

# Constrained recommendations for query visualizations

Ibrahim A. Ibrahim<sup>1</sup> · Abdullah M. Albarrak<sup>1</sup> ·  
Xue Li<sup>1</sup>

Received: 10 December 2015 / Revised: 24 August 2016 / Accepted: 26 September 2016 /  
Published online: 6 October 2016  
© Springer-Verlag London 2016

**Abstract** The improvement of data storage and data acquisition techniques has led to huge accumulated data volumes in a variety of applications. International research enterprises such as the Human Genome and the Digital Sky Survey Projects are generating massive volumes of scientific data. A major challenge with these datasets is to glean insights from them to discover patterns or to originate relationships. The analysis of these massive, typically messy, and inconsistent volumes of data is indeed crucial and challenging in many application domains. Hence, the research community has introduced a number of visualizations tools to guide and help analysts in exploring the data space to extract potentially useful information. However, when working with high-dimensional datasets, identifying visualizations that show interesting variations and trends in data is not trivial: the analyst must manually specify a large number of visualizations, explore relationships among various attributes, and examine different subsets of data before discovering visualizations that are interesting or insightful. Though, exploring all possible visualizations involves complex challenges. It is a costly and time-consuming process especially when the dimensionality is high. Furthermore, the rapid growth of databases becomes multifaceted in their channels and dimensionality; thus, the transition from static analysis to real-time analytics represents a fundamental paradigm shift in the field of Big Data. Motivated by the above challenges, we propose an efficient framework called *real-time scoring engine (RtSEngine)* that assists analysts to limit the exploration of visualizations for a specified number of visualizations and/or certain execution time quote to recommend a set of visualizations that meet analysts' budgets. To achieve that, *RtSEngine* incorporates our proposed approaches to prioritize and score attributes that form all possible visualizations in a dataset based on their statistical properties such as

---

✉ Ibrahim A. Ibrahim  
i.ibrahim@uq.edu.au

Abdullah M. Albarrak  
a.albarrak@uq.edu.au

Xue Li  
xueli@itee.uq.edu.au

<sup>1</sup> School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia

selectivity, data distribution, and number of distinct values. Then, *RtSEngine* recommends the visualizations created from the top-scored attributes. Moreover, we present visualizations cost-aware techniques that estimate the retrieval and computation costs of each visualization so that analysts may discard high-cost visualizations. We show and evaluate the effectiveness and efficiency of our proposed approaches, and assess the quality of visualizations and the overhead obtained by applying our techniques on both synthetic and real datasets.

**Keywords** Query visualization · Aggregate queries · Visual analytics

## 1 Introduction

Data visualization is one of the most common tools for identifying trends and finding anomalies in Big Data. However, with high-dimensional datasets, identifying visualizations that effectively present interesting variations or patterns in the data is a non-trivial task: analysts typically build a large number of visualizations optimizing for a range of visualization types, appealing features, and more before arriving at one that shows something valuable.

For datasets with large number of dimensions, it is extremely exhaustive for analysts to manually study all the dimensions; hence, interactive data visualization needs to be boosted with automated visualizations recommendation techniques. Interactive visualization analytics tools such as Tableau, ShowMe, and Fusion Tables [9,24,31] provide some features for automatically recommending the best visualization for a dataset. However, these features are restricted to a set of esthetic rules (e.g., color, fonts, and styles) that guide which visualization is most appropriate.

Profiler [18] is another visualization tool which explores all data space to detect anomalies in data and recommends the best binning for the horizontal  $x$ -axis of a visualization. It decides which granularity is appropriate to bin on to depict the most interesting relationships among data. Profiler [18] maintains a data cube in memory and uses it to support rapid user interactions. While this approach is possible when the dimensionality and cardinality are small, it cannot be used with large tables and ad hoc queries with high-dimensional data, which is the norm of scientific databases.

In the biomedical data analysis domain, INVISQUE [11,34] was proposed as a visual sense-making system to support information analysis for medical diagnosis. INVISQUE illustrates the similarity between the information analysis during intelligence analysis and medical diagnosis based on a sense-making loop and a data-frame model. To overcome the challenges of exploring high-dimensional patients data, SubVIS [13] was recently proposed as a visualization tool to interactively explore biomedical data by utilizing subspace analysis algorithms to cluster data into subclusters and show the relationships that exist among them.

Another example of tools that recommend visualizations is VizDeck [19]. VizDeck recommends visualizations based on the statistical properties of small datasets and adopts a card game metaphor to help organize the recommended visualizations into interactive visual dashboard.

For large-scale datasets, SeeDB [33] was proposed to automatically recommend interesting visualizations based on distance metrics which compute deviations among the probability distributions of the visualizations. SeeDB presents different levels of optimizations to decrease the latency and maintain the quality of visualizations such as sharing computations and combined query executions.

Table: GoCard											
Dimension Attributes									Measure Attributes		
Trip_ID	Operators	Operations_date	run_ID	Route	Service_no	Direction	Boarding_stop	Alighting_stop	Trip length	Passengers no	Passengers by_route
1	Brisbane Transport	3-Jan-13	2360	P384	7394	Outbound	Leichhardt St - The Ridge (Stop 164)	West Ashgrove Waterworks Road	9.1	1	205
2	Surfside Buslines	4-Jan-13	715	715	2228	Inbound	Town Centre Drive	Labrador Park Shopping Centre	12.8667	1	1354
3	Surfside Buslines	4-Jan-13	703	703	1239	Inbound	Turpin Road	Harbourtown Stop B	5.56667	2	1175

**Fig. 1** Snippet from the GoCard relational database schema with a representative sample. Each row represents one trip with a bus, a ferry, or a train with 12 dimensions describing the details of that trip. The database’s dimensions are classified into two: dimension attributes and measure attributes, in order to generate meaningful 2-dimensional visualizations, e.g., bar charts

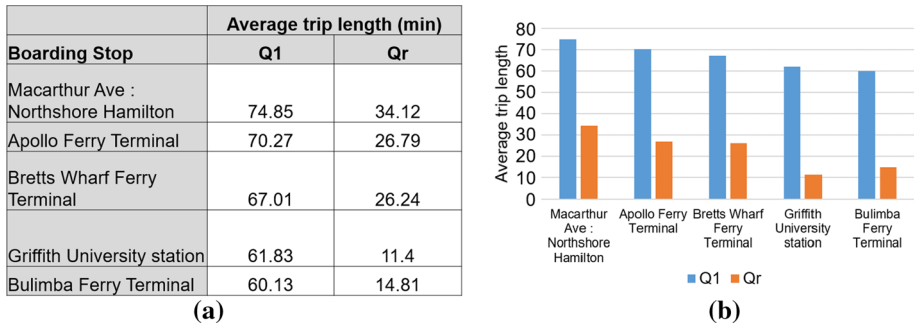
Although these analytic tools present various approaches and measures to assess the interestingness of data, they still have to explore all possible visualizations to recommend a subset of interesting visualizations. Exploring the entire data space and all visualizations is almost impossible with the limited time and resources, especially when data are growing in both the dimensionality and cardinality. As a result, shifting from static analytics to real-time analytics is essential because of the rapid data accumulation when compared with a constant human cognitive capacity. Indeed this is a challenging problem. An interactive visualizations recommendation tool needs to explore the data space intelligently by discounting unnecessary visualizations and recommend only the essential ones while preserving the quality of the results.

The following example illustrates the need for an automatic visualizations technique to identify interesting visualizations from a real, large, and structured database called GoCard which represents trips details of the public transportation system of the Brisbane city in Australia. Figure 1 shows a snippet of the GoCard database schema and a small sample from the database out of the 4.4 million tuples. Each tuple is a record that represents a trip using either a bus, a ferry, or a train, with 12 dimensions describing that trip with more details.

*Example 1* Consider a transportation analytic team that is undertaking a study for a particular alighting stop: *University of Queensland (UQ)*. This stop has received a lot of passengers complaints due to poor performance; hence, it is being investigated by the team. Suppose that the team uses the GoCard database to generate 2-dimensional visualizations (e.g., bar charts) which summarize all recorded trips using different dimensions, then search for the ones that appear to explain the increase in received complaints. To accomplish that, an analyst would begin by using a program’s GUI or a custom query language to execute the equivalent of the following SQL query and pull all data from the database for the alighting stop UQ:

```
Q = SELECT * FROM GoCard
WHERE alighting stop = "University of Queensland";
```

Next, the analyst would use an interactive GUI interface to generate all possible visualizations of the query result. For instance, the analyst may visualize average trip length grouped by route, total daily passengers grouped by direction, maximum trip length by boarding stop, and so on. Hence, the analyst would manually study all these visualizations to find



**Fig. 2** Average trips length in minutes by boarding stop of  $Q_1$  and the reference  $Q_r$  result into high utility value, i.e., high *interestingness*. **a** Sample results of query  $Q_1$  and  $Q_r$ , **b** 2-D bar chart for query  $Q_1$  and  $Q_r$

interesting insight or visualizations that might explain the reason behind the increase in complaints. Indeed, exploring and studying all visualizations is challenging especially for high-dimensional datasets. Hence, an automatic visualization recommendation technique should show the analyst the most interesting visualization based on the alighting stop UQ.

Consider the visualization for the average trip length by boarding stop: it is generated by running an operation equivalent to the following SQL query:

```
Q1 = SELECT boarding stop, AVG(trip length) FROM GoCard
      WHERE alighting stop = "University of Queensland" GROUP
      BY boarding stop;
```

Figure 3 shows the visualization of  $Q_1$ 's result. Consequently, the visualization in Fig. 3 happened to be the most interesting visualization. The reason is when  $Q_1$ 's result is compared with entire data, it depicts long average trip length in some boarding stops which travels toward UQ that are significantly different from the equivalent average of the trip lengths (equals 17.6min) in the entire dataset. Specifically,  $Q_1$ 's result is compared against the following reference query  $Q_r$ :

```
Qr = SELECT boarding stop, AVG(trip length) FROM GoCard
      GROUP BY boarding stop;
```

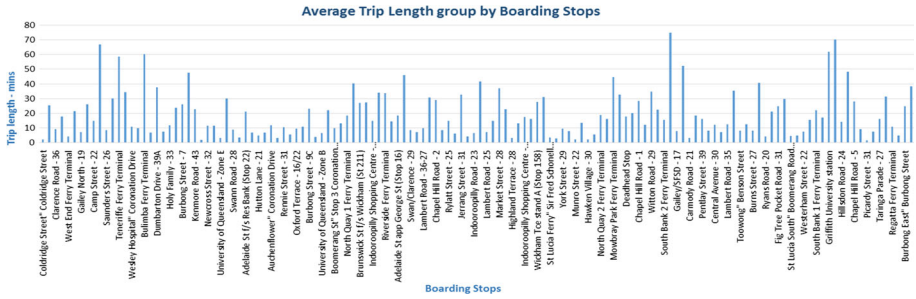
Figure 2a, b shows a sample results of  $Q_1$  and  $Q_r$ . □

Example 1 above suggests that visualizations which portray trend deviations from a reference are potentially remarkable and of high interest.

Here, the average trip length grouped by boarding stops (Fig. 3) is considered as the top interesting visualization, among other visualizations such as total daily passengers grouped by direction, and maximum trip length grouped by boarding stop. The reason is it depicts long average trip length in some boarding stops which travels toward UQ that are significantly different from the equivalent average of the trip lengths (equals 17.6 min) in the entire dataset. As listed in Fig. 2a, ferry terminals scored longer trips to UQ than bus stops because ferries often take longer waiting times among stops than buses.

We summarize our contributions as follows:

- Proposing a new problem which address the limitation of current visualizations recommendation tools. Particularly, we include budget constraints to automatically recommend top- $K$  interesting visualizations according to an input query within the specified budget.



