SeeDB: WhatÕs interesting about my query?

ABSTRACT

Data scientists analyzing large amounts of data often rely on visualizations to identify interesting trends and gain insights. However, picking the right visualization is a manual and tedious task. Given a query, the analyst would like to know what makes the results of a query "interesting" compared to the underlying dataset. In this paper, we demonstrate a prototype of a system SeeDB a system that automatically discovers statistical differences between the query results and the underlying dataset, and visualizes the differences to aid data exploration.

1. INTRODUCTION

Data scientists analyze large amounts of data to find new insights, trends and anomalies. Given the large amount of data and large bandwidth available for visual analysis, analysts often use visualizations to identify areas of interest. The general workflow, which is usually ad-hoc, can be represented as follows:

1. Step 1: Analyst selects a subset of the data they are interested in; for instance, the analysts may select patients that had costs more than twice the standard deviation of all patients, or the genes that belong to a particular group, or a set of products of a certain kind.

2. Step 2: The analyst processes and visualizes the data in various ways, e.g. via binning, aggregates, groupbys etc. and then building scatterplots, histograms etc., and studies these "views" of the data.

3. Step 3: Finally, the analyst studies all the views and identifies the interesting ones. He/she then drills down into these views or further explores the properties that show unusual trends. This can be done using sophisticated statistics or machine learning.

This iterative process and repeated for different subsets of the data. Since a large number of views of the data are possible, this process of exploration is tedious and time-consuming. In Step 3 above, we further posit that *what makes a view interesting is deviations from expected behavior.* Therefore, given the subset of data selected using a relational query, what makes the data interesting are trends or properties that are different for this subset compared to the underlying complete dataset.

Ultimately, the analyst must decide if deviations in the data are interesting but we can use deviation as a means to eliminate the laborious process of stepping through all possible views of a dataset.

■<Example■>

In this demo paper, we describe our prototype of SeeDB, a system to automatically identify and visually highlight interesting aspects of a dataset.

While various ways of obtaining "views" of a dataset are possible, for the purpose of this work, we focus on views obtained by adding a single group-by and a single aggregate clause to the query used to select subsets of the data. This restriction also allows us to focus on a limited set of visualization types to show trends in the data.

Formally, the problem can be described as follows: given a database and a query Q, SeeDBidentifies the most significant deviations between the subset of the data selected by the query and the underlying dataset. SeeDBdoes so by considering the space of views obtained by adding a single group-by and aggregate to the input query, and then computing the discriminating power of each of these views. Discriminating power or utility measures how different the distribution of data in the subset is compared to the underlying data. SeeDBproduces the top-k views having the highest dicriminating power. For the utility measure, SeeDBprovides the option of using distance metrics such as earth movers distance, L2 norm etc.

EXAMPLE 1.1. *We focus on a database with a traditional star schema. We operate on a single fact table D, containing information about sales. The schema of D comprises three dimension attributes: Product, Location, and Year, and one measure attribute: Sales.*

Let us assume our analyst has entered a query with a single selection predicate: $Q \equiv \sigma_{(Product\ =\ Staplers)}$. The result contains too many tuples to examine individually, and hence the analyst has to rely on some appropriate visualization in order to glean interesting insights about the overall query result.

SeeDB *searches over all possible discriminating views that can be obtained by adding a single aggregate and group by operator. We initially focus on these two-attribute discriminating views because they are easy to visualize using his-*

Proceedings of the VLDB Endowment, Vol. 7, No. 4

Location Aggregates: $R_1(Q(D))$			
Boston: 30	Seattle: 40	New York: 40	San Francisco: 90
Year Aggregates: $R_2(Q(D))$			
2009: 50	2010: 40	2011: 60	2012: 50

Table 1: Aggregates for Product = 'Staplers'

Location Aggregates: $R_1(D)$			
Boston: 300	Seattle: 300	New York: 300	San Francisco: 700
Year Aggregates: $R_2(D)$			
2009: 100	2010: 200	2011: 500	2012: 800

Table 2: Original Aggregates

*tograms or line plots. (*SeeDB *also considers more general views.) One of these queries is $Q'_1 = R_1(Q)$ where $R_1 \equiv \gamma_{Location,\ sum(Sales)}$. This query tracks the sum of Sales over Location. A possible result, $R_1(Q(D))$, is the discriminating view shown in the top part of Table 2. Another possible query is $Q'_2 = R_2(Q)$, where $R_2 \equiv \gamma_{Year,\ sum(Sales)}$, tracking the sum of Sales over Year. The bottom part of Table 2 shows a possible result of this second view.*

