THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Active Learning View Recommendation for Visual Data Exploration

PhD Confirmation

Candidate : Rischan Mafrur
Email : r.mafrur@uq.edu.au

Principle Supervisor : Mohamed A. Sharaf
Email : m.sharaf@uq.edu.au
Associate Supervisor : Hina A. Khan
Email : h.khan3@uq.edu.au

School of Information Technology and Electrical Engineering
The University of Queensland, Australia
October 2018

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 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 (*views*) that reveal interesting and useful insights. There are three main challenges to support automatically recommend views which we want to focus in this study are: automatically present the most important views from high dimensional datasets, support an iterative exploration model to discover and present the most important views that relevant to user interest, provide an interactive performance to deal with multiple users and the limitation of the users 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 views and provide efficient pruning scheme which can reduce processing cost significantly compared to the baseline methods.

Table of Contents

1. 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 enthusiast* [1]. Generally, data enthusiast 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, she must manually specify different combinations of dimensions, measures and aggregate functions before finally generating a visualization (*view*) that reveals some insights from the dataset. In fact, manually looking for insights in each view is a labour-intensive and time-consuming process. Moreover, current visualization tools which mentioned above are focus on question answering which assume that the users has a good knowledge about the dataset which is this assumption is not always true.

Such challenge motivated multiple research efforts (e.g., [2]–[8]) that focused on developing views recommendation that can automatically recommend views based on some metrics that capture the utility of a recommended views. In order to develop views 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.

1.1. Research Problem

Given a high dimensional dataset such as has a number of attributes and measures, how to develop an interactive data exploration scheme that can automatically present the most important views which match with the user interest? The aim is to propose a novel interactive data exploration scheme that can accuratly and efficiently recommend the most important views which relevant to the user interest.

1.2. Challenges

There are three main challenges to support automatically recommend data visualizations are as follows:

- **Challenge 1: Automatically present the most important views from high dimensional datasets.** A large number of possible views are generated and ranked according to some metric of importance, then the top-k most important views are recommended to the user. There are two main issues in this challenge: a) metric of importance that can discover the most important views; b) the solution should be able to deal with the large number of generated views and it must an efficient approach that has an interactive performance.
- **Challenge 2: Support an iterative exploration model to discover and present the most important views that relevant to user interest.** To present the most important views such as in the challenge 1, it only needs a single iteration. However, in order to recommend views which important as well as relevant to user interest, it needs user feedback. To capture user interest, we use an active learning apporach to present some sample of views to get user feedback. The solution should support multi-iteration exploration and it should able to build user exploration profile that can be used for the future recommendation.
- **Challenge 3: Provide an interactive performance to deal with the limitation of the users wait time.** To discover the most important views, a large number of views must be generated where only a small fraction of those views are actually of interest and are candidates to be included in the top-k set. Moreover, in order to get the user feedback, an iterative data exploration model is needed. Hence, the solution should have an interactive performance as the limitation of user wait time.

1.3. Expected Research Outcomes

The expected of our research outcomes as follows:

- We propose *Diversifying View Recommendation for Visual Data Exploration (DiVE)* to automatically present the most important views effectively and efficiently.
- We propose *Active Learning View Recommendation for Visual Data Exploration (ALiVE)* for providing an iterative exploration model which based on active learning approach to get user feedback and present the most important and relevant views to her.
- We design *ALiVE* to handle a high dimensional dataset and provide an interactive performance which is less than few seconds in each iteration of exploration.

1.4. Report Structures

The rest of the report is organised as the following. Section 2 explains background in data driven machine learning for critical care and its related work in time series mining, multivariate feature learning, and real-time recommendation. Section 3 provides formal problem formulation and

details of our proposed methods. Experimental evaluation and preliminary results are given in Section 4. Summary and future works are discussed in Section 5 followed by timeline of remaining PhD project in Section 6.

2. Related Work

2.1. View Recommendation

There are a lot of visualization tools which commonly used by users such as MS Excel, Tableau, and Spotfire until tools which need programing language capability to use it such as Rstudio, Spider, D3js, etc. However, most visualization tools that exist today do not support auto-generated recommended views, as we know Tableau starts to provide the feature 'Show Me'[9], that provides recommended chart type of visualization. In terms of visualization recommendation system, there are three approaches as follows:

- Data-driven approach: presenting recommended views based on data characteristics or data characteristics oriented.
- User-driven approach: presenting recommended views based on given information by the user related to her preference or user characteristics oriented.

In data-driven approach, there are [10] proposed Polaris and later [11] used it as a formal declarative visual language for data visualization called VizQL. Tableau that we said before, used VizQL to automatically generates recommended chart types. It happens when the user starts to select the attribute of the dataset, Tableau 'Show Me'[9] provides recommended charts that match the selected attributes. However, this feature only recommends chart types not recommend the subset which has interesting trends. The recent work of data-driven approach is Voyager [12] which based on Vega-lite [13], Voyager uses statistical properties of the data to generate recommended views. Vega-lite is new high-level specification language, it using JSON object to describe the data source, it also called a new grammar of interactive graphic. Another recent work that purely using data-driven approach is SeeDB [2], [3], the authors using a statistical method to