**Fig. 3** 2-Dimensional bar chart visualization generated by query  $Q_1$ . The x-axis represents the boarding stop, while the y-axis represents the average trip lengths in minutes, toward University of Queensland stop

- Designing an efficient framework called *real-time scoring engine (RTSEngine)* that limits the exploration of visualizations by assessing their priorities of the recommended views according to their deviation utilities and costs.
- Proposing efficient algorithms which utilize statistical features of the views such as number of distinct values, selectivity ratios, and data distribution, to early prioritize the views.
- Proposing efficient algorithms to approximate the retrieval and computations costs of the generated visualizations and evaluate their estimated costs against their deviation utilities to recommend high-accuracy views in the specified budgets.
- Conducting extensive experiments that demonstrate the efficiency and effectiveness of our proposed algorithms on real and synthetic dataset.

This paper is organized as follows: Sect. 2 describes related works on query visualization. Then, Sect. 3 provides preliminary details on recommendation of query visualizations and presents our problem statement. Then, we present our framework *RTSEngine* in Sect. 4 that contains two main modules: Priority Evaluator and Cost Estimator, which recommend a set of visualizations efficiently within the specified constraints. Section 5 shows experiment results for our proposed algorithms on two real datasets.

## 2 Related work

Interactive data visualization tools have interested the research community over the past few years, and it has presented a number of interactive data analytics tools such as ShowMe, Polaris, and Tableau [6, 18, 19, 24]. Similar visualization specification tools have also been introduced by the database community, including Fusion Tables [9] and the Devise [22] toolkit. Unlike SeeDB, which recommends visualizations automatically by exploring the entire views space, these tools place the onus on the analyst to specify the visualization to be generated. For datasets with a large number of attributes, it is unfeasible for the analyst to manually study all the attributes; hence, interactive visualization needs to be augmented with automated visualization techniques.

A few recent systems have attempted to automate some aspects of data analysis and visualization. Profiler is one such automated tool that allows analysts to detect anomalies in data [18]. Another related tool is VizDeck [19], in given a dataset, which depicts all possible 2-D visualizations on a dashboard that the user can control by reordering or pinning visualizations. Given that VizDeck generates all visualizations however, it is only meant for

small datasets, and VizDeck does not discuss techniques to speed up the generation of these visualizations.

To support visual sense making in medical diagnosis, INVISQUE [11,34] is an interactive visualization system proposed such as physical index cards on a two-dimensional workspace. INVISQUE provides some features to support annotating, revisiting, and merging two clusters. It discusses essential problems in designing medical diagnostic displays that can improve the review of a patient's medical history [11]. A recent work, SubVIS [13] is a visualization tool which assists the user to analyze and interactively explore computed subspaces to discover insights into highly dimensional and complex patient's datasets. SubVIS [13] introduces an analysis workflow to visually explore subspace clusters from various perspectives, and it tackles some subspace clustering challenges such as difficulty of interpretation patient results, redundancy detection in subspaces and clusters, and multiple clustering results for different parameter settings.

Statistical analysis and graphing packages such as R, SAS, and MATLAB could also be used generate visualizations, but they lack the ability to filter and recommend 'interesting' visualizations.

OLAP: there has been some work on browsing data cubes, allowing analysts to variously find *explanations* for why two cube values were different, to find which neighboring cubes have similar properties to the cube under consideration, or get suggestions on what unexplored data cubes should be looked at next [15,27,28].

Database Visualization Work: Fusion tables [9] allow users to create visualizations layered on top of Web databases; they do not consider the problem of automatic visualization generation. Devise [10] translated user-manipulated visualizations into database queries.

Although the aforementioned approaches provide assistance in query visualization, they lack the ability to automatically recommend interesting visualizations, except SeeDB which provides different optimization techniques to automatically recommend interesting visualizations while avoiding unnecessary visualizations by utilizing two kinds of optimization techniques as explained next.

*Visualizations pruning in SeeDB* SeeDB implemented an execution engine to reduce latency in assessing the collection of aggregate views which it applies two kinds of optimizations: sharing, where aggregate view queries are combined to share computation as much as possible, and pruning, where aggregate view queries corresponding to low-utility visualizations are dropped from consideration without scanning the whole dataset. SeeDB developed a phased execution framework; each phase operates on a subset of the dataset. Phase  $i$  of  $n$  operates on the  $i$ th of  $n$  equally sized partitions of the dataset. The execution engine begins with the entire set of aggregate views as follows: during phase  $i$ , the SeeDB [33] modifies partial results for the views still under consideration using the  $i$ th fraction of the dataset. The execution engine applies sharing-based optimizations to minimize scans on this  $i$ th fraction of the dataset. At the end of phase  $i$ , the execution engine uses pruning-based optimizations to determine which aggregate views to discard. The partial results of each aggregate view on the fractions from 1 through  $i$  are used to estimate the quality of each view, and the views with low utility are discarded.

The execution engine uses pruning optimizations to determine which aggregate views to discard. Specifically, partial results for each view based on the data processed so far are used to estimate utility and views with low utility are discarded. SeeDB execution engine supports two pruning schemes. The first uses confidence-interval techniques to bound utilities

of views, while the second uses multi-armed bandit allocation strategies to find top utility views.

- *Confidence-interval-based pruning* The first pruning scheme uses worst-case statistical confidence intervals to bound views utilities. This technique is similar to top-k-based pruning algorithms developed in other contexts [29]. It works as follows: during each phase, it keeps an estimate of the mean utility for every aggregate view  $V_i$  and a confidence interval around that mean. At the end of a phase, it applies the following rule to prune low-utility views: if the upper bound of the utility of view  $V_i$  is less than the lower bound of the utility of  $k$  or more views, then  $V_i$  is discarded.
- *Multi-armed bandit pruning* Second pruning scheme employs a multi-armed bandit strategy (MAB) [2,33]. In MAB, an online algorithm repeatedly chooses from a set of alternatives over a sequence of trials to maximize reward. This variation is identical to the problem addressed by SeeDB: the goal is to find the visualizations (arms) with the highest utility (reward). Specifically, SeeDB adapts the Successive Accepts and Rejects algorithm from [2] to find arms with the highest mean reward. At the end of every phase, views that are still under consideration are ranked in order of their utility means. Then, it computes two differences between the utility means:  $\Delta l$  is the difference between the highest mean and the  $k + 1$ st highest mean, and  $\Delta n$  is the difference between the lowest mean and the  $k$ th highest mean. If  $\Delta l$  is greater than  $\Delta n$ , the view with the highest mean is accepted as being part of the top-k (and it no longer participates in pruning computations). On the other hand, if  $\Delta n$  is higher, the view with the lowest mean is discarded from the set of views in the running. [6] proves that under certain assumptions about reward distributions, the above technique identifies the top-k arms with high probability.

However, SeeDB pruning schemes experience some limitations, as they assume fixed data distribution [32,33] for sampling to estimate the utility of views and require large samples for pruning low-utility views with high guarantees. Moreover, aggregate functions MAX and MIN are not docile to sampling-based optimizations.

*Offline visualizations in SeeDB* SeeDB prunes redundant views [33]: for each table, it first determines the entire space of aggregate views. Next, it prunes all aggregate views containing attributes with 0 or low variance since corresponding visualizations are unlikely to be interesting. For each remaining view  $V_i$ , SeeDB computes the distribution for reference views on the entire dataset. The resulting distributions are then clustered based on pairwise correlation. From each cluster, SeeDB selects one view to compute as a cluster representative and store stubs of clustered views for subsequent use. At run time, the view generator accesses previously generated view stubs, removes redundant views, and passes the remaining stubs to the execution engine.

### 3 Preliminaries

In this section, we present background details on visualizations in the context of structural databases. We start by explaining how a visualization (or a view) is constructed by an SPJ SQL query. Then, we define our scope of visualizations that our framework is focused on, and how to measure the interestingness of a visualization based on a model proposed by [33] and another model that we believe is important. Then, we formally present our problem statement.

### 3.1 Background and scope

A visualization  $V_i$  is constructed by an SQL select-project-join query with a group-by clause over a database  $D$ . The attributes in a database table are classified into two sets: dimension attributes set  $A = \{a_1, a_2, \dots\}$  and measure attributes set  $M = \{m_1, m_2, \dots\}$ . The set  $F = \{f_1, f_2, \dots\}$  contains all aggregation functions. Hence, each visualization  $V_i$  is represented as a triple  $(a, m, f)$ , where  $a$  is a group-by attribute applied to the aggregation function  $f$  on a measure attribute  $m$ .

We limit our scope of visualizations on the basic components found on most 2-dimensional visualization systems such as bar charts and line charts, as they satisfy a wide range of applications requirements [17]. For instance, Fig. 2b represents a 2-D bar chart for the table in Fig. 2a.

As an example,  $V_i(D)$  visualizes the results of grouping the data in  $D$  by  $a$  and then aggregating the  $m$  values using  $f$ . This view is called the *reference view*. Consequently,  $V_i(DQ)$  represents a similar visualization applied to the result set denoted as  $DQ$  for a given user query  $Q$  and is called the *target view*. An example of a target view is shown in Fig. 3 where  $a$  is the boarding stops,  $m$  is the trip length, and  $f$  is the average aggregation function.

Any combination of  $(a, m, f)$  represents a view. Accordingly, we can define the total number of possible views as follows:

$$\text{View Space (SP)} = 2 \times |A| \times |M| \times |F| \tag{1}$$

*Example 2* Using the GoCard database in Example 1, the dimensions within that database can be classified as follows: the set of dimension attributes is  $A = \{\text{Operators, Operation date, Route, Boarding stop, Alighting stop, Direction}\}$ , while the set of measure attributes is  $M = \{\text{trip length, passengers by route, passengers no}\}$ , and the set of aggregate functions is  $F = \{\text{count, sum, avg, max, min}\}$ , as shown in Fig. 1. Therefore, the view space of GoCard database is:  $2 \times 6 \times 3 \times 5 = 180$ .

Though, in the context of Big Data,  $SP$  is potentially a very large number. Hence, there is a need to automatically score all these  $SP$  views so that exploring them become efficient and practical.

### 3.2 Views utility

Each view is associated with a *utility* value. The utility of a visualization is measured as its deviation from a reference dataset  $D_R$ . For instance, visualizations that show different trends in the query dataset (i.e.,  $DQ$ ) compared to a reference dataset  $D_R$  are supposed to have high utility. The reference dataset  $D_R$  may be defined as the entire underlying dataset  $D$ , the complement of  $DQ(D - DQ)$  or data selected by any arbitrary query  $Q'(DQ')$ .

Given an aggregate view  $V_i$  and a probability distribution for a target view  $P(V_i(DQ))$  and a reference view  $P(V_i(D_R))$ , the utility of  $V_i$  is the distance between these two normalized probability distributions. The higher the distance between the two distributions, the more likely the visualization is to be interesting and therefore higher utility value. Formally:

$$U(V_i) = S(P(V_i(DQ)), P(V_i(D))) \tag{2}$$

where  $S$  is a distance function (e.g., Euclidean distance and earth mover’s distance). In addition,  $S$  can be the Pearson’s correlation coefficient to capture interesting trends in visualizations.

Hence, the problem of visualizations recommendation is as follows [33]:



**Definition 1** Given a user-specified query  $Q$  on a database  $D$ , a reference dataset  $D_R$ , a utility function  $U()$ , and a positive integer  $K$ . Find top- $K$  aggregate views  $V_1, V_2, \dots, V_K$  that have the highest utilities among all views while minimizing total computation time.

Now, we are in place to present our problem formulation for visualization recommendations.

### 3.3 Problem formulation

Our proposed problem for visualization recommendations incorporates two limits (i.e., input parameters) to overcome the limitation of exploring all views.

**Definition 2** Given a user-specified query  $Q$  on a database  $D$ , a reference dataset  $D_R$ , a utility function  $U()$ , a positive integer  $K$ , an execution time limit  $tl$  or a views number limit  $R$  where  $K \leq R \leq SP$ . Find top- $K$  aggregate views  $V \equiv (a, m, f)$  which have maximum utilities  $U(V)$  among all possible views in the specified limits  $R$  or  $tl$  while maximizing the accuracy among all top- $K$  views chosen from all  $SP$  views.

