for pruning such low-utility views without incurring the high cost for evaluating their importance score (Sec. 5.1 and Sec. 5.2). Moreover, we also propose an adaptive pruning method based on *non-parametric predictive intervals* (Sec. 5.3). Such method is able to balance the tradeoff between minimizing the number of executed queries and maximizing the quality of recommendation.

5.1 Pruning for DiVE-Swap

In Sec. 4.3, we presented two variants of DiVE-Swap: 1) *DiVE-iSwap*, and 2) *DiVE-dSwap*. While those two variants incur the same cost, they offer substantially different performance when combined with our pruning techniques. Particularly, consider DiVE-iSwap, which is initialized with the k views that maximize importance. To select those views, all possible views have to be generated first, which in turn requires processing all their corresponding queries. Hence, DiVE-iSwap simply eliminates all opportunities for pruning as all views are executed in the initialization phase. To the contrary, DiVE-dSwap is initialized with the k views that maximize diversity. To select that initial set, no query execution is needed and the processing is limited to computing the context-based deviation distances, which incurs a significantly lower processing cost compared to query execution. Hence, DiVE-dSwap provides a valuable opportunity for pruning low-utility views.

Recall that under DiVE-dSwap, in each iteration a view $X_i \in X$ is selected to replace a view $S_j \in S$. The criterion for that selected view is to improve $F(S)$. That is, $F(S \backslash S_j \cup X_i) > F(S)$. Hence, the task is to find that *top-1* pair of views $< X_i, S_j >$ that provides the maximum improvement in $F(S)$ once interchanged. Without pruning, that requires iterating through S and X simultaneously and computing F for each pair, which requires processing and generating each view in X. To avoid such expensive processing and enable pruning, the following steps are taken.

A list L is created for all possible swap pairs $< X_i, S_j >$, where L is sorted based on the diversity achieved if the swap is to be made. Notice that up to this point the only processing needed is to compute diversity without any query execution to evaluate the importance of any X_i. Given that setting, the task is clearly similar to top-k query processing, for which numerous optimization tehniques are proposed (e.g.,[6, 9]). Particularly, to find the *top-1* view, each view X_i is initially assigned an importance equal to the upper bound I_u(Sec. 3.3). In turn, the upper bound of $F(S)$ achieved by X_i is computed as: $maxF(S \backslash S_j \cup X_i)$, which is based on the actual diversity achieved by the swap, and the upper bound on importance. As such, $maxF(S \backslash S_j \cup X_i)$ is compared against $F(S)$, leading to one of the following two cases: If $maxF(S \backslash S_j \cup X_i) > F(S)$, then the swap $< X_i, S_j >$ can "potentially" improve $F(S)$. Hence, at that stage the view X_i needs to be generated in order to evaluate its actual importance $I(X_i)$. Otherwise, the pair $< X_i, S_j >$ is pruned if:

$$maxF(S \backslash S_j \cup X_i) < F(S).$$

Simply put, if the upper bound $maxF$ achieved by that swap is still less than the current $F(S)$, then the actual $F(S \backslash S_j \cup X_i)$ is guaranteed to be less than $F(S)$ and the pair $< X_i, S_j >$ can be safely ignored. More importantly, since the L is sorted by diversity, then the next views are also guaranteed to provide no improvement and that iteration of DiVE-dSwap reaches *early termination*. Hence, for all the remaining views no query processing is needed, which significantly reduces the overall cost.

5.2 Pruning for DiVE-Greedy

The pruning technique described above is directly applicable to DiVE-Greedy. Particularly, recall that under DiVE-Greedy, in each iteration the view with the highest utility score is added to the partial set S until $|S| = k$. As described in Eq. 4, such utility score $U(X_i)$ is a weighted sum of two measures: 1) the importance score of X_i (i.e., $I(X_i)$), and 2) the distance of X_i from S (i.e., $setDist(X_i, S)$). Thus, while the goal in DiVE-dSwap is to find the pair $< X_i, S_j >$ that provides the maximum improvement in $F(S)$, the goal for DiVE-Greedy is to find the view with the highest utility $U(H)$. Hence, similar

to DiVE-dSwap, a list L of all views in X is created such that each X_i is assigned an importance equal to the upper bound I_u and L is sorted based on the diversity score of each view X_i to the current set S. Then, the highest utility $U(H)$ is initialized to a default value of 0.0, and the list L is traversed in order. For each visited view X_i, the upper bound on the utility achieved by X_i (i.e., $maxU(X_i)$) is computed using its actual diversity score and the upper bound on its importance. If $maxU(X_i) > U(H)$, then X_i is generated and its actual utility $U(X_i)$ is calculated. Accordingly, if $U(X_i) > U(H)$, then $U(H)$ is set to be equal to $U(X_i)$. However, if $maxU(X_i) < U(H)$, then early termination is reached.