The next step is to score each view based on its utility, i.e., its ability to show an interesting property of the query result. For this purpose, SeeDB *obtains aggregate statistics for Sales for Location and Year for the original full database. (As we will see later, there are interesting optimization opportunities if we can integrate this step with the processing of Q.) Table 1 shows the full database aggregates, i.e., $R_1(D)$ and $R_2(D)$ corresponding to our sample views. (Note the missing Q.) For the Location attribute, both $R_1(D)$ and $R_1(Q(D))$ have similar distributions. (The fact that sales are uniformly lower in $R_1(Q(D))$ is not surprising since $R_1(Q(D))$ only considers a fraction of the data.) However, notice that the distribution across Years is very different for tuples satisfying Product = 'Staplers'. That is, demand for 'Staplers' seems to have not gone up, unlike the other products. This unexpected behavior will be detected by* SeeDB *when it computes the utility of $R_2(Q)$, and hence $R_2(Q)$ (and not $R_1(Q)$) will be suggested to the analyst for further human evaluation.*

Thus, there are several technical challenges that need to be addressed:

- For a given query, n, the total number of discriminating views, (even if we restrict ourselves to views that append a group-by and an aggregation) is likely to be very large to explore exhaustively and precisely. Generating each of $R_1(Q(D))$, ..., $R_n(Q(D))$, scoring them on utility, and then picking the best one is certainly not feasible for most databases. Thus, we need mechanisms to prune the space of views and compute their utility approximately.
- Generating and scoring the discriminating views $R_i(Q(D))$ one-by-one may miss interesting optimization opportunities: First, we may share computation between discriminating views. For example, the results of two views with different aggregates but the same group-by may be computed together in one query, followed by projecting out to reveal the two individual views. Second, by evaluating the discriminating views in a deliberate order, we may be able to prune views with low utility (without evaluation) that are definitely not going to be recommended to the analyst.
- Since visualizations tend to convey approximate information, e.g., a trend in a line plot may be more important than knowing the exact coordinates of each point, we can introduce approximations as part of SeeDB. Thus, the utility of a discriminating view may be computed approximately but efficiently, and the recommended discriminating views can be populated with approximate results, based on synopses of the base data or of the query result, that can be generated much more efficiently.

Over the past few years, there has been a significant effort from the visualization community to provide interactive tools for data analysts. In particular, tools such as ShowMe, Polaris, and Tableau [16, 10] provide a canvas for data analysts to manipulate and view data, tools such as Wrangler [7] allow data analysts to transform and clean data, and tools such as Profiler [8] allow users to visualize simple anomalies in data. However, unlike SeeDB, these tools have little automation; in effect, it is up to the analyst to generate a two-column result (like the result of the discriminating view) to be visualized. Other related areas of work include OLAP and database visualization tools. 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 [12, 14, 13].

Fusion tables [4] allows users to create visualizations layered on top of web databases; they do not consider the problem of automatic visualization generation. Devise [9] translated user-manipulated visualizations into database queries.

2. SYSTEM OBJECTIVES

We now state the goals of SeeDB more formally, to provide a blueprint for the system design sections that follow. When doing so, we deliberately focus on a simple setting to ground our discussion. However, the setting we consider is an important use-case that occurs often in practice, and was the focus of our illustrative example. We will consider advanced variants in Section 4.

Concrete Goals: We consider a database D with a snowflake schema, with dimension attributes A, measure attributes M, and potential aggregate functions F over the measure attributes. We limit the class $\mathcal{Q}$, of queries posed over D, to be those that select a horizontal fragment of the fact table. The selection of the fragment can be done with selection predicates on the fact table, or on dimension tables through key/foreign-key joins. Intuitively, the idea is that the analyst specifies their interest in examining facts that satisfy specific conditions. SeeDB will identify visualizations that show some interesting properties of these facts.