The limits  $tl$  and  $R$  in Definition 2 are added explicitly to overcome the limitation of exploring all views. The former is a time budget that any algorithm should not exceed, while the latter is an upper bound on the number of views to be explored. For instance,  $tl$  can be set to zero, and  $R = SP$ . That is, no limit on the execution time and no limit on the number of generated views.

While those limits can be tuned by any valid value, an algorithm should output the same views as if there were no limits. This requirement makes the problem non-trivial; hence, we address it by presenting our optimization techniques encapsulated within the *RtSEngine* framework.

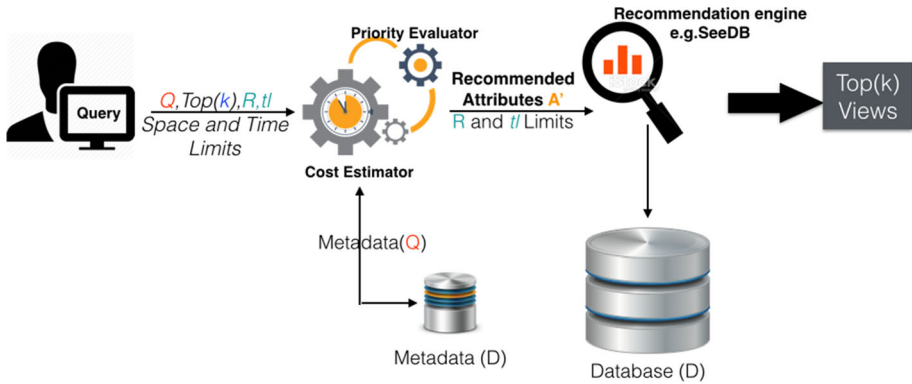
## 4 Methodology: *RtSEngine* framework

The goal of *RtSEngine* is to recommend a set of aggregate views that are considered interesting because of their abnormal deviations. To achieve that, *RtSEngine* utilizes the following key idea: recommend views that are created from grouping high-ranked dimension attributes  $A'$  within the set  $A$ . The attributes ranks in  $A'$  are computed using our proposed prioritizing techniques discussed later in the following sections. Essentially, those techniques evaluate the priorities of all dimension attributes according to their statistical features gathered from the metadata, e.g., number of selected values, data distribution, and selectivity. Then, by reordering all dimension attributes according to their priorities, only a subset of high-priority attributes are passed to the execution engine, hence limiting the number of examined views and execution time.

Conceptually, *RtSEngine*<sup>1</sup> is designed as a recommendation plug-in that can be applied to any visualization engine, e.g., Tableau and Spotfire. However, in this work, we built *RtSEngine* as a standalone end-to-end system on top of SeeDB which allows users to pose arbitrary queries over data and obtain recommended visualizations. *RtSEngine* is comprised of two main modules (see Fig. 4):

1. *Priority Evaluator* An underlying module in front of any recommendation engine. Used to evaluate the dimension attributes that form visualizations according to a priority function  $Pr$  computed using our proposed techniques.

<sup>1</sup> Implementations and data are available at: [https://github.com/ibrahimDKE/Cdb\\_RtsEngine\\_DKE\\_UQ](https://github.com/ibrahimDKE/Cdb_RtsEngine_DKE_UQ).



**Fig. 4** *RtSEngine*: real-time evaluation architecture for automatic recommendation

2. *Cost Estimator* This module is supposed to run in parallel with the Priority Evaluator to estimate the retrieval and computation costs of each visualization using our estimation approaches. Estimating the visualization costs in real time improves the efficiency by discounting high-cost and low-priority visualizations. Note that this module is an awareness cost approach which incorporates the estimated costs to assess visualizations based on their priorities and costs.

We define a notion of benefits *Benefit* ( $V_i$ ) of a view  $V_i$  as the gains from each view represented as the utility of view  $U(V_i)$ , compared with the time spent *Cost* ( $V_i$ ) to compute the view  $V_i$ . Formally:

$$Benefit(V_i) = \frac{U(V_i)}{Cost(V_i)} \tag{3}$$

Cost estimations of visualizations are discussed later in Sect. 4.2. Both modules (Priority Evaluator and Cost Estimator) read information by querying metadata to collect information about dimension attributes, e.g., number of distinct values and cardinality. Next, we describe the two modules in details.

### 4.1 Priority evaluator: dimension attributes prioritizing

In this section, we discuss the proposed approaches for prioritizing the dimension attributes in the both results set  $DQ$  and reference set (e.g., the entire dataset  $D$ ) and suggest a set of visualizations that are likely to be interesting and score high- deviation utilities in certain real-time limits such as maximum number of explored visualizations and execution time. The proposed approaches are based on our observations about the difference between the number of distinct values in the dimension attributes in the results set  $DQ$  and the entire dataset  $D$  affects on the deviation measures. In addition, other statistical features may also affect such as data distribution and selectivity; such features will be discussed in more detail in the next subsections. The following example illustrates this observation and describes how our strategies are agnostic for any recommendation system.

*Example 3* Suppose a flights database keeps flights records which contains two-dimension attributes such as destination airport name and airlines and one metric is arrival delays. Given the large size of the database (millions of records) contains 100 airports and 20 airlines companies, the analyst will study the average delays visualizations grouped by airports

and airlines using a recommendation tool, e.g., SeeDB, and comparing these views with a reference set to glean insights about all flights departure from *Origin1*. These views can be expressed as SQL queries:

- $V_1$ : select AVG (arrival delays), airport from flights where origin='Origin1' group by airport;
- $V_2$ : select AVG (arrival delays), airlines from flights where origin='Origin1' group by airlines; □

For instance, both visualizations  $V_1$  and  $V_2$  in Example 3 have the same number of distinct values: 10 destinations airports and 10 airlines operators. Eventually, aggregate views  $V_1$  and  $V_2$  will be compared to the corresponding reference views (i.e., the entire dataset  $D$ ) according to a metric. In [33] for instance, it uses a deviation-based metric that calculates the distance between the normalized distributions between the target and reference views. In our Example 3, the average arrival delays of 10 destinations airports in view  $V_1$  are evaluated against the average arrival delays of 100 destinations airports in the entire dataset  $D$ . Similarly, the average arrival delays of the 10 airlines operators in view  $V_2$  are compared against the average arrival delays of the all 20 airlines operators in the entire dataset  $D$ .

Thus, only 10 distinct values in view  $V_1$  will be compared with equivalent values in the reference view, while the remaining 90 distinct values would have no equivalent values in the target view. As a result, those remaining 90 distinct values will be compared with zeros.

Furthermore, in view  $V_2$  there are only 10 airlines operators that would be compared with zeros. This illustration arises a question about the impact of the difference in distinct values of views and their data deviations according to distance-based metrics.

Formally,  $Dval(V_i(DQ))$  is defined as the number of distinct values in a target view  $V_i$ . Consequently,  $Dval(V_i(D))$  is the number of distinct values in the corresponding reference view  $V_i$ . In Example 3,  $Dval(V_1(DQ)) = 10$  and  $Dval(V_1(D)) = 100$ . As mentioned previously, the deviation of each visualization is captured by a distance-based metric that computes the distance between two probability distributions of views. That is the deviation of a visualization  $V_i$  is its utility defined in Eq. 2:  $U(V_i) = S(P(V_i(DQ)), P(V_i(D)))$ . The distance metric  $S()$  is a distance function such as Euclidean and earth mover’s distance.

We discuss the influence of the difference in distinct values on computing the view utility  $U(V_i)$  using Euclidean distance (although our experiments are using earth mover’s distance function as the default deviation measure). As shown in Eq. 4,  $L_2$ -norm distance evaluates all aggregated values (points) in both views  $V_i(DQ)$  and  $V_i(D)$  to find the utility  $U(V_i)$ . Hence,  $V_1$ ’s utility in Example 3 is obtained by computing the  $L_2$ -norm distance between the average arrival delays (values) of destination airports (points) in  $V_1(DQ)$  and all airports in  $V_1(D)$  the entire dataset. Formally:

$$U(V_i) = \sqrt{\sum_{j=1}^n (V_i D(y_j) - V_i DQ(x_j))^2} \tag{4}$$

where  $n > 0$  is the maximum number of points among  $V_i D$  and  $V_i DQ$ . Since the compared views (i.e., target and reference view) may contain different number of distinct values, we denote  $n'$  as the number of records in  $V_i(DQ)$  and  $n''$  as the number of records in  $V_i(D)$ . Hence, we can rewrite the utility equation of view  $V_i$  as follows:

$$U(V_i) = \sqrt{\sum_{j=1}^{n'} (V_i D(y_j) - V_i DQ(x_j))^2 + \sum_{j=n'+1}^{n''} (V_i D(y_j) - 0)^2} \tag{5}$$

where  $n' < n''$ , and  $n = n' + n''$ . Because there are only  $n'$  values in the target view  $V_i DQ$ , then all subsequent points in the reference view  $V_i D$ , i.e.,  $n'' - n'$  values, would be compared with zeros. The higher the difference between distinct values in corresponding views forces much remaining values to be compared with zeros and increases the distance among views. In Example 3, the number of records  $n'$  of both target views  $V_1(DQ)$  and  $V_2(DQ)$  equals 10. However, the number of records in the reference views, i.e.,  $n''$ , is  $V_1(D) = 100$  and  $V_2(D) = 20$ .  $V_1$  is expected to show higher distance (deviation) than  $V_2$  when computing  $L_2$  norm distance because 90 airports would be evaluated to zeros in  $V_1$  but there are only 10 airlines operators with zero values in view  $V_2$ . Since every view is an aggregate group-by query over a dimension attribute as described earlier, then the number of records in each view equals the number of distinct values in the grouped dimension attribute.

Such observations can be utilized to early asses these (visualizations) views before executing the underlying queries to avoid computational costs (i.e., retrieval and deviation measure costs) by evaluating dimension attributes that contribute in creating visualizations. Furthermore, evaluating dimension attributes can also be done using other statistical properties such as selectivity and data distribution.

Further discussion of utilizing these features in our proposed approaches is presented in the next sections.

#### 4.1.1 Ranking dimension attributes based on distinct values

Scoring dimensions based on difference of distinct values is the first class of prioritizing algorithms. This approach is referred to as *Diff<sub>D</sub>Val*, and it is based on the basic observation about the number of distinct values of the dimension attributes in the results set  $DQ$  and the entire database  $D$ . The *Diff<sub>D</sub>Val* algorithm scores the dimension attributes according to the difference between the normalized distinct values of attributes in the result set  $DQ$  and the entire database  $D$ . Algorithm 1 inputs a query  $Q$ , a set of dimension attributes  $A$ , maximum views limit  $R$ , and/or execution time limit  $tl$ . Then, *Diff<sub>D</sub>Val* obtains the number of distinct values for all dimension attributes in both results sets  $DQ$  and reference dataset  $D$  by posing underlying queries to select the count of distinct values. After getting the number of distinct values, *Diff<sub>D</sub>Val* computes the priority of each dimension attribute as the difference between each normalized values. Then, *Diff<sub>D</sub>Val* sorts all dimension attributes based on their priorities. Based on Eq. 1, *Diff<sub>D</sub>Val* computes the required number of dimension attributes  $G$  that creates the limit number of views  $R$ , and then, it returns the set  $\mathcal{H}$  of size  $G$  that contains a group of high-priority attributes.

In case there is an execution time limit  $tl$ , *Diff<sub>D</sub>Val* returns an ordered set of all dimension attributes based on their priorities, and then, it passes the time limit  $tl$  to the recommendation visualization engine to limit the executions.

#### 4.1.2 Scoring dimension attributes based on selectivity

In this section, we discuss another variation of scoring the dimension attributes by capturing the data distribution in terms of query size and selectivity. Selectivity estimation is at the heart of several important database tasks. It is essential in the accurate estimation of query costs and allows a query optimizer to characterize good query execution plans from unnecessary ones. It is also important in data reduction techniques such as in computing approximated answers to queries [1, 8]. Databases have relied on selectivity estimation methods to generate fast estimates for result sizes [3, 5, 25, 26].