5.2.1 DiVE-dSwap vs. DiVE-Greedy. At this point, it is especially important to examine and contrast the pruning power achieved by each of DiVE-Greedy and DiVE-dSwap. For DiVE-Greedy, recall that pruning is attained for those views where: $maxU(X_i) < U(H)$. However, since DiVE-Greedy is a constructive algorithm, the set S is incrementally constructed iteration by iteration until $|S| = k$. Hence, in the first iterations S has a very small number of selected views with a minimum of $|S| = 2$. Naturally, when S is a small set, then most of the remaining unselected views in X are expected to exhibit high diversity, since the majority of them will be very dissimilar from the small set of views in S. Accordingly, for most of the views in X, the value $maxU(X_i)$ will be relatively high, due to achieving high score on diversity. Hence, most views will fail to satisfy the pruning condition and are consequently executed incurring high query processing cost.

To the contrary, DiVE-dSwap is initiated with a set S of k diverse views. Hence, it will initially have a reasonably high $F(S)$. Moreover, many views in X will be "close" to some view in S. Hence, the swaps that involve those views will score low on diversity, and in turn low $maxF$. Since pruning happens when $maxF(S \backslash S_j \cup X_i) < F(S)$, the combination of those two factors above (i.e., high F and low $maxF$) allows for many views satisfying the pruning condition, which improves the pruning power of DiVE-dSwap. That pruning power can be further improved by relaxing the assumption about maximum importance I_u, as described next.

5.3 Predictive Interval for Adaptive Bounds

In general, both the pruning schemes provided by DiVE-Greedy and DiVE-dSwap rely on the fundamental idea of evaluating the upper bound of the benefit provided by a view V_i towards the objective F. If that maximum benefit is still not enough to consider V_i to join S, then V_i is pruned and its query processing cost is saved. Moreover, to evaluate that upper bound, both schemes compute the actual diversity offered by V_i and instead of computing its actual importance, it is substituted with the maximum attainable importance score I_u. Naturally, overestimating $I(V_i)$ leads to overestimating its benefit and consequently limited pruning power is achieved. Meanwhile, for most datasets, I_u is in fact an overestimation of $I(V_i)$. Hence, our goal in this section to provide a tighter bound on $I(V_i)$, which allows for maximum pruning while maintaining the quality of the solution.

Recall that I_u is achieved when for each group a_i, the corresponding value $\frac{g_i}{G}$ in $P[V_i(D_R)]$ or $P[V_i(D_Q)]$ is zero. Hence, I_u is a theoretical bound for the maximum importance achieved by any view in any dataset. For most real datasets, however, that condition is rarely satisfied and the actual upper bound I_{au} is typically much smaller than I_u. Meanwhile, a hypothetical pruning scheme that utilizes that actual upper bound I_{au} is expected to deliver more pruning power than the schemes using the theoretical upper bound I_u, especially when $I_{au} \ll I_u$. In practice, however, that hypothetical scheme is not achievable since obtaining the value I_{au} requires executing all the possible views, which is clearly in conflict with the goal of pruning.

Accordingly, rather than using overestimated I_u or obtaining the actual I_{au}, our goal is to estimate I_{au} with high accuracy and minimum number of query executions. In particular, given the set of possible views $\mathbb{V}$, the goal is to estimate the maximum importance $\bar{I}_{au}$ given by some view in $\mathbb{V}$. However, estimating the maximum value of a population is known to be a challenging problem, as opposed to estimating other statistics such as average or sum [7]. That challenge is further emphasized when the values exhibited by the population are skewed and do not follow a typical normal distribution, which is typically the case for the importance value of views.

Thus, instead of estimating I_{au}, we rely on *non-parametric predictive interval* models to determine its value with certain level of confidence without any assumption on the population [7]. To apply that model, some sample views are executed and the maximum importance observed in that sample is recorded as $\bar{I}_{au}$. To determine the number of samples, a *Predictive Interval (PI)* is to be defined, such that: $PI = \frac{(m-1)}{(m+1)}$, where m is the number of samples.

For instance, setting $m = 19$, results in PI = 90%. That is, 90% of the time, the importance value of an unseen view V_i will be less than the maximum importance seen so far. Clearly, the higher the PI value, the higher the accuracy of $\bar{I}_{au}$, but also requires executing more views. In this work, we find that a value of $PI = 97\%$ is able to strike a fine balance between minimizing the number of executed queries and maximizing the objective F, as shown next.