Given a query Q in $\mathcal{Q}$, we define $\mathcal{R}_Q$, the set of all discriminating views, to be the set of views that perform a group-by and some aggregation over the results of Q. For simplicity, we assume that a discriminating view R in $\mathcal{R}_Q$ performs a group-by on a single attribute $a \in A$, and applies an aggregation function $f \in F$ on a single measure attribute $m \in M$. A view in this class corresponds to a two-column table that shows how the value of $f(m)$ varies with values of attribute a. This table can be directly visualized using a histogram, a bar chart or a line plot. (We consider generalizations in Section 4.)

We also assume the existence of a function $U(R)$ that can characterize the utility of each view $R(Q)$ in $\mathcal{R}_Q$ (higher is better). For now, we focus on picking discriminating views that optimize $U(R)$ with latency as low as possible: we return to more general objectives in Section 4. Thus, our concrete goal is:

GOAL 2.1. *Given $Q \in \mathcal{Q}$ and a positive integer K, find K discriminating views $R_i \in \mathcal{R}_Q$, such that the R_i have the largest values of $U(R_i)$ among those in $\mathcal{R}_Q$, and the total latency is minimized.*

Operationalizing Utility: One of the key challenges behind SeeDB is formalizing the utility function $U(R)$ for a discriminating view R. There are many choices for U and we expect SeeDB to recommend views that score high on several metrics. As discussed previously, the proposed metric tries to capture the idea of "deviation" between distributions, i.e., a view has high utility if its contents show a trend that deviates from the corresponding trend in the original database.

We first define some notation. For any discriminating view R_i in the class defined above, we note that $R_i(D)$ and $R_i(Q(D))$ are both two column tables. A two-column table can be represented using a weight vector. We let the weight vector $W_{a,f(m)}$ represent the result of $R_i(D) = \gamma_{a,f(m)}(D)$, i.e., distribution of the aggregate function f on the measure quantity m across various values of the attribute a. Going back to our example, it follows that

$$W_{\text{Year, sum(Sales)}} = (2009 : 100, 2010 : 200, 2011 : 500, 2012 : 800)$$

Here m is Sales, f is the sum, and a is the Year attribute. Then, we let $W^Q_{a,f(m)}$ represent the (changed) distribution of $R_i(Q(D))$, the aggregated quantity m across values of the attribute a, when restricted to the result of the query Q. Thus,

$$W^{\sigma_{\text{Prod = 'Staplers'}}}_{\text{Year, sum(Sales)}} = (2009 : 50, 2010 : 40, 2011 : 60, 2012 : 50)$$

The utility U of a discriminating view $\gamma_{a,f(m)}$ is defined to be the distance between $W^Q_{a,f(m)}$, and $W_{a,f(m)}$: $U(\gamma_{a,f(m)}) = S(W^Q_{a,f(m)}, W_{a,f(m)})$ where S is a distance metric. The higher S is, the more useful a discriminating view is. Common distance metrics used in visualization literature include K-L divergence [19], Jenson-Shannon distance [18, 17], and earth mover distance [20]. Wang [17] provides a good overview of the metrics used in scientific visualizations, while [20] provides a summary of probability-based distance metrics. As discussed earlier, we do not prescribe any specific distance metrics, instead, we plan to support a whole range of distance metrics, which can be overridden by the data analyst.

We show in our extended paper [11] that in our example, we get the right discriminating view (i.e., that of Year) when we consider the metrics mentioned above: the utility for Year is significantly higher than that for Location for each of the metrics considered.

3. INITIAL DESIGN

Our initial design for SeeDB is as a simple wrapper over an existing database system. One straightforward workflow is as follows:

- *Step 1:* Enumerate all discriminating views $R \in \mathcal{R}_Q$ and evaluate utility $U(R)$ (i.e., compute $W^Q_{a,f(m)}$ and $W_{a,f(m)}$) by issuing the corresponding counting queries to the DBMS. Select the K views with highest $U(R)$.
- *Step 2:* Compute the results of these K views using the DBMS, then forward the results to the visual engine.

It is clear that this workflow suffers from several inefficiencies. We now discuss potential optimizations, some of which require new query processing schemes specialized for the problem at hand.

Approximate Utility Computation: We can speed up Step 1 by computing the utilities of discriminating views approximately. Naturally, we would want the approximations to be accurate enough to select a good set of views for Step 2.