---

**Algorithm 1:** *Diff<sub>D</sub>Val*

---

**Input:** Attributes  $A(a_1, a_2, \dots, a_n)$ , Query  $Q$ , Views limit  $R$   
**Output:** Set  $\mathcal{H}$ : Highest priorities of dimension attributes  
 $C = \phi$  Set of all dimension attributes priorities ;  
**for**  $i = 1$  **to**  $n$  **do**  
     $DvalD(a_i) \leftarrow$  number of distinct values of  $a_i$  in  $D$  ;  
     $DvalDQ(a_i) \leftarrow$  number of distinct values of  $a_i$  in  $DQ$  ;  
     $Pr(a_i) = |DvalD(a_i) - DvalDQ(a_i)|$ ;  
     $C \leftarrow Pr(a_i)$ ;  
Sort  $C$ ;  
 $G = \lfloor \frac{R}{M \times F} \rfloor$  Calculate the required dimension number;  
**for**  $i = 1$  **to**  $G$  **do**  
     $\mathcal{H} \leftarrow C.get(i)$ ;  
**return**  $\mathcal{H}$ ;

---

The selectivity ratio [20] is defined as follows:

**Definition 3** The degree to which one value can be differentiated within a wider group of similar values.

The selectivity ratio also known as the number of distinct unique values in a column divided by its cardinality [19]. Formally, the selectivity ratio of attribute  $a_i$  is:

$$Sel_{a_i}^B = \frac{\text{Number of distinct values of } a_i \text{ in } B}{\text{Cardinality of } a_i \text{ in } B} \tag{6}$$

where  $B$  is either the result set  $DQ$  or the reference dataset  $D$ , and  $0 < Sel_{a_i}^B \leq 1$ .

For the flight database in Example 3, both the result set  $DQ$  and the reference set  $D$  have a fixed number of records, which reveals that the selectivity ratio of the airlines column is usually low because we cannot do much filtering with just the 20 values. In contrast, the selectivity ratio of the airports column is high since it has a lot of unique values. Our proposed approach *Sela* utilizes the number of distinct values in the dimension attributes and incorporates the query size to identify priorities of these dimensions by calculating a priority function  $Pr()$  for each dimension attribute. Then, *Sela* reorders the dimension attributes based on the priority.

Using selectivity ratio and the number of distinct values for assessing visualizations in  $D$  and  $DQ$  gives closer insights about the data characteristics such as the size (number of records) of aggregated views generated from group-by attributes and the uniqueness degree of data in each dimension attribute. Again, in the flights database Example 3,  $DQ$  has 10 distinct airports out of 100 airports in the airports column. This means any visualization constructed by grouping airports column in result set  $DQ$  contains only 10 aggregated records. Hence, using the query size assists on quantifying how many records would be aggregated in each view that formed from grouping a dimension attribute. However, capturing the change of both number in distinct values and the number of aggregated records in each dimension attribute in result set  $DQ$  and reference set  $D$  is essential to identify visualizations that produce high deviations among all possible visualizations. Thus, we modified the priority function  $Pr()$  in *Sela* to consider the number of records in each dimension attribute  $a_i$  and its selectivity ratio. Formally:

$$Pr(a_i) = DvalDQ(a_i) * Sel_{a_i}^{DQ} + \left( \frac{DvalDQ(a_i)}{DvalD(a_i)} \right) * Sel_{a_i}^D \tag{7}$$

The attribute priority  $Pr(a_i)$  evaluates the number of distinct values for each dimension attribute in result set  $DQ$  multiplied by its selectivity ratio. This identifies the distinct values variations and the diversity through each dimension attribute when compared with the number of records. Furthermore, the same number of distinct values is assessed in the corresponding dimension attribute of the reference set  $D$  while considering the number of records.

In *Sela*, high-priority dimension attributes are assumed to produce aggregate views (i.e., target views) that contain many groups (i.e., points) which are aggregated from records in the result set  $DQ$ . Also, the same high-priority dimension attributes are assumed to produce aggregate views (reference views) by aggregating larger number of records in reference set  $D$ . This has direct effect on the aggregated values and the number of groups in both target and reference views which is expected to score high-deviation utilities.

Although the *DiffDval* approach prioritizes dimension attributes (aggregate views) according to the number of distinct values, it is limited since it is incompetent to prioritize dimension attributes (aggregate views) when the number of distinct values remains stable in both result set  $DQ$  and reference set  $D$ . Moreover, *DiffDval* does not consider the data distribution within the attributes. To overcome this limitation, *Sela* utilizes the number of records and the selectivity ratios of dimension attributes in both datasets  $DQ$  and  $D$ .

The proposed algorithm *Sela* firstly computes the priority of each dimension attribute based on Eq. 7. Then, it sorts the dimension attributes based on the assigned priority to create a set  $\mathcal{H}$  of the top  $G$  dimension attributes. In case of execution time limit  $tl$ , *Sela* returns an ordered set of attributes with the highest priorities and passes time limit  $tl$  to the recommendation visualization engine to limit the executions.

#### 4.1.3 Prioritizing dimension attributes based on histograms

We proposed *Sela* and *DiffDVal* approaches to automatically recommend views with the highest deviations based on a priority for each dimension attributes in a star schema database  $D$ . Specifically, the proposed approaches relay on the number of the distinct values and the selectivity ratio of each dimension attribute in the compared datasets (i.e.,  $DQ$  and  $D$ ) to compute the attributes priorities.

However, the limitation of the proposed approaches is using the selectivity ratio to reflect the degree of variations of data in the dimension attributes while ignoring the distribution of data itself. In addition, it is difficult to prioritize dimensions that have the same distinct values or the same selectivity ratio.

Hence, we propose the *DimsHisto* approach which attempts to capture data distribution inside the dimension attributes by creating frequency histograms and directly measuring the distance among corresponding histograms to evaluate these dimension attributes. *DimsHisto* firstly generates frequency histograms for all dimension attributes in each dataset. Then, it computes the deviation in each dimension by calculating the normalized distances between each corresponding dimension attribute. For any star schema database  $D$ , a dimension attribute  $a_i \in A = \{a_1, a_2, \dots, a_n\}$  can be represented as two frequency histograms:  $H_{D(a_i)}$ , and  $H_{DQ(a_i)}$ . Those two histograms are created by executing the following queries:

$H_{D(a_i)}$ : Select *count* ( $a_i$ ) from  $D$  group by  $a_i$ ;

$H_{DQ(a_i)}$ : Select *count* ( $a_i$ ) from  $DQ$  group by  $a_i$ ;

Then, after normalizing these histograms, the priority of each dimension attribute is computed as the distance between these two histograms:

$$Pr(a_i) = S(H_{D(a_i)}, H_{DQ(a_i)}) \quad (8)$$

where  $S()$  is a distance metric. Eventually, the dimension attributes are sorted according to their priorities.

A constructed histogram  $H_{D(a_i)}$  is equivalent to all aggregate views created by aggregating any measure attribute (using aggregate function *Count*) and grouped by the dimension attribute  $a_i$  in the dataset  $D$ . Such a histogram assists in improving the performance of recommendation engines by avoiding the construction and computation of aggregate views along all measure attributes.

*DimsHisto* has to submit  $2 \times |A|$  queries to compute the histograms of all dimensions and the computations of the distance metric. However, this step can be optimized to only  $|A|$  by computing the histograms of all dimensions for the entire database offline. While *DimsHisto* can use any distance metric to compute the deviation among the views, we suggest to use the same metric to unify the metric of the deviations.

All proposed algorithms *DiffDVal*, *Sela*, and *DimsHisto* have the same number of queries as the cost of retrieving data. While *DimsHisto* has additional cost for distance computations, it shows high accuracy for most of the aggregate functions such as sum, avg, and count, because these functions are relative to the data frequencies. Though, *DimsHisto* is less descriptive to other aggregate functions such as Min and Max, as they are not amenable for sampling-based optimizations.

## 4.2 Cost estimator: visualizations cost estimation

The previous approaches rank dimension attributes according to their priorities and recommend visualizations while being oblivious to the retrieval and computational costs of those visualizations. However, visualizations created using different dimension and measures attributes have different retrieval and execution costs according to the query size, type of the aggregate functions, number of groups in each attribute, and the time used to compute the deviation among all values in the corresponding visualizations.

This urges the need to only generate visualizations with high deviations and avoid the computation costs of the low-deviation ones. Besides differences in deviation utilities among different visualizations, each visualization exhibits different execution and retrieval costs. Furthermore, some visualizations may take long computations and retrieval time to only yield small deviation distances. The trade-off between gaining high utilities of the visualizations and their computations and querying costs is challenging because it involves the optimizations of finding high-utility visualizations while considering their costs.

The cost estimation step is essential to determine the cost of running and computing the deviations of a visualization to evaluate its costs against the utility obtained by measuring the deviation among visualizations. To improve the performance of recommendation applications, it is vital to discard visualizations that are expected to consume much retrieval and computation time while returning low-deviation distances.

The cost estimation modules approximate CPU and I/O costs to combine them into an overall metric that is used for comparing alternative plans. The problem of choosing an appropriate technique to determine CPU and I/O costs requires considerable care. An early study [23] identified key roles for accurate cost estimation, such as the physical and statistical properties of data. Cost models take into account relevant aspects of physical design, e.g., co-location of data and index pages. However, the ability to do accurate cost estimation and propagation of statistical information on data remains one of the difficult open issues in query optimization [4].

We determine the cost of a view  $Cost(V_i)$  as the sum of the following:

- Cost of running view  $V_i(a, m, f)$  on dataset  $D$ .

- Cost of running view  $V_i(a, m, f)$  on dataset  $DQ$ .
- Computation cost of the distance function  $S(V_iDQ, V_iD)$ .

Formally:

$$Cost(V_i) = C(V_iDQ) + C(V_iD) + C(S(V_iDQ, V_iD)) \quad (9)$$

As mentioned previously, the cost of running a view  $V_i(a, m, f)$  on a database is affected by various factors. For instance, access paths and indices that are used to execute the view determine the proper execution plan, which reflects the view execution cost.

*Running cost of views*  $C(V_iDQ)$  and  $C(V_iD)$  refers to the retrieval cost of the results of both views  $V_iDQ$  and  $V_iD$  as discussed earlier.

*Computation cost of*  $C(S(V_iDQ, V_iD))$  is considered as the time spent on calculating the distance measure  $S()$  for each value in both corresponding views.

The number of points that are compared in the corresponding views  $V_iDQ$  and  $V_iD$  is the maximum number of groups (bins) among these two views, and it is denoted as  $n$ . Alternatively, it equals the maximum number of distinct values in  $V_iDQ$  and  $V_iD$  attribute dimension.

Note that the cost of distance measures vary according to their computational complexity. For example, the Euclidean distance is faster than the earth mover's ( $EMD$ ) distance function. This is because  $EMD$  has a very high complexity  $O(n^3 \log n)$  [16], while the complexity of the Euclidean distance is  $O(n)$ .

Since the computation cost depends on  $n$  and also depends on the computational complexity of the distance measure, we propose the following view cost equation:

$$C(S(V_iDQ, V_iD)) = O_d(n) \times d_t \quad (10)$$

where  $O_d$  is the complexity of the distance measure and  $d_t$  is the computation time used to compute a single point.

#### 4.2.1 Retrieval costs of visualizations

In our context, the execution cost of views can be obtained using two different methods:

- *Actual cost* actual costs of the views are obtained by executing all queries to get their exact I/O costs and calculating the deviation among the corresponding views.
- *DB estimates* reading the estimates of each view directly from the database engine (i.e., query optimizers).

However, our proposed cost method is not restricted to a certain cost estimation approach including methods based on sampling (e.g., [12,21]), histograms (e.g., [14]), and machine learning (e.g., [7,30]) which can be used to obtain the retrieval cost from independent estimation models.