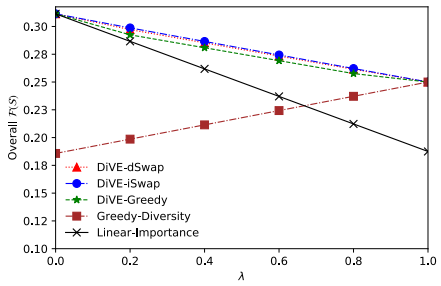
6 EXPERIMENTAL EVALUATION

Table 1 summarizes the different parameters used in our evaluation (default values are in bold). We conducted our experiments over the following datasets: 1) Heart Disease Dataset [2]: This dataset is comprised of 9 dimensional attributes and 5 measure attributes, using four aggregate functions, resulting in a total of $9 \times 5 \times 4 = 180$ possible views, and 2) Airline (Flights) Dataset [3]: This dataset is comprised of 7 dimensional attributes and 4 measure attributes for a total of $7 \times 4 \times 4 = 112$ possible views. While its dimensionality

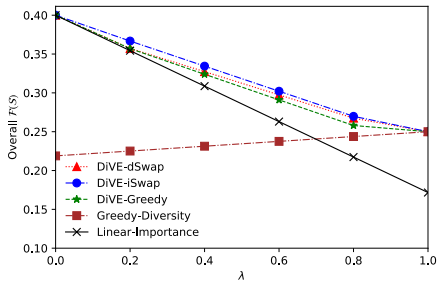
[2] http://archive.ics.uci.edu/ml/datasets/heart+Disease
[3] http://stat-computing.org/dataexpo/2009/the-data.html

Table 1: Parameters testbed in the experiments

Parameter	Range (**default**)
datasets	**Heart disease**, Flights
diversity weight ratio	**3**(A) **: 2**(M) **: 1**(F)
tradeoff weight λ	0.0, 0.2, 0.4, **0.5**, 0.6, 0.8, 1.0
result set (size of k)	**5**, 15, 25, 35
prediction interval %	80 , 85, 90, 95, **97**, 98



(a) Heart disease dataset



(b) Flights dataset

Figure 4: Impact of λ on $F(S)$, k = 5

is lower than the heart disease data, it is a relatively large dataset of almost one million tuples, which helps in evaluating the incurred query processing time. For each experiment, the performance measures are averaged over a query workload of ten random queries submitted to select ten different subsets.

The impact of λ on F: Figure 4 shows how the performance of each scheme in terms of $F(S)$ is effected as the value of λ varies from 0 to 1. Clearly, for the lower values of λ, the highest $F(S)$ is achieved by Linear-Importance. To the contrary, the Greedy-Diversity method achieves highest values of $F(S)$ as the λ approaches 1. Hence, there is a crossover between the two schemes. However, our proposed DiVE schemes have stable performance for all values of λ and outperforms Linear-Importance and Greedy-Diversity.

Execution time evaluation In this experiment, we measure the cost of *DiVE* schemes. Figure 5 plots the execution time for Flights dataset with $k = 5$ and $\lambda = 0.5$. The total execution time is split into the query execution time C_Q and the diversification cost C_D. It is clear from Figure 5 that the total execution time C_T is dominated by the cost of generating

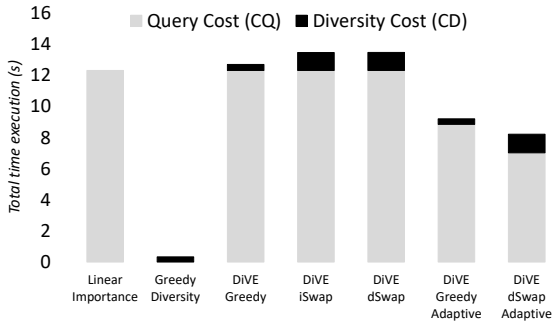


Figure 5: Cost of DiVE on Flights dataset, k =5, λ = 0.5

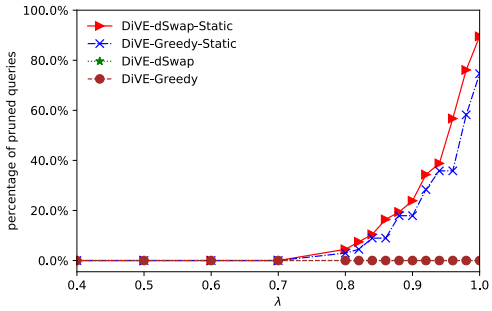


Figure 6: Impact of Static pruning

the views C_Q. Hence, the minimum cost is incurred by the Greedy-Diversity which only computes diversity. For other methods without adaptive pruning, the C_Q is same as all views are generated only once. However, the cost of diversification C_D is slightly higher for DiVE-iSwap and DiVE-dSwap as compared to the DiVE-Greedy due the higher number of iterations. Moreover, the Figure also shows the cost of both schemes with adaptive pruning. It shows that adaptive pruning can reduce the C_Q cost significantly.