Sampling is one possible approximation method: We construct a sample of the query result $Q(D)$, and the underlying data D, and use these samples to compute approximate weight vectors $W^Q_{a,f(m)}$ and $W_{a,f(m)}$. (In fact, the latter, which does not depend on Q, can be computed before any queries are issued.) The question of how large a sample of Q is necessary to enable the weight vectors $W^Q_{a,f(m)}$, for all a, f, m, to have high accuracy is, to the best of our knowledge, still open. Techniques from sampling for aggregation [3], and more generally, approximate query processing [1, 2] may be relevant here.

Ideally, we would want error bounds on the resulting approximate utilities, in order to enable SeeDB to select views of provably high utility and avoid views of provably low utility. These guarantees may depend on the specific metric used. For instance, approximation guarantees on the magnitudes of the weight vectors may directly translate to guarantees on the earth mover distance (a simpler metric), but not the Jenson-Shannon distance (a complex metric).

Searching the Space of Discriminating Views: The space of discriminating views may be too large to search exhaustively in an efficient manner, particularly if SeeDB relies on exact utility computation. Instead, it may be possible to prune the search space by leveraging relationships between discriminating views in terms of utility. For instance, functional dependencies among grouping attributes can help us infer that certain views will have the exact same utility by virtue of having the same groups. Furthermore, the search strategy that navigates the space of views may also take into account inter-dependencies: for instance, investing computational resources to determine that a view has provably low utility would be useful, if this determination will lead to the pruning of several other views correlated with the specific view.

Multi-Query Optimization: Step 2 comprises the evaluation of K queries and so raises opportunities for multi-query optimization [15]. For instance, if we are recommending several discriminating views with the same group-by attribute, we may combine the computation of the views into a single group-by query with multiple aggregations (one per view). We expect that the opportunities to share computation will increase with more complex queries and views.

Multi-query optimization may also be used to optimize utility computation (Step 1). For instance, multi-query optimization may reveal that $W^Q_{a,f_i(m)}$ may be computed together for all i. Additionally, since we make repeated calls to evaluate $W^Q_{a,f(m)}$ for different a, f, m, we can instead first materialize the query result $Q(D)$ and then compute the weight vectors $W^Q_{a,f(m)}$ by issuing queries on the materialized result. Materializing $Q(D)$ can also help in evaluating $R(Q(D))$ for each selected discriminating view R in Step 2.

Fusing the Two Steps: Up to this point we considered view selection and view computation as two separate stages inside SeeDB. Alternatively, we may fuse the two stages in order to share work between the computation of utilities (Step 1) and the evaluation of view queries (Step 2), thus reducing end-to-end latency.

Computing $U(R)$ requires knowledge about the contents of R, and therefore, we compute $U(R)$ and $R(Q(D))$ together. Since it may be prohibitively expensive to compute these quantities for each R, and since the end goal is to rec-

ommend only K views, we may employ approximate utility computation coupled with a pruning rule. Specifically, suppose that is it possible to schedule the computation of $Q(D)$ and receive its output in a random order (e.g., as described in [6]). SeeDB will then observe a sample of increasing size as it consumes the output of $Q(D)$. Processing the output involves two tasks: (a) updating a running estimate of the utility of each view (leveraging the fact that the observed output is a sample of $Q(D)$), and (b) updating the current contents of unpruned views using hash-based aggregation. SeeDB can use the running utility estimates to prune views of low utility. Overall, as SeeDB processes $Q(D)$ it can make progress towards both selecting the top-K views and computing their contents. Thus, this specialized query-processing strategy can reduce SeeDB's end-to-end latency but it comes with higher resource requirements (since many group-by queries need to be processed concurrently).

4. ENHANCEMENTS AND EXTENSIONS

We now present some enhancements that may further improve the analyst's user experience beyond those suggested previously.

Reducing Perceived Latency: To reduce perceived latency, SeeDB can first produce the discriminating view result in the top-K that takes the least amount of time to execute. That way, the user can peruse the first visualization as soon as possible. Then, SeeDB can generate the remaining views. Further, SeeDB can, in the background, compute the result of the current top-K views while considering other views. If a view is no longer in the current top-K, then it is replaced with another (better) view, which begins executing.

Latency Threshold: We may wish to incorporate a user-specified overall latency threshold in our goal (i.e., find top-Kviews such that total latency is bounded), so that the analyst does not have to wait too long to see visualizations. Dealing with a latency threshold is certainly more challenging, and will require new techniques. One simple heuristic is to discard any views (without computation) whose cost is estimated to be large. Also, we may be able to leverage inter-dependencies between views for further pruning based on cost.