Our proposed estimation algorithm *ViewsEstimate* is illustrated in Alg.4. *ViewsEstimate* takes dimensions, measures attributes, and the aggregated functions as input. Then, it estimates I/O and computation time for each view  $V_i$  for both datasets and it returns the estimated costs of each view. The estimated I/O time for each view is obtained by reading the estimation of queries from the database query optimizer or using an independent cost estimation model. Then, *ViewsEstimate* calculates the computations costs of the distance measure between the corresponding views according to equation Eq. 10 to find the total estimated cost. Afterward, *ViewsEstimate* adds up the computations cost and the I/O cost for  $V_i$  and then stores it into set  $\mathcal{S}$ .



**Table 1** Experiments parameters and their default values

Parameter	Range	Default values
Top- $K$	1–70	10, 20
Views limit $R$	1–100	70
Dimension attributes $ A $	–	9, 10
Measure attributes $ M $	–	3, 10
Aggregate functions $ F $	count, sum, avg, min, max	–

Cost Estimator utilizes the set  $S$  by defining a benefit of a dimension attribute  $Benefit(a_i)$  as the priority of  $a_i$  divided by the maximum estimated cost of any view created using dimension attribute  $a_i$ , formally:

$$Benefit(a_i) = \frac{Pr(a_i)}{Cost(a_i)} \tag{11}$$

where  $Cost(a_i)$  is the maximum estimated cost of any view created by grouping by  $a_i$ .

Finally, *Dims Estimate* ranks dimension attributes depending upon their benefits as computed by Eq. 11. As shown in Alg.4, *Views Estimate* inputs a set of dimensions and a visualization number limit  $R$ , and then, it iteratively calculates the priority and the cost of each dimension attribute to compute the benefit of each attribute. *Views Estimate* computes the number of dimension attributes  $G$  that create the limit  $R$  and then outputs a set of high *Benefit* attributes of size  $G$ .

## 5 Experiments setup

Before presenting our results, we describe the details of the conducted experiments including the used datasets, the proposed algorithms, and the performance metrics which we use to measure the effectiveness and efficiency. Table 1 shows the parameters used throughout the experiments.

### 5.1 Datasets

We used the following real world datasets:

1. *Flights database* The Flights database contains flights delays in the year 2008. It was obtained from the US Department of Transportation’s Bureau of Transportation Statistics (BTS).<sup>2</sup> The database contains 250k tuples with a total of 20 dimensions: 10 dimension attributes and 10 measure attributes.
2. *GoCard database* This is the database we introduced in Example 1. It has 4.4 million tuples with a total of 13 dimensions.

### 5.2 Algorithms

We have implemented the following algorithms:

<sup>2</sup> <http://www.transtats.bts.gov/>.

**Algorithm 2:** *Sela***Input:** Attributes  $A(a_1, a_2, \dots, a_n)$ , Query  $Q$ , Views limit  $R$ **Output:** Set  $\mathcal{H}$ : Highest priorities of dimension attributes $C = \phi$  Set of all dimension attributes priorities ;**for**  $i = 1$  **to**  $n$  **do**     $DvalD(a_i) \leftarrow$  number of distinct values of  $a_i$  in  $D$  ;     $DvalDQ(a_i) \leftarrow$  number of distinct values of  $a_i$  in  $DQ$  ;     $Pr(a_i) = DvalDQ(a_i) * Sel_{a_i}^{DQ} + (\frac{DvalDQ(a_i)}{DvalD(a_i)}) * Sel_{a_i}^D$ ;     $C \leftarrow Pr(a_i)$ ;Sort  $C$ ; $G = \lfloor \frac{R}{M \times F} \rfloor$  Calculate the required dimension number;**for**  $i = 1$  **to**  $G$  **do**     $\mathcal{H} \leftarrow C.get(i)$ ;**return**  $\mathcal{H}$ ;**Algorithm 3:** *DimsHisto***Input:** Attributes  $A(a_1, a_2, \dots, a_n)$ , Query  $Q$ , Views limit  $R$ **Output:** Set  $\mathcal{H}$ : Highest priorities of dimension attributes $C = \phi$  Set of all dimension attributes priorities ;**for**  $i = 1$  **to**  $n$  **do**    Compute  $HD(a_i)$  and  $HDQ(a_i)$ ;     $Pr(a_i) = d(HD(a_i), HDQ(a_i))$ ;     $C \leftarrow Pr(a_i)$ ;Sort  $C$ ; $G = \lfloor \frac{R}{M \times F} \rfloor$  Calculate the required dimension number;**for**  $i = 1$  **to**  $G$  **do**     $\mathcal{H} \leftarrow C.get(i)$ ;**return**  $\mathcal{H}$ ;

1. *SeeDB baseline* State-of-the-art algorithm [33] that processes the entire data without discarding any view. It thus provides an upper bound on latency and accuracy and lower bound on the error distance.
2. *SeeDB Rnd* A modified version of SeeDB which returns a random set of  $K$  aggregate views as the result. This strategy gives a lower bound on accuracy and upper bound on error distance: for any technique to be useful, it must do significantly better than SeeDB Rnd.
3. *DiffDVal* It prioritizes dimensions based on the number of distinct values in each dimension (Algorithm 1).
4. *Sela* Our proposed algorithm (Algorithm 2).
5. *DimsHisto* Our proposed algorithm (Algorithm 3).

Note that the Priority Evaluator module in our proposed RtSEngine utilizes DiffDVal, Sela, and DimsHisto algorithms to prioritize visualizations. On the other hand, the Cost Estimator module implements the same three algorithms while utilizing the cost estimations approaches described earlier in Sect. 4.2.

### 5.3 Performance metrics

We used two metrics for evaluating the results of our proposed approaches. One of these metrics is used by SeeDB [33] to evaluate the quality of the recommended views. To evaluate the quality and correctness of the proposed algorithms, we used the following metrics:

**Algorithm 4:** *ViewsEstimate*

---

**Input:** Attributes  $A(a_1, a_2, \dots, a_n)$ , Measures  $M(m_1, m_2, \dots, m_o)$ , Functions  $F(f_1, f_2, \dots, f_x)$   
**Output:**  $S$ : set of views estimated costs

```

for  $i = 1$  to  $n$  do
    for  $j = 1$  to  $o$  do
        for  $p = 1$  to  $x$  do
             $C(VDQ) = EstCostDQ(a_i, m_j, f_p)$ ;
             $C(VD) = EstCostD(a_i, m_j, f_p)$ ;
             $C(d(VDQ, VD))$  Eq.10;
             $Cost = C(VDQ) + C(VD) + C(d(VDQ, VD))$ ;
             $S.add(Cost, (a_i, m_j, f_p))$ ;
    return  $S$ ;
```

---

1. *Accuracy* If  $\{VS\}$  is the set of aggregate views with the highest utility, and  $\{VT\}$  is the set of aggregate views returned by the baseline SeeDB, then the accuracy is defined as:

$$Accuracy = \frac{1}{|VT|} * \sum x, \text{ where } \begin{cases} x = 1 & \text{if } VT_i = VS_i \\ x = 0 & \text{otherwise} \end{cases}$$

i.e., accuracy is the fraction of true positions in the aggregate views returned by SeeDB.

2. *Distance error* Since multiple aggregate views can have similar utility values, we use the utility distance as a measure of how far SeeDB results are from the true top- $K$  aggregate views. Formally, SeeDB [32] defines distance error as the difference between the average utility of  $\{VT\}$  and the average utility of  $\{VS\}$ :

$$Distance\ error = \frac{1}{K} (\sum_i U(VT_i) - \sum_i U(VS_i))$$

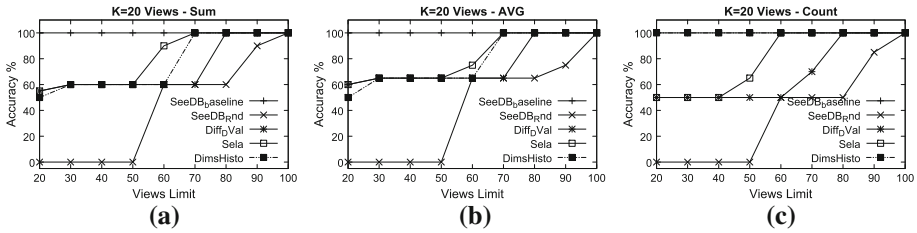
All experiments were run on a PC machine with Windows 10, Intel CPU 2.8Ghz, and 8 GB of RAM memory. The RtSEngine and the algorithms were coded using the Java programming language, and datasets were loaded into a Postgres DBMS. The datasets along with the implementation are available online as a GitHub repository at [https://github.com/ibrahimDKE/Cdb\\_RtsEngine\\_DKE\\_UQ](https://github.com/ibrahimDKE/Cdb_RtsEngine_DKE_UQ).

## 6 Experiments results

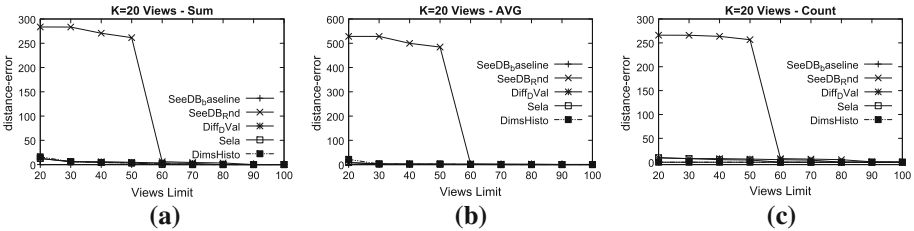
Next, we present our results which demonstrates the efficiency and effectiveness of our proposed algorithms. Firstly, we test the quality of the results produced by our algorithms on the Flights database in Sect. 6.1. Then, we perform similar experiments on the GoCard database in Sect. 6.2, to show that the results are consistent. Later, we present our detailed results on the efficiency of our algorithms in Sect. 6.3. Finally, we show our experiments on the time limit parameter in Sect. 6.4 and cost estimation in Sect. 6.5.

### 6.1 Quality evaluation across aggregate functions

In these experiments, we evaluated the quality of the recommended visualizations produced by our proposed techniques across different aggregate functions, namely `count`, `sum`, `avg`, `min`, `max`. The dataset used is the Flights database with 10 dimension attributes and 10 measure attributes. We run these experiments to assess the quality of the recommended



**Fig. 5** Accuracy on varying view space  $R$  and  $K = 20$ . **a** Sum, **b** avg, **c** count



**Fig. 6** Distance error on varying view space  $R$  and  $K = 20$ . **a** sum, **b** avg, **c** count

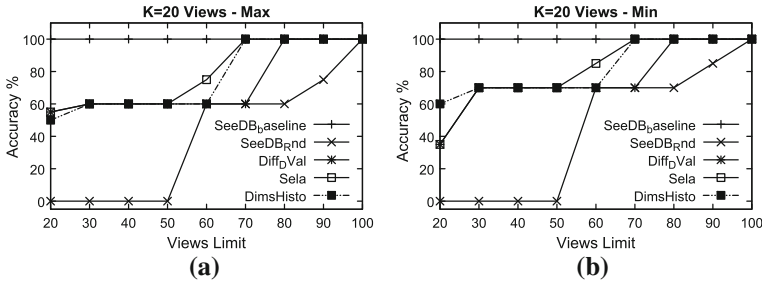
views over each aggregate function separately with a view space size  $SP = 1 \times 10 \times 10 = 100$  possible views. A utility of a view is measured using the earth mover’s distance (EMD).