Impact of static Pruning In this experiment, we present the performance of our proposed pruning techniques in terms of the number of pruned queries. The higher number of pruned queries result in the higher cost savings in the total query execution time. Figure 6 shows the performance of our static pruning technique using the theoretical upper bound of importance score I_u. In this and next experiments, DiVE-iSwap is not evaluated as it executes all view queries for the initial set selection and any pruning afterwards is not possible. Moreover, due to space limit, we use only the heart disease dataset in the next experiments. For both schemes DiVE-Greedy-Static and DiVE-dSwap-Static, since the I_u value far from the actual importance scores of individual views, the percentage of pruned queries is 0 for lower values of λ. Only for λ close to 0.9 some queries get pruned. For instance, at $\lambda = 0.9$, DiVE-Greedy-Static prunes almost 20% queries while DiVE-dSwap-Static prunes 25% queries.

Impact of Adaptive Pruning In this experiment, we analyze the performance of adaptive pruning technique under different values of λ and prediction interval PI. As shown in Figure 7, DiVE-Greedy-Adaptive is able to prune more queries compared to DiVE-Greedy-Static (Figure 6). It is also able to

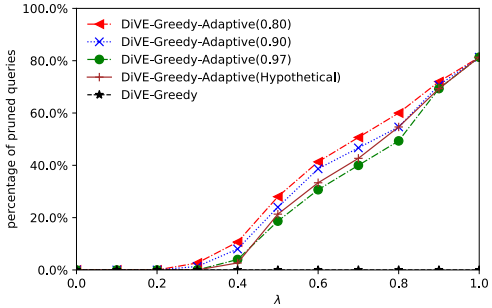


Figure 7: DiVE-Greedy with Adaptive Pruning

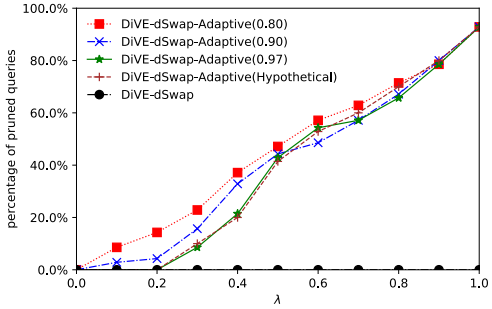


Figure 8: DiVE-dSwap with Adaptive Pruning

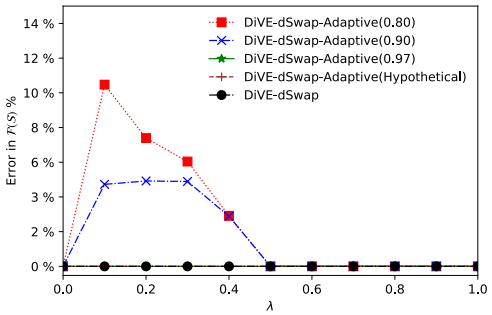


Figure 9: Impact of Adaptive pruning on $F(S)$

prune even for lower λ. The number of queries pruned increase significantly for higher values of λ. Figure 8 shows the performance of DiVE-dSwap-Adaptive with different values of PI. In comparison to DivE-Greedy-Adaptive, the number of pruned queries by DiVE-dSwap-Adaptive are much higher for all values of λ. The interesting observation is the fact that DiVE-dSwap-Adative is able to prune 15% queries for $\lambda = 0.2$. For higher values of λ the percentage of pruned queries is between 60% and 90%. Similar to DiVE-Greedy-Adaptive, highest number of queries are pruned for $PI = 0.80$.