General Settings: SeeDB can be generalized to handle more elaborate discriminating views with little or no change. For instance, SeeDB can easily handle multiple group-by attributes, resulting in multi-column views that can be visualized as stacked bar-charts. Our discriminating views could include additional selection predicates (in addition to a group-by and an aggregation); for instance, in our staplers example, perhaps the trend line of total sales of staplers in California is an interesting visualization, because it differs from the total sales of staplers in the rest of the country. Overall, there is a rich space of general discriminating views we can consider.

Refinement of Visualizations: Since analysts are rarely interested in absolute values in their visualizations, we may be able to leverage ideas similar to those used in online aggregation [5] to produce visualizations that become more accurate over time.

Selecting Diverse Views: Our goal simply selects the K views with the highest utility, ignoring the fact that the discriminating views are related. For instance, an analyst may prefer one visualization each of sales and revenue, instead of two visualizations of sales, since the former covers more measure attributes. Incorporating such personalized preferences requires new models of discriminating view-diversity (leveraging metrics from recommendation systems) and hence new computation methods.

5. DEMO WALKTHROUGH

6. CONCLUSIONS

We outlined our vision for SeeDB: a system that provides analysts with visualizations highlighting interesting aspects of the query result. We defined several concrete problems in architecting SeeDB, relating to areas ranging from multi-query optimization and approximation to multi-criteria optimization. We believe that the general area of bringing visualizations closer to the DBMS is a challenging, yet, important direction for database research in the future.

7. REFERENCES

[1] K. Chakrabarti et al. Approximate query processing using wavelets. In *VLDB*, pages 111–122, 2000.

[2] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.

[3] P. B. Gibbons. Distinct sampling for highly-accurate answers to distinct values queries and event reports. In *VLDB*, pages 541–550, 2001.

[4] H. Gonzalez et al. Google fusion tables: web-centered data management and collaboration. In *SIGMOD Conference*, pages 1061–1066, 2010.

[5] J. M. Hellerstein, P. J. Haas, and H. J. Wang. Online aggregation. In J. Peckham, editor, *SIGMOD 1997*, pages 171–182. ACM Press, 1997.

[6] C. Jermaine et al. Scalable approximate query processing with the dbo engine. *ACM Trans. Database Syst.*, 33(4), 2008.

[7] S. Kandel et al. Wrangler: interactive visual specification of data transformation scripts. In *CHI*, pages 3363–3372, 2011.

[8] S. Kandel et al. Profiler: integrated statistical analysis and visualization for data quality assessment. In *AVI*, pages 547–554, 2012.

[9] M. Livny et al. Devise: Integrated querying and visualization of large datasets. In *SIGMOD Conference*, pages 301–312, 1997.

[10] J. D. Mackinlay et al. Show me: Automatic presentation for visual analysis. *IEEE Trans. Vis. Comput. Graph.*, 13(6):1137–1144, 2007.

[11] A. Parameswaran, N. Polyzotis, and H. Garcia-Molina. SeeDB: Visualizing Database Queries Efficiently. *Stanford Infolab*, 2013.

[12] S. Sarawagi. Explaining differences in multidimensional aggregates. In *VLDB*, pages 42–53, 1999.

[13] S. Sarawagi. User-adaptive exploration of multidimensional data. In *VLDB*, pages 307–316, 2000.

[14] G. Sathe and S. Sarawagi. Intelligent rollups in multidimensional olap data. In *VLDB*, pages 531–540, 2001.

[15] T. K. Sellis. Multiple-query optimization. *ACM TODS*, 13(1):23–52, 1988.

[16] C. Stolte et al. Polaris: a system for query, analysis, and visualization of multidimensional databases. *Commun. ACM*, 51(11):75–84, 2008.

[17] C. Wang and H.-W. Shen. Information theory in scientific visualization. *Entropy*, 13(1):254–273, 2011.

[18] Wikipedia. Jensen shannon divergence — wikipedia, the free encyclopedia, 2013. [Online; accessed 16-July-2013].

[19] Wikipedia. Kullback leibler divergence — wikipedia, the free encyclopedia, 2013. [Online; accessed 16-July-2013].

[20] Wikipedia. Statistical distance — wikipedia, the free encyclopedia, 2013. [Online; accessed 16-July-2013].