We report the accuracy and distance error of views produced by our proposed algorithms by varying the limited number of views  $R$  while  $K = 20$ . In these experiments, we use the following query as our target view:

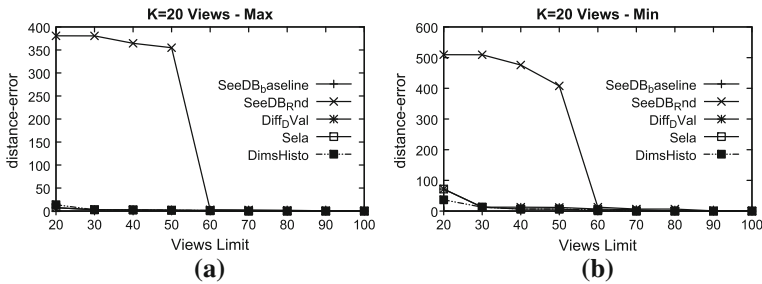
```
Q : SELECT * FROM Flights WHERE uniquecarrier
    = 'American Airlines Inc.'
```

In summary, Sela and DimsHisto algorithms both produce results with accuracy  $> \% 80$  for all aggregate functions, especially when  $R = 60$ , as shown in Fig. 5a–c. Moreover, they produce results with  $\% 100$  accuracy when  $R > 60$ . Sela does slightly better than DimsHisto in terms of accuracy, as Sela evaluates the recommended views by capturing the change of the selectivity ratios of dimension attributes that create views in both result set and reference set. However, DimsHisto scores  $\%100$  accuracy in Fig. 5c for aggregate function count because the generated histograms from this algorithm are similar to the views created by counting dimension attribute values across different measure attributes. Algorithm Diff<sub>D</sub>Val has the lowest accuracy and the highest distance error among the other algorithms specially for aggregate functions  $\max$ ,  $\min$  as shown in Fig. 7a, b as it assess recommended views based on the difference of the distinct values only (Fig. 6).

As shown in Fig. 6a–c, the proposed algorithms produce results near-zero distance error for all aggregate functions compared with lower baseline strategy SeeDB<sub>Rnd</sub> which produce views with low quality; however, the quality of the recommended views produced by the proposed algorithms is almost near to the same utilities of views output by the top baseline SeeDB<sub>baseline</sub>. The distance error of results in the first view limits=20 and 30 views as shown in Fig. 8a, b is high specially for the aggregate function  $\min$  because functions such as  $\min$ ,  $\max$  are not docile for sampling, but the proposed algorithms still score very low distance error.



**Fig. 7** Accuracy on varying view space  $R$  and  $K = 20$ . **a** max, **b** min



**Fig. 8** Distance error on varying view space  $R$  and  $K = 20$ . **a** max, **b** min

The proposed techniques recommend high-quality views in different views limits. Furthermore, the accuracy is increasing without fluctuating along various views limits  $R$ , and similarly, the distance error is declining while increasing the number of explored views  $R$ . In the worst cases, the accuracy and the distance error remain constant while increasing the number of explored views  $R$ .

In the following experiments, we vary  $K$  and fix the number of explored visualizations as  $R = 70$  and measure the accuracy, and error distance for each of our strategies along different aggregate functions. We pay special attention to  $K = 10$  and  $K = 20$  because empirically these  $K$  values are used most commonly. Figures 9a–c and 11a, b show that *Sela* and *DimsHisto* algorithms both produce results with accuracy %100 and zero distance error when  $K = 10$  and  $K = 20$  for all aggregate functions. Moreover, *DiffDVal* algorithm scored accuracy %100 in the first number of recommended views  $K = 10$ . Although *DiffDVal* obtains the same accuracy as *SeeDBRnd* for all aggregate functions, the *DiffDVal* scores much better distance error than *SeeDBRnd*, as shown in Figs. 10a–c and 12a, b. As discussed in the previous experiment, the *DimsHisto* algorithm scores accuracy %100 specifically when the aggregate function is *count*. It also succeeds to recommend views with %100 accuracy and zero distance error for aggregate functions *count*, *sum*, *avg* as shown in Fig. 9a–c. In addition, we found that *Sela* and *DimsHisto* algorithms produce high-quality views with %100 accuracy and zero distance error for *max* aggregate function. Also, they obtain  $> \%75$  and  $< 0.2$  distance error for *min* aggregate function when  $K = 70$  as shown in Fig. 12a, b, respectively.

Figures 10a–c and 12a, b show that the *DiffDVal* approach scores the same accuracy produced by *SeeDBRnd* and obtains very low distance error along all aggregate functions when compared with *SeeDBRnd*. Hence, our proposed approaches boost the accuracy of the recommended views for the mostly common used  $K$  values. Moreover, the *Sela* and

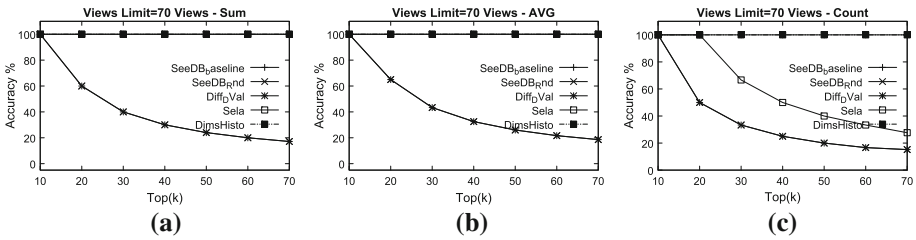


Fig. 9 Accuracy while varying  $K$  and  $R = 70$ . **a** sum, **b** avg, **c** count

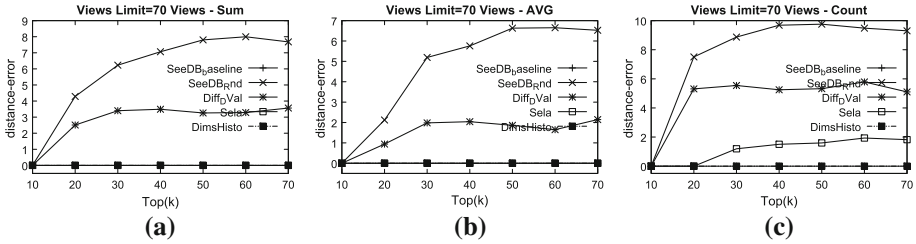


Fig. 10 Distance error on varying  $K$  and  $R = 70$ . **a** sum, **b** avg, **c** count

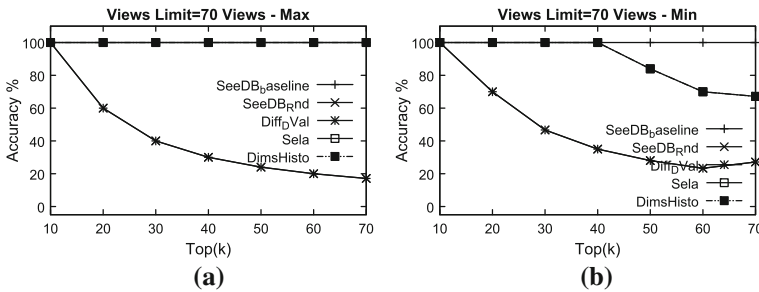


Fig. 11 Accuracy on varying  $K$  and  $R = 70$ . **a** max, **b** min

*DimHisto* algorithms achieve better quality results than *Diff<sub>D</sub>Val* because they capture the data distribution in the dimension attributes by using selectivity ratios and frequency histograms.

### 6.2 Accuracy evaluation

We present now our results on the GoCard database for all aggregate functions *count*, *sum*, *avg*, *min*, *max*. Hence, the view space  $SP = 5 \times 9 \times 3 = 135$  views. Similar to the previous experiments, we used earth mover’s distance (EMD) as the deviation metric for computing the utility of a view. Also, we use the following query as our target view:

```
Q: SELECT * FROM GoCard WHERE alightingstop
    ='University of Queensland';
```

Figure 13a shows the accuracy of the results produced by algorithms *Sela*, *Diff<sub>D</sub>Val*, *DimHisto*, and *SeeDB<sub>Rnd</sub>* to find top 25 views comparing with different view space  $R$  values. As shown, the proposed algorithms *Sela* and *Diff<sub>D</sub>Val* scored the same accuracy

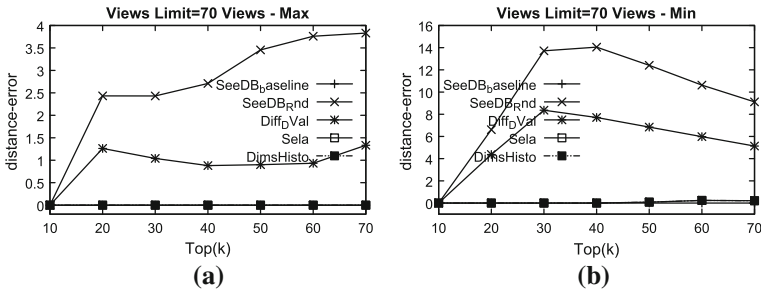


Fig. 12 Distance error on varying  $K$  and  $R = 70$ . **a** max, **b** min

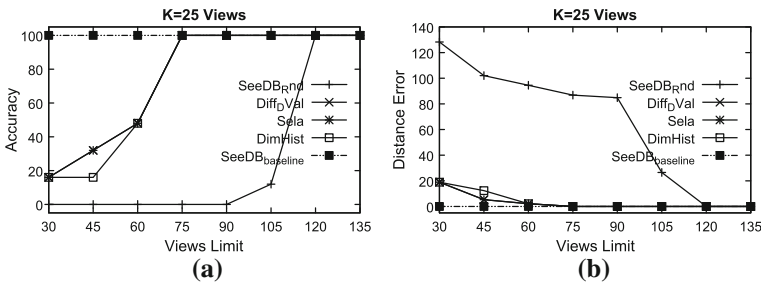


Fig. 13 Results quality while varying view space  $R$  and  $K = 25$ . **a** Accuracy, **b** distance error

in the first 30 explored views. However, *DimHisto* shows lower accuracy than *Sela* and *DiffDVal* when the number of explored views is 45. The reason is *DimHisto* evaluates dimension attributes according to their frequencies; hence, it is less descriptive to some aggregate functions such as *max*, *min*. Note that the accuracy of the proposed algorithms increases with  $R$ , as shown in Fig. 13. Finally, *SeeDB<sub>Rnd</sub>* obtains the lowest accuracy while varying  $R$ , except when it considers almost all the views, i.e., when  $R$  approaches  $SP$ .

Figure 13b reports the distance error produced by algorithms *Sela*, *DiffDVal*, and *DimHisto* to find top 25 ( $K = 25$ ) views across different values of  $R$ . As shown, our proposed algorithms succeed to minimize the distance error as quickly as *SeeDB<sub>baseline</sub>*, especially when expanding the space size  $R$ . Although algorithm *DimHisto* obtains lower accuracy than *Sela* and *DiffDVal* as shown in Fig. 13a when  $R = 45$ , the distance error at the same view space is low. This is because *DimHisto* recommends different views with high utility values to minimize the distance error. *SeeDB<sub>Rnd</sub>* shows high distance error even when the space size is large, i.e.,  $R = 90$ .

To sum up, the proposed algorithms evaluate the dimension attributes according to different priorities methods. Then, by recommending a set of views which increases the quality of the view space limit  $R$  in terms of minimizing the distance error and enhancing the accuracy, as explained earlier by Fig. 13a, b.