Further, we evaluate the effectiveness of adaptive pruning in terms of $F(S)$. Figure 9 shows the loss on $F(S)$ in comparison to $F(S)$ achieved by Hypothetical methods. The loss for DiVE-dSwap-Adaptive is 0% for $PI = 0.97$, With a larger sample size the accuracy of approximated importance score is higher. For a smaller sample size of $PI = 0.80$, there is 0% loss while $\lambda = 0$ because at the moment there are no pruned queries. However, there is a maximum loss of 10% at $\lambda = 0.1$. The loss on $F(S)$ decrease as λ increases as the impact of importance score becomes smaller in the hybrid objective function. Meanwhile, starting $\lambda \geq 0.5$ the loss is 0%.

7 CONCLUSIONS

In this work, we propose the *DiVE* scheme for view recommendation in visual data exploration. DiVE combines importance and diversity into a hybrid utility function to provide full coverage of the possible insights to be discovered. Moreover, DiVE also leverages the properties of both the importance and diversity metrics to prune a large number of query executions without compromising the quality of recommendations.

Acknowledgments: This research is partially supported by the Indonesia Endowment Fund for Education (LPDP) and an Advanced Queensland Research Grant.

REFERENCES

[1] A. M. Albarrak and M. A. Sharaf. 2017. Efficient schemes for similarity-aware refinement of aggregation queries. *World Wide Web* 20, 6 (2017), 1237–1267.
[2] C. L. A. Clarke et al. 2008. Novelty and diversity in information retrieval evaluation. In *SIGIR*.
[3] M. Drosou and E. Pitoura. 2010. Search result diversification. *SIGMOD Record* 39, 1 (2010), 41–47.
[4] H. Ehsan et al. 2016. MuVE: Efficient Multi-Objective View Recommendation for Visual Data Exploration. In *ICDE*.
[5] H. Ehsan et al. 2018. Efficient Recommendation of Aggregate Data Visualizations. *TKDE* 30, 2 (2018), 263–277.
[6] R. Fagin et al. 2003. Comparing top k lists. In *ACM-SIAM*.
[7] Y. Hu et al. 2009. Estimating aggregates in time-constrained approximate queries in Oracle. In *EDBT*.
[8] Z. Hussain et al. 2015. Diversifying with Few Regrets, But too Few to Mention. In *ExploreDB*.
[9] I. F. Ilyas et al. 2008. A survey of top-k query processing techniques in relational database systems. *ACM Comput. Surv.* 40, 4 (2008), 11:1–11:58.
[10] S. Kandel et al. 2012. Profiler: integrated statistical analysis and visualization for data quality assessment. In *AVI*.
[11] V. Kantere. 2016. Query Similarity for Approximate Query Answering. In *DEXA*.
[12] V. Kantere et al. 2015. Query Relaxation across Heterogeneous Data Sources. In *CIKM*.
[13] A. Key et al. 2012. VizDeck: self-organizing dashboards for visual analytics. In *SIGMOD*.
[14] H. A. Khan and M. A. Sharaf. 2015. Progressive diversification for column-based data exploration platforms. In *ICDE*.
[15] D. Rafiei et al. 2010. Diversifying web search results. In *WWW*.
[16] T. Sellam et al. 2016. Ziggy: Characterizing Query Results for Data Explorers. *PVLDB* 9, 13 (2016), 1473–1476.
[17] T. Sellam and M. L. Kersten. 2016. Fast, Explainable View Detection to Characterize Exploration Queries. In *SSDBM*.
[18] J. Seo and B. Shneiderman. 2006. Knowledge Discovery in High-Dimensional Data: Case Studies and a User Survey for the Rank-by-Feature Framework. *TVGC* 12, 3 (2006), 311–322.
[19] B. Smyth et al. 2001. Similarity vs. Diversity. In *ICCBR*.
[20] Q. T. Tran and C. Y. Chan. 2010. How to ConQueR why-not questions. In *SIGMOD*.
[21] M. Vartak et al. 2014. SEEDB: Automatically Generating Query Visualizations. *PVLDB* 7, 13 (2014), 1581–1584.
[22] M. Vartak et al. 2015. SEEDB: Efficient Data-Driven Visualization Recommendations to Support Visual Analytics. *PVLDB* 8, 13 (2015), 2182–2193.
[23] F. B. Viegas et al. 2007. Many Eyes: A site for visualization at internet scale. *TVGC* (2007), 1121–1128.
[24] M. R. Vieira et al. 2011. On query result diversification. In *ICDE*.
[25] E. Wu et al. 2014. The Case for Data Visualization Management Systems. *PVLDB* 7, 10 (2014), 903–906.
[26] C. Yu et al. 2009. It takes variety to make a world: diversification in recommender systems. In *EDBT*.
[27] M. Zhang and N. Hurley. 2008. Avoiding monotony: improving the diversity of recommendation lists. In *RecSys*.