Assignment3 **CMSC 828D**

Trisha Mittal, Abhishek Bassan

SeeDB

The paper proposes SEEDB, a visualization recommendation engine to facilitate fast visual analysis. Given a subset of data to be studied, SEEDB intelligently explores the space of visualizations, evaluates promising visualizations for trends, and recommends those it deems most useful. They also try to take into account the teo major obstacles in this recommendation process. The two major obstacles in recommending interesting visualizations are (a) scale: evaluating a large number of candidate visualizations while responding within interactive time scales, and (b) utility: identifying an appropriate metric for assessing interestingness of visualizations. SEEDB introduces pruning optimizations to quickly identify high-utility visualizations and sharing optimizations to maximize sharing of computation across visualizations. They adopt a deviation based metric for visualization utility, i.e. a visualization is likely to be interesting if it displays large deviations from some reference.

Existing Evaluation

They present an evaluation of SEEDB both in terms of performance when returning visualizations and user studies for determiniting how valid the utility metric is.

Performance

They answer the following two questions.
1. How well do sharing and pruning optimisation improve latency?
2. How well do pruning optimisation affect the accuracy?
They look at 4 real-world datasets. For each dataset, they show the latencies obtained on the ROW and COL store by the basic SEEDB framework (NO OPT), by our sharing optimizations (SHARING), and by the combination of our sharing and pruning optimizations (COMB). We also show latencies for early result generation with COMB (COMB EARLY), where we return approximate results as soon as the top-k visualizations have been identified. They also show that the pruning doesn't compromise with the accuracy especially when the K value is high.

User Studies

They assess the utility metric of SEEDBs recommendations with real users. First, they perform a study to validate the deviation-based utility metric. They show that although simple, their deviation-based metric can find visualizations users feel are interesting. Second, they compare SEEDB to a manual charting tool without visualization recommendations. They show that SEEDB can enable users to find interesting visualizations faster and can surface unexpected trends.

Validating Utility Metric

To validate deviation as a utility metric, they obtain ground truth data about interestingness of visualizations and evaluated SEEDB against it. To obtain ground truth, they present 5 data analysis experts with the Census dataset and the analysis task of studying the effect of marital status on socioeconomic indicators. They presented experlts with the full set of potential

aggregate visualizations and asked them to classify each visualization as interesting or not interesting in the context of the task. They do the same with SeeDB's utility metric and plot a ROC.

SeeDB vs Manual Visualisation Tool
They recruit participants with some prior data analysis experience and visualization experience (e.g. R, matplotlib or Excel). They use the Housing and Movies datasets. Each participant was given a dataset and after a tutorial with SeeDB and the manual visualisation tool they were asked to perform two visualisation tasks. Over the course of each study session, they collected data by three means: interaction logs from each tool, responses to surveys, and exit interview notes. The interaction logs capture the number of visualizations constructed, the number of visualizations bookmarked, bookmark rate, and interaction traces. SEEDB and MANUAL both support the construction of different types of charts such as bar charts, scatterplots etc.

Strengths & Weaknesses

1. **Strengths**

   (a) They are using AUROC (Area under ROC) for judging the accuracy of their utility function. This is a better measure of a classification power of the model rather than just giving the accuracy and precision numbers which can be manipulated by choosing appropriate threshold of the utility value.

   (b) They are using movies and housing data set for their user study which can be easily understood by all the test subjects. This eliminates any domain knowledge based advantage or disadvantages that the test subjects might have.

   (c) In addition to reporting the number of bookmarks, they recognized that it depends on total number of visualization explored, and also included the bookmark rate.

2. **Weaknesses**

   (a) To establish the ground truth for the evaluation of utility metrics they used majority voting to determine the "goodness" of the visualization. This model uses a binary system to judge the utility of a visualization, a more granular system would be better representation of the truth.

   (b) They used a interface with no recommendation as the control set. This might not be a fair comparison as it takes more time to generate a new visualization then to select a visualization from a list. So the increase in number of explored visualization could be due to ease in selection.

   (c) They did not do any analysis on utility of visualization during user studies. This data could also be used to validate their claims.

