

**Active Learning View Recommendation for**

**Visual Data Exploration**

**Rischan Mafrur**

**44667649**

**Principal Advisor: Dr Mohamed A. Sharaf**

# Abstract

Data visualization is one of the most important parts of interactive data exploration. It is often used as the opening step in performing various analysis tasks. There are several powerful current data visualization tools which widely used such as Tableau and Microsoft Power BI. However, those data visualization tools still require manual effort and trial-error process to specify visualizations that is a labour-intensive and time-consuming process. Moreover, those current data visualization tools focus on question answering visualization which is assume that users have good knowledge of the datasets whereas this assumption is not always true. In order to support effective interactive data exploration, there has been a growing interest in developing solutions that can automatically recommend data visualizations that reveal interesting and useful insights. There are three main challenges to support automatically recommend data visualizations which we want to focus in this study are: automatically present the most important visualizations from high dimensional datasets, support an iterative exploration model to discover and present the most important visualizations that relevant to user interest, provide an interactive performance due to it is dealing with the limitation of the user wait time. To overcome those challenges, this study presents two novel schemes: Diversifying view recommendation for visual data exploration (DiVE) and Active learning view recommendation for visual data exploration (ALiVE). Current experimental results show that our proposed DiVE scheme is able to improve the quality of recommended visualizations and provide efficient pruning scheme which can reduce processing cost significantly compared to the baseline methods.

Contents

[Abstract 2](#_Toc519612294)

[Introduction 4](#_Toc519612295)

[Research Problem 4](#_Toc519612296)

[Challenges 4](#_Toc519612297)

[Expected Research Outcomes 4](#_Toc519612298)

[Report Stuctures 4](#_Toc519612299)

[Approach and Methodology 4](#_Toc519612300)

[Data-driven approach 5](#_Toc519612301)

[User-driven approach 5](#_Toc519612302)

[Hybrid approach 6](#_Toc519612303)

[Schedule and Timeline 7](#_Toc519612304)

[References 8](#_Toc519612305)

# Introduction

In the recent years with an exponential growth of available data in various domains, there has been an increase in the number of people who try to gain insights from the dataset called *data analyst* [1]. Generally, data analyst uses visualization tools such as Spotfire, Tableau, Google Table Fusion, Microsoft Power BI, Qlik, and etc to perform several analytical tasks. Despite those tools provide a powerful data analysis toolbox, however, without any prior knowledge about the data, the analyst must manually specify different combinations of dimensions, measures and aggregate functions before finally generating a visualization that reveals some insights from the dataset. In fact, manually looking for insights in each visualization is a labour-intensive and time-consuming process.

Such challenge motivated multiple research efforts (e.g., [2]–[8]) that focused on developing view recommendation that can automatically recommend views based on some metrics that capture the utility of a recommended views*.* In order to develop view recommendation, there are two approaches can be broadly classified as user-driven approach and data-driven approach. User-driven solution recommend set of views that focus on user intent or task. For instance, VizDeck [2] that generate all possible views and request the feedback from the user. User feedback is used as the base knowledge to understand the user preference for the future recommendation. Other previous approaches which working on user-driven such as Profiler [8], Rank-by-Feature Framework [9], and etc. Meanwhile, the data-driven approach focuses on the discovery of insights from the dataset and recommend visualizations based on data characteristics [4]–[7], [10].

To that end, the contributions of this study will be divided into three parts: 1) data-driven approach, 2) user-driven approach, 3) hybrid approach. In the data-driven approach, we improve SeeDB work [4] by employing diversification and applying the efficient pruning scheme. Meanwhile, in the user-driven approach, we will focus on how to identify the user’s intent or the task, it may use the explicit approach such as provide the options in advanced (e.g. ask the analyst what kind of task that she wants to do) or by using the sequence of action log which performed by the analyst. While the hybrid approach is the combination of our proposed data-driven and user-driven.

## Research Problem

Given a high dimensional dataset that has a number of attributes and measures, how to develop an intelligent scheme that can automatically present the most important visualizations? The aims is to propose a novel scheme

## Challenges

There are three main challenges to support Given a high dimensional dataset that has a number of attributes and measures, how to develop an intelligent scheme that can automatically present the most important visualizations? The aims is to propose a novel scheme

## Expected Research Outcomes

## Report Stuctures

# Approach and Methodology

## Data-driven approach

In terms of data-driven approach, recent case studies have shown that "**a deviation-based metric**" to be effective in providing the “most important” visualization (*top-k views*) [4], [5]. In this work, we adopt a deviation-based metric to expose the quality of the individual view. However, the drawback of only rely on deviation-based metric is that often deliver redundant recommended views, which leads to presents limited insights of results. To address that limitation, in this work we posit that employing diversification techniques in the process of view recommendation allows eliminating that redundancy and provides concise coverage of the possible insights to be discovered. In fact, novelty and diversity are one of the fundamental characteristics of any effective recommendation systems [11]–[19].