Figure 14a shows the accuracy of the compared algorithms in a fixed space size  $R = 90$  while varying  $K$ . As shown, all algorithms score %100 accuracy in the first top 45 views, which form half of the explored views. We observe that the accuracy declines while increasing  $K$  in a fixed space limit  $R$ . This is because when one-dimension attribute is incorrectly prioritized, it will consequently affects all recommended views that are created from that dimension attribute. However, the accuracy is above %50 when  $K = 90$  (i.e., the

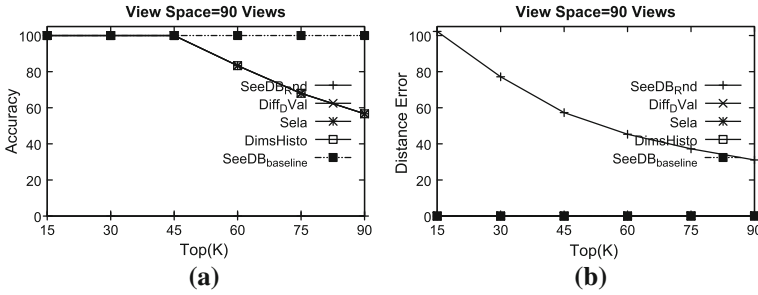


Fig. 14 Results quality while varying  $K$  and  $R = 90$ . **a** Accuracy, **b** distance error

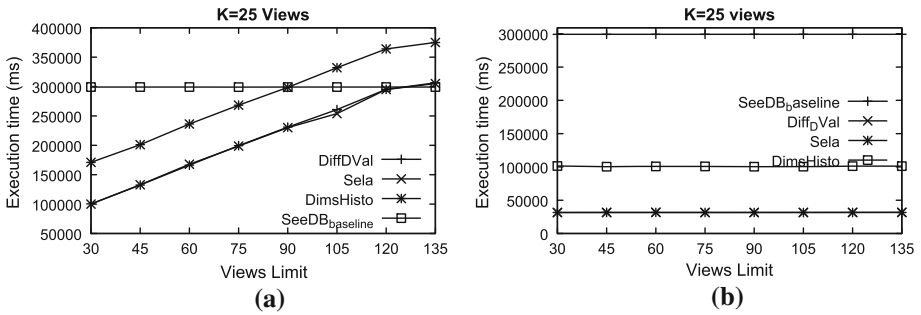


Fig. 15 Algorithms performance while varying  $R$ . **a** Execution time while varying  $R$ , **b** average overhead

entire view space limit) as shown in Fig. 14a. Furthermore, analysts are usually interested in recommending a small number of visualizations, i.e.,  $K = 25$ .

In Fig. 14b, the distance error of the compared algorithms is shown while varying  $K$  and  $R$  is fixed to 90 views. All algorithms produce small distance error for  $K = 60$ ; however, *DiffDVal* shows the smallest distance error across different  $K$  values. Both *Sela* and *DimHisto* report growing distance error with respect to top- $K$  required by the analyst in certain view space  $R = 90$ .

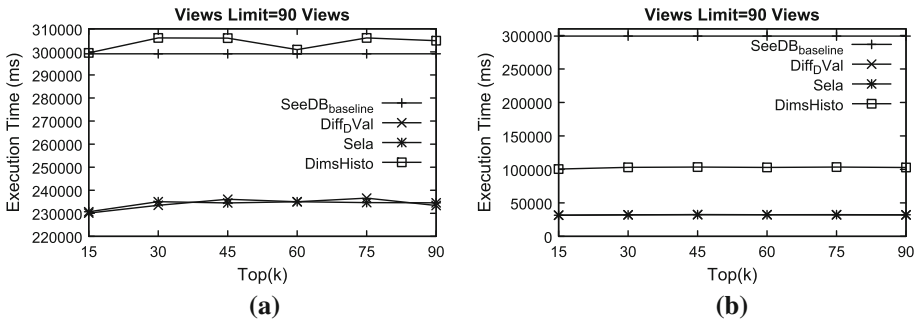
While the discussed algorithms show high accuracy and low distance errors along different  $R$  and  $K$  values, as demonstrated above, these algorithms differentiate on the quality measures. For instance, *Sela* and *DiffDVal* obtain higher accuracy when compared with *DimHisto* as shown in Figs. 13a and 14a, but *DiffDVal* obtains the lowest distance error, as shown in Figs. 13b and 14b.

### 6.3 Efficiency evaluation

In this section, we evaluated the efficiency of our prioritizing algorithms in terms of the overhead added to the automatic recommendation engine RtSEngine. We report the overhead as the execution time averaged over 5 runs. Similar to previous experiments, we vary  $K$  and  $R$  and compare with the actual execution of SeeDB engine as a baseline.

As shown in Fig. 15a, the total execution time of the algorithms is compared with the original SeeDB baseline. As shown, the improvements in the performance by using the proposed algorithms are significant when compared with the baseline. Furthermore, the execution time of our proposed algorithms increases linearly with  $R$ .





**Fig. 16** Algorithms performance on varying  $K$ . **a** Total execution time while varying  $K$  and  $R = 90$ , **b** average overhead while varying  $K$  and  $R = 90$

The execution time shown in Fig. 15b is the extra overhead needed by our proposed algorithms. As shown, the average overhead is almost stable along different  $R$  values. This is because our algorithms evaluate a fixed set of dimension attributes every time, regardless of the value of  $R$ . The high cost of *DimsHisto* is due to its nature: it processes a number of queries to create histograms for computing the distance among them.

The following experiments discuss the efficiency of the proposed algorithms along different  $K$  values. As shown in Fig. 16a, the proposed algorithms show improvements in the execution. More than %40 when compared with the SeeDB baseline execution time. As shown above, *DimsHisto* shows the highest cost among the algorithms *Sela* and *DiffDVal*.

Figure 16b shows the average overhead of the algorithms while varying  $K$ . The overhead is almost constant while increasing  $K$ . This is because the space limit  $R$  is constant too.

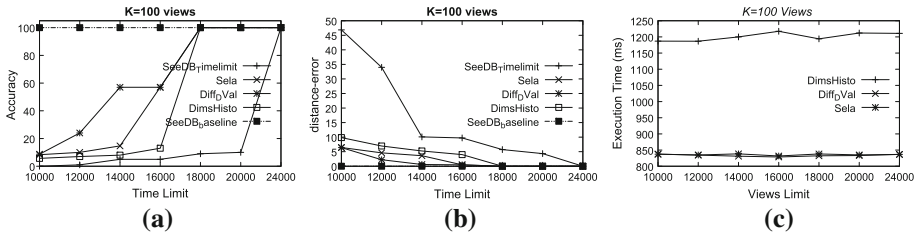
### 6.4 Time limit ( $tl$ )

In these experiments, we evaluated the quality of the recommended visualizations produced by our proposed techniques across different aggregate functions, namely *count*, *sum*, *avg*, *min*, *max*. The dataset used is the Flights database with 10 dimension attributes and 10 measure attributes. We run these experiments to assess the quality of the recommended views over each aggregate function separately with a view space size  $SP = 5 \times 10 \times 10 = 500$  possible views. A utility of a view is measured using the earth mover’s distance (EMD).

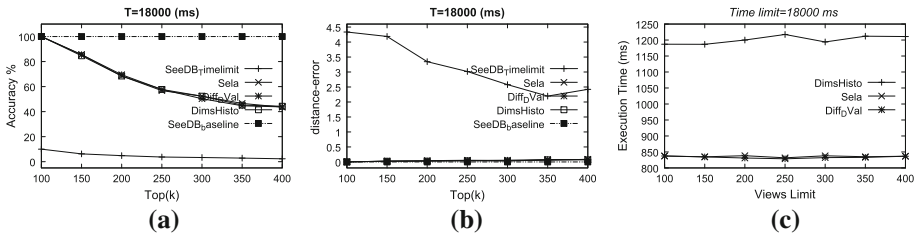
We report the accuracy, distance error, and efficiency of views produced by our proposed algorithms by varying the time limits  $tl$ , number of views  $R$  and  $K$ . In these experiments, we use the following query:

```
Q: SELECT * FROM Flights WHERE dimmonth IN
    ('APR', 'MAY', 'JUN');
```

The query  $Q$  represents the second quarter of the database, so that we can compare with the entire database to find different  $K$  views while varying the time limit  $tl$ . In addition, we evaluated the quality of the top- $K$  views produced by each algorithm with those produced by SeeDB baseline, i.e., without any time limits or optimizations used. We implemented *SeeDBTimeLimit* algorithm which processes the entire data and views in a specified execution time limit and then recommends top views that are processed in that time limit. This strategy represents a lower bound on accuracy and an upper bound on distance error.



**Fig. 17** Performance of *Sela*, *DiffDVal*, *DimHisto*, and *SeeDBTimelimit* on different time limits while  $K = 100$ . **a** Accuracy, **b** distance error, **c** average overhead



**Fig. 18** Performance of *Sela*, *DiffDVal*, *DimHisto*, and *SeeDBTimelimit* on varying  $K$  and  $tl = 18s$ . **a** Accuracy, **b** distance error, **c** average overhead

In Fig. 17a, b, the accuracy and distance error of the results produced by algorithms *Sela*, *DiffDVal*, *DimHisto*, and *SeeDBTimelimit* to find a top 100 ( $K = 100$ ) views are compared with *SeeDB* baseline on different execution time limit  $tl$ . These algorithms output an ordered set of dimension attributes based on their priorities and submit the ordered set to the execution engine. Then, it processes all views generated according to the ordered set that produced by algorithms. As shown, *SeeDBTimelimit* shows high distance error and very low accuracy as well while the algorithms *Sela* and *DiffDVal* score higher accuracy than *DimHisto*. Although the proposed algorithms show a growing accuracy while extending the time limit, they achieved %100 accuracy for  $tl > 18000$  ms. For a big database as the one used here (i.e., 500 different views), 18 seconds is considered reasonable.

The algorithms boosted the performance by more than %30 and preserved the quality of views. On the other side, Fig. 17c describes the execution costs referred to as the overhead time of the proposed algorithms on the same experiment. *DimHisto* algorithm execution time is about 1200ms, while the algorithms *Sela* and *N - N'* have almost the similar execution time, about 825ms. This shows that *Sela* and *DiffDVal* algorithms are faster by %66 than *DimHisto*. As discussed previously, the additional histograms distance computations are the cause of the extra overhead in algorithm *DimHisto*.

To show the effects of varying  $K$  with time limits, Fig. 18a, b show the accuracy of the algorithms *Sela*, *N - N'*, *DimHisto*, and *SeeDBTimelimit* in a certain time limit  $tl = 18000$ ms. As shown, all algorithms score %100 accuracy in the first top 100 views. However, the accuracy declines with increasing  $K$  while  $tl$  is fixed, but the proposed algorithms score very small distance error for large values of  $K$ , while *SeeDBTimelimit* shows very low accuracy and huge distance error. As illustrated in Fig. 17c, the overhead costs of the proposed algorithms remain stable on different time limits. In short, the proposed algorithms improve the quality of the results, thanks to the evaluation metrics that are used along different  $K$ ,  $R$ , and time limits  $tl$  values. Moreover, the algorithms overhead is comparatively small with the total execution time of baseline *SeeDB*.

### 6.5 Cost estimation evaluation

In the next experiments, we evaluate our proposed cost estimation methods discussed previously in Section 4.2 on the GoCard database with  $SP = 5 \times 9 \times 3 = 135$ . Similar to previous experiments, we include the aggregate functions *count*, *sum*, *avg*, *min*, *max* and use the earth mover’s distance (EMD) as our deviation metric for computing the utility. We use the following query for the next experiments:

```
Q: SELECT * FROM GoCard WHERE alightingstop
    ='University of Queensland';
```

We are interested in evaluating the results of the cost estimation methods based on the classical effectiveness and efficiency. For effectiveness, we assess the quality of views outputted by the proposed prioritizing algorithms *DiffDVal*, *Sela*, and *Dimshisto* along different cost estimation methods (i.e., DB estimate and actual costs) comparing with SeeDB baseline. We implemented two baseline strategies: SeeDB baseline which processes the entire data and evaluates all views without any cost considerations. Thus, it provides upper bounds on latency and accuracy and a lower bound on distance error. The other baseline strategy we implemented is actual costs that computes the actual execution time of all views and also the actual computational time for computing the utility of views.

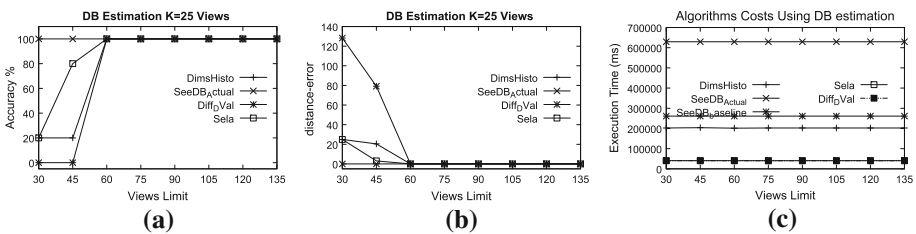
We measure the quality of results based on the accuracy and distance error. However, the efficiency of estimating methods is captured by showing the execution time across the proposed prioritizing algorithms *DiffDVal*, *Sela*, and *Dimshisto*.