   (d) Not strictly a weakness of the evaluation study, but the authors should have shared the data for user study

Proposed Evaluation Study

Re-design Goals

SeeDB is a visualization recommendation system. The main metric that we should be measuring is how good are the recommendations made by the system and how much does these recommendations help the end user. But, these recommendations require some prepossessing to

generate the recommendations. This time should be kept minimal but if it is less than certain threshold of acceptability, it would not affect the experience of the end user. The metrics used by authors to measure the optimization techniques are very comprehensive and covers all the necessary metrics that needs to be computed.

For measuring the effectiveness of the utility metrics the authors used AUROC metrics from a user survey of 5 data analysis expert. This is a good metric for judging a classifier but recommendation systems are closer to a ranking system than to a classifier. A good way of judging a recommendation systems system is by using Normalized Discounted Cumulative Gain (NDCG). For computing this metric, instead of telling user to classify a visualization as good or bad, they are told to rank them on a scale of 1-5. Then we compare the order of visualization obtained by average scores of the user with the order obtained by using the utility metric.

For the user study the authors are using a slightly different interface for treatment and control. The MANUAL Interface is same as the SeeDB interface with only difference being that the recommendation bar is removed. We think this approach gives a unfair advantage to SeeDB interface as users can create more visualization in same amount of time because it is faster to click on a recommendation than create a new visualization. So, we propose that control trials also have a recommendation bar but with randomly selected visualization. The users should not be told that one of the recommendation system is random. This will tell us how much better is the recommendation system from a random algorithm.

The SeeDB recommendation system is task agnostic but during the experimentation a specific task is assigned to each user. This could be problematic as we will have task dependent bias in the experiment. We propose to remove this bias by giving a more open ended task. We would also calculate the accuracy of the utility function by computing a truth table for each participant for the bookmarked and explored visualization. For example, if a user has explored 20 visualizations then we restrict our truth table to only those 20 visualization. In the end we can add up all truth tables to find the accuracy and precision of our utility function. This is not the primary metric for utility function (as we have discussed above, we believe it is closer to a ranking problem than a classification) but just a validation of what we had explored before. Also we would like to share the data for the user studies while complying with all the privacy policies. We found that some of the descriptions of user study by the author were very vague for example the participants were told to "use the specified tool to find visualizations supporting or disproving a specific hypothesis."In this case we do not know what kind of hypothesis was presented and whether it could be biased to do well for their recommendation. Also, they have a data set labeled by experts which can be used by others for evaluating their recommendation algorithms.

Potential Participant pool

In the original study, they recruit participants who have some prior data analysis experience. We expect this tool to be used by professionals and students. To be able to get a comprehensive evaluation, we would want to recruit participants from this pool. We would propose to recruit around 30 people

Study Setup

We would be using in-person study. This allows us to control the variables external to the experiments like speed of the machine and network speed. The experiment part of the user

study would take approximately 20 minutes. We would also like to set 15 minute of time aside for a brief exit interview and tool specific survey . We estimate that each user should be assigned 45 minutes of time. We would give them a standard compensation of $15.

Relevant Tasks

In the existing study, they check for a validity of a hypothesis. They provide every participant with a dataset and a hypothesis to validate. The recommendation system is agnostic to the visualization task and we want to setup a broader objective function which can test the general utility of visualization rather than the utilty in the context of a task. So, to better assess the utility of SeeDB interface we propose that we would give the participant 8 minute of time to explore the dataset with the objective of finding interesting facts. We will tell them that they can bookmark the interesting visualization. After this time period they will have access to only the bookmarked visualization and they can list down the interesting facts that they have found and why they think that these are interesting. While we will measure metrics for only the experiment but this additional data can be used to come up with new recommendation system.

Test Dataset

The two datasets that are being used in the original study for the evaluation is a Housing and the Movies dataset. These datasets were chosen by the authors because they were easy to understand and comparable in size and number of potential visualizations.We believe this is a good choice of datasets as the participants do not require any prior understanding to be able to perform visualisation analysis. For the housing dataset, they use the Boston Housing Dataset, which has around 500 records, 4 attributes and 10 measures. This is a dataset derived from information collected by the U.S. Census Service concerning housing in the area of Boston Mass. The dataset can be found here https://www.cs.toronto.edu/ delve/data/boston/bostonDetail.html. For the Movie sales dataset, we were unable to trace the actual source that they claim to use.