In order to recommend views that consider both importance and diversity, we propose the hybrid objective function that integrates two components: 1) the total importance score of the recommended set of views and 2) the diversity score of the recommended set of views. Speciﬁcally, an objective function is formulated as the linear weighted combination of the importance score, and diversity function which is deﬁned as:

where is employed to control the tradeoff between Importance and diversity. The higher values of result in a set of more diverse views whereas lower values of generate a set of the most important views that might be similar to each other.

In fact, existing research has shown that recommending views based on deviation-based approach is a computationally expensive task [4], [7]. Moreover, integrating diversification into the view recommendation problem further increases that computational cost. To address that challenge and minimize the query processing cost, we propose an integrated scheme that leverages the properties of both the importance and diversity to prune a large number of low-utility views without reducing the quality of recommendations.

## User-driven approach

As mentioned above that deviation-based metric [4], [5] can be the efficient way to provide the most important views based on data-driven approach but it only solves one of the user task types. While analyst does not has any idea and she wants blindly to know the most interesting visualizations from the dataset, the deviation-based metric can be the appropriate solution. However, data-driven approach (e.g. deviation-based metric) is not able to handle three main issues: 1) what kind of tasks that analyst wants, 2) the analyst expertise (e.g. the analyst knowledge, whether she is an expert or not), and 3) the analyst preferences. Hence, the user-driven approach can be the solution.

In this work, we focus on three main issues on user-driven approach as mentioned above. In order to understand the user’s intent, preference, and expertise, we need to understand the types of the task while analyst performing the analysis. This work [20] has been explained three aspects of task are: a) style of analysis (e.g. comparative, exploratory, predictive, and targeted), b) subject of analysis (e.g. the analyst has a target of attributes that want to be analysed), c) goal of analysis (e.g. the analyst has a specific goal). Based on these following aspects, deviation-based metric only solve one of the types of tasks that is while the style of the analyst is exploratory and she does not have b) and c).

To address that limitation, in this work have two techniques as follows:

1. The explicit technique, it can be defined on the application interface. We may utilize the drop-down menu to show the options of style analysis to the analyst, this technique can help us understand the state of the analyst (e.g. tasks types and expertise). For instance, while analyst has target attributes or the specific goal, it may be considered that analyst has an expertise in the dataset. Another popular way in the explicit technique to get the user preference is by asking the feedback. We may use approaches such as in [2], [3] to get the feedback from the analyst.
2. The implicit technique, we may record all user’s actions (e.g. click) and use machine learning technique to build the user preference model.

We will use active learning [21], [22] to process all data from the explicit and implicit techniques and build the user preference model. The user preference model can be used as the base for recommending views.

## Hybrid approach

In the data-driven approach, we adopt the deviation-based metric and employ diversification to avoid the redundancy in the recommended set of views. Moreover, we propose an efficient pruning scheme to prune a large number of low-quality views without reducing the quality of the results. Meanwhile, we also propose a user-driven approach using the explicit and implicit techniques to understand the user state, then using active learning to build the user preference model. A hybrid approach utilizes the combination of both data-driven and user-driven to build the effective and efficient view recommendation.

1. Kita kan punya SeeDB yang mana modelnya adalah explartory dan user gak reti opo-opo langsung direkomendasikan views. Iki yowis titik. Dadi paling interesting terus ditambah diversification ben gak redundant tur luwih efficient mergo pruning

2. Dari SeeDB kita tak bisa belajar user interestnya. Intinya pingin coba untuk mempersembahkan ke user sample views terus nanti minta feedbacknya.

3. DeepEye kan juga user masukin parameter, db, bins dsb abracadabra langsung metu hasil.

Nah kita pingin coba schema mirip kayak AIDE, yaitu coba sampling, terus nganggo active learning

An Active learning-Based Approach for View Recommendation in Data Exploration

Liat video nya AIDE jadi paham. Klo di implementasikan di visualization, tiap vis di klik kan kedetek x, y dan pattern nah itu yang dijadikan patokan rekomendasi sebelumnya. Dan intinya kan yg direkomendasikan pertama banyak sampling kayak AIDE

Nah coba baca yang VizDeck sama yang Profiler atau yg lain, bedanya apa

Rischan Mafrur received a Master Degree in Electrical and Computer Engineering from Chonnam National University, South Korea in 2015. He has commenced his PhD studies at the School of ITEE, The University of Queensland in October 2017. His research interests include Data Visualization, Data Exploration, and Recommender Systems.

# Schedule and Timeline

|  |  |
| --- | --- |
| **Tasks** | **Date** |
| Data-driven view recommendations   * Read and summarize literature reviews related to a data-driven approach to view recommendations. * Find the gap from the previously proposed approach * Propose an idea that can improve the quality of the recommended views. * Propose diversification technique to avoid redundancy while recommending views. * Propose new objective function for recommending views which are based on relevance and diversity. * Propose new technique to reduce the cost while generating recommended views. * Write and submit a paper to CIKM 2018 (May 23, 2018) * Prepare the extended version of this work for the Journal submission. * Write and submit a paper to IEEE Journal TKDE | October 2017 – September 2018 |
| **Confirmation Milestone** | October 2018 |
| User-driven view recommendations   * Read and summarize literature reviews related to user-driven approach on view recommendations. * Find the gap from the previously proposed approach. * Find a way to propose a new idea on user-driven view recommendations without any human involvement. (avoid to deal with an ethic clearance which may take time) * Propose a new technique that can improve the quality of recommended views and the efficiency in terms of user-driven approach. * Write and submit a paper to ICDE 2020 | November 2018 – September 2019 |
| **Mid-candidature Review Milestone** | October 2019 |
| Data-driven and user-driven view recommendations.   * Combine between our proposed approach on data-driven and our proposed approach on user-driven to improve the quality of recommended views. * Write and submit a paper of our combination approach to IEEE Journal. | November 2019 – December 2020 |
| **Thesis Review Milestone** | January 2021 |
| **Thesis write up and submission** | January – June 2021 |

# References

[1] K. Morton, M. Balazinska, D. Grossman, and J. Mackinlay, “Support the Data Enthusiast: Challenges for Next-Generation Data-Analysis Systems,” *Proc. VLDB Endowment, Vol. 7, pp. 453–456, 2014*, vol. 7, pp. 453–456, 2014.

[2] A. Key, B. Howe, D. Perry, and C. R. Aragon, “VizDeck: self-organizing dashboards for visual analytics,” *SIGMOD Conf.*, pp. 681–684, 2012.

[3] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon, “Many Eyes: A site for visualization at internet scale,” *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1121–1128, 2007.

[4] M. Vartak, S. Rahman, S. Madden, A. Parameswaran, and N. Polyzotis, “SEEDB : Efficient Data-Driven Visualization Recommendations to Support Visual Analytics,” *VLDB Proc. VLDB Endow.*, vol. 8, no. 13, pp. 2182–2193, 2015.

[5] M. Vartak and S. Madden, “S EE DB : Automatically Generating Query Visualizations,” *Proc. 40th Int. Conf. Very Large Data Bases*, vol. 7, no. 13, pp. 1581–1584, 2014.

[6] H. Ehsan, M. Sharaf, and P. K. Chrysanthis, “Efficient Recommendation of Aggregate Data Visualizations,” *IEEE Trans. Knowl. Data Eng.*, vol. 4347, no. c, pp. 1–1, 2017.

[7] H. Ehsan, M. A. Sharaf, and P. K. Chrysanthis, “MuVE: Efficient Multi-Objective View Recommendation for Visual Data Exploration,” *2016 IEEE 32nd Int. Conf. Data Eng. ICDE 2016*, pp. 731–742, 2016.

[8] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, “Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment.”

[9] Jinwook Seo and B. Shneiderman, “A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections,” in *IEEE Symposium on Information Visualization*, pp. 65–72.

[10] P. Hanrahan, “VizQL: a language for query, analysis and visualization,” *Proc. 2006 ACM SIGMOD Int. Conf. Manag. data - SIGMOD ’06*, p. 721, 2006.

[11] B. Smyth and P. Mcclave, “Similarity vs . Diversity,” no. Section 2, pp. 347–361, 2001.

[12] C. Yu, L. Lakshmanan, and S. Amer-Yahia, “It takes variety to make a world: diversification in recommender systems,” *EDBT ’09 Proc. 12th Int. Conf. Extending Database Technol. Adv. Database Technol.*, pp. 368–378, 2009.

[13] M. Zhang and N. Hurley, “Avoiding Monotony: Improving the Diversity of Recommendation Lists,” *Proc. 2008 ACM Conf. Recomm. Syst.*, pp. 123–130, 2008.

[14] C. L. A. Clarke *et al.*, “Novelty and diversity in information retrieval evaluation,” *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR ’08*, p. 659, 2008.

[15] D. Rafiei, K. Bharat, and A. Shukla, “Diversifying web search results,” *Proc. 19th Int. Conf. World wide web WWW 10*, p. 781, 2010.

[16] M. R. Vieira *et al.*, “On query result diversification,” *Proc. - Int. Conf. Data Eng.*, pp. 1163–1174, 2011.

[17] S. Gollapudi and A. Sharma, “An Axiomatic Framework for Result Diversification.,” *IEEE Data Eng. Bull.*, vol. 32, no. 4, pp. 7–14, 2009.

[18] G. Adomavicius and Y. Kwon, “Diversity Using Ranking-Based Techniques,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 896–911, 2012.

[19] K. Zheng, H. Wang, Z. Qi, J. Li, and H. Gao, “A survey of query result diversification,” *Knowl. Inf. Syst.*, vol. 51, no. 1, 2017.

[20] M. Vartak, S. Huang, T. Siddiqui, S. Madden, and A. Parameswaran, “Towards Visualization Recommendation Systems,” *ACM SIGMOD Rec.*, vol. 45, no. 4, pp. 34–39, 2017.

[21] Y. Diao *et al.*, “AIDE: An Automatic User Navigation System for Interactive Data Exploration.”

[22] K. Dimitriadou, O. Papaemmanouil, and Y. Diao, “AIDE: An Active Learning-Based Approach for Interactive Data Exploration,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 2842–2856, 2016.