The first experiment evaluates the results of the top 25 views using the DB estimates (reading the costs from the database optimizer) along different space limits *R* while comparing the estimated costs of the recommended views with the baseline. Figure 19a shows that the accuracy of the results produced by *Sela*, *DiffDVal*, and *Dimshisto* while reading the costs of the recommended views from the database optimizer to find a top 25 views while varying *R* is almost %100 starting from *R* = 60.

While *Sela* algorithm has the highest accuracy and the lowest error distance among all proposed algorithms as shown in Fig. 19b, the accuracy of *DiffDVal* is very low when  $R \leq 60$  because it evaluates views according to the difference of distinct values only and does not consider the query size, while *Sela* does.

Consequently, the error distance is higher than *Sela* and *Dimshisto*.

The following experiment illustrates the average overhead of using different cost estimation methods along our prioritizing algorithms added to the actual SeeDB baseline. In Fig. 19c, the average overhead of implementing the algorithms *Sela*, *DiffDVal*, and *Dimshisto* and reading the costs from database optimizer is shown on the y-axis. As shown, computing



**Fig. 19** Results quality and average overhead using DB Estimation. **a** Accuracy, **b** distance error, **c** average overhead

actual costs is much expensive than running SeeDB itself. This is because SeeDB does not execute all aggregate queries. For example, the average function `avg` of a view is computed by dividing the total (`sum` aggregate function) on their frequency (`count` aggregate function). Moreover, SeeDB combines the aggregate queries of the datasets  $D$  and  $DQ$ . All algorithms have a stable performance on different space limits  $R$  because the algorithms evaluate the same set of dimension attributes  $A$  and outputs a subset  $A'$  of top-scored dimension attributes. As shown, *DimsHisto* shows a considerable time cost since it create and assess histograms; however, both algorithms *Sela* and *DiffDVal* have nearly equal execution costs.

## 7 Conclusion

Finding top interesting visualizations by exploring a specified number of visualizations or an execution time budget, while persevering the quality and the accuracy of the recommended views is a challenging and emerging problem. In this paper, we addressed this problem and proposed an efficient framework called real-time scoring engine (RtSEngine) that assist data analysts in the exploration of visualizations generated from structured databases.

Specifically, RtSEngine supports analysts by efficiently recommending visualizations while meeting analysts budgets: certain number of visualizations or execution time quote. RtSEngine accomplishes this by incorporating inventive approaches to prioritize and score attributes that form all possible visualizations in database based on their statistical proprieties such as selectivity ratio, data distribution, and number of distinct values. Then, RtSEngine recommends the views created from top-scored attributes.

In addition, we presented visualizations cost-aware techniques that estimate the retrieval and computation costs of all visualizations. Those estimated costs are then fed to RtSEngine to recommend views while considering their costs to guarantee the efficiency and effectiveness of the recommendation process.

Finally, we conducted comparative experiments and demonstrated the quality of visualizations and the overhead obtained by applying our techniques on both synthetic and real datasets. The experiments showed superior effectiveness and efficiency of our proposed approaches on different time and space limits.

## References

1. Barbará D, DuMouchel W, Faloutsos C, Haas PJ, Hellerstein JM, Ioannidis YE, Jagadish HV, Johnson T, Ng RT, Poosala V, Ross KA, Sevcik KC (1997) The New Jersey data reduction report. *IEEE Data Eng Bull* 20(4):3–45
2. Bubeck S, Wang T, Viswanathan N (2013) Multiple identifications in multi-armed bandits. In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16–21 June 2013*, pp 258–265. <http://jmlr.org/proceedings/papers/v28/bubeck13.html>
3. Charikar M, Chaudhuri S, Motwani R, Narasayya VR (2000) Towards estimation error guarantees for distinct values. In: *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Dallas, TX, USA, 15–17 May 2000*, pp 268–279
4. Chaudhuri S (1998) An overview of query optimization in relational systems. In: *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, PODS '98, ACM, New York, NY, USA, pp 34–43. doi:10.1145/275487.275492*
5. Chaudhuri S, Motwani R, Narasayya VR (1998) Random sampling for histogram construction: How much is enough? In: *SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, Seattle, Washington, USA, 2–4 June 1998*, pp 436–447
6. Fisher D (2007) Hotmap: looking at geographic attention. *IEEE Trans Vis Comput Graph* 13(6):1184–1191

7. Getoor L, Taskar B, Koller D (2001) Selectivity estimation using probabilistic models. In: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, SIGMOD '01, ACM, New York, NY, USA, pp 461–472. doi:[10.1145/375663.375727](https://doi.org/10.1145/375663.375727)
8. Gilbert AC, Kotidis Y, Muthukrishnan S, Strauss M (2001) Optimal and approximate computation of summary statistics for range aggregates. In: Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Santa Barbara, CA, USA, 21–23 May 2001
9. Gonzalez H, Halevy AY, Jensen CS, Langen A, Madhavan J, Shapley R, Shen W, Goldberg-Kidon J (2010) Google fusion tables: web-centered data management and collaboration. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, IN, USA, 6–10 June 2010, pp 1061–1066
10. Hellerstein JM, Haas PJ, Wang HJ (1997) Online aggregation. In: SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 13–15 May 1997, pp 171–182
11. Holzinger A, Simonic K (eds.) (2011) Information Quality in e-Health - 7th Conference of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2011, Graz, Austria, 25–26 Nov 2011, Lecture Notes in Computer Science, vol 7058. Springer
12. Hou WC, Ozsoyoglu G (1991) Statistical estimators for aggregate relational algebra queries. *ACM Trans Database Syst* 16(4):600–654. doi:[10.1145/115302.115300](https://doi.org/10.1145/115302.115300)
13. Hund M, Böhm D, Sturm W, Sedlmair M, Schreck T, Ullrich T, Keim DA, Majnaric L, Holzinger A (2016) Visual analytics for concept exploration in subspaces of patient groups. *Brain Inf* 1–15. doi:[10.1007/s40708-016-0043-5](https://doi.org/10.1007/s40708-016-0043-5)
14. Ioannidis Y (2003) The history of histograms (abridged). In: Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29, VLDB '03, pp 19–30. VLDB Endowment. <http://dl.acm.org/citation.cfm?id=1315451.1315455>
15. Jagadish HV (1999) Review—explaining differences in multidimensional aggregates. *ACM SIGMOD Digit Rev* 1:1–11
16. Jang MH, Kim SW, Faloutsos C, Park S (2011) A linear-time approximation of the earth mover's distance. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, ACM, New York, NY, USA, pp 505–514. doi:[10.1145/2063576.2063652](https://doi.org/10.1145/2063576.2063652)
17. Jugel U, Jerzak Z, Hackenbroich G, Markl V (2016) VDDA: automatic visualization-driven data aggregation in relational databases. *VLDB J* 25(1):53–77. doi:[10.1007/s00778-015-0396-z](https://doi.org/10.1007/s00778-015-0396-z)
18. Kandel S, Parikh R, Paepcke A, Hellerstein JM, Heer J (2012) Profiler: integrated statistical analysis and visualization for data quality assessment. In: Proceedings of the International Working Conference on Advanced Visual Interfaces, ACM, pp 547–554
19. Key A, Howe B, Perry D, Aragon CR (2012) Vizdeck: self-organizing dashboards for visual analytics. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, 20–24 May 2012, pp 681–684
20. Lahdenmaki T, Leach M (2005) Relational database index design and the optimizers. Wiley-Interscience
21. Lipton RJ, Naughton JF, Schneider DA (1990) Practical selectivity estimation through adaptive sampling. In: Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, SIGMOD '90, ACM, New York, NY, USA, pp 1–11. doi:[10.1145/93597.93611](https://doi.org/10.1145/93597.93611)
22. Livny M, Ramakrishnan R, Beyer KS, Chen G, Donjerkovic D, Lawande S, Myllymaki J, Wenger RK (1997) Devise: integrated querying and visualization of large datasets. In: SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 13–15 May 1997, pp 301–312
23. Mackert LF, Lohman GM (1986) R\* optimizer validation and performance evaluation for local queries. *SIGMOD Rec* 15(2):84–95. doi:[10.1145/16856.16863](https://doi.org/10.1145/16856.16863)
24. Mackinlay JD, Hanrahan P, Stolte C (2007) Show me: automatic presentation for visual analysis. *IEEE Trans Vis Comput Graph* 13(6):1137–1144
25. Mannino MV, Chu P, Sager T (1988) Statistical profile estimation in database systems. *ACM Comput Surv* 20(3):191–221
26. Piatetsky-Shapiro G, Connell C (1984) Accurate estimation of the number of tuples satisfying a condition. In: SIGMOD '84, Proceedings of Annual Meeting, Boston, Massachusetts, 8–21 June 1984, pp 256–276
27. Sarawagi S (2000) User-adaptive exploration of multidimensional data. In: VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, Cairo, Egypt, 10–14 Sept 2000, pp 307–316
28. Sathe G, Sarawagi S (2001) Intelligent rollups in multidimensional OLAP data. In: VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases, Roma, Italy, 11–14 Sept 2001, pp 531–540
29. Serfling RJ (1974) Probability inequalities for the sum in sampling without replacement. *Ann Stat* 2(1):39–48. doi:[10.1214/aos/1176342611](https://doi.org/10.1214/aos/1176342611)

30. Stillger M, Lohman GM, Markl V, Kandil M (2001) Leo - db2's learning optimizer. In: Proceedings of the 27th International Conference on Very Large Data Bases, VLDB '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 19–28. <http://dl.acm.org/citation.cfm?id=645927.672349>
31. Stolte C, Hanrahan P Polaris A (2000) system for query, analysis and visualization of multi-dimensional relational databases. In: Proceedings of the IEEE Symposium on Information Visualization 2000, INFOVIS '00, IEEE Computer Society, Washington, DC, USA, p 5. <http://dl.acm.org/citation.cfm?id=857190.857686>
32. Vartak M, Madden S, Parameswaran A, Polyzotis N Seedb: towards automatic query result visualizations. Tech. rep., Technical Report, data-people. cs.illinois.edu/seedb-tr.pdf
33. Vartak M, Madden S, Parameswaran AG, Polyzotis N (2014) SEEDB: automatically generating query visualizations. PVLDB 7(13):1581–1584
34. Wong BLW, Chen R, Kodagoda N, Rooney C, Xu K (2011) INVISQUE: intuitive information exploration through interactive visualization. In: Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Extended Abstracts Volume, Vancouver, BC, Canada, 7–12 May 2011, pp 311–316. doi:10.1145/1979742.1979720



**Ibrahim A. Ibrahim** is currently a Ph.D. candidate in the School of Information Technology and Electrical Engineering at the University of Queensland Brisbane, Australia. He obtained his Master in Computer Science in from Ain Shams University, Egypt in 2009. His research interests include Big Data Exploration, Query Visualizations, and Query Explanations in Databases.



**Abdullah M. Albarrak** is currently a Ph.D. candidate in the School of Information Technology and Electrical Engineering at the University of Queensland Brisbane, Australia. He obtained his Master in Computer Science from the University of Queensland in 2012. His research interests are focused on interactive data exploration techniques and their applications.



**Xue Li** is a Professor in the School of Information Technology and Electrical Engineering at the University of Queensland (UQ) in Brisbane, Queensland, Australia. He is honored as one of the most powerful people in Australia on Big Data by the Financial Review—the Power Issue 2015. His major areas of research interests and expertise include: Data Mining, Social Computing, Database Systems, and Intelligent Web Information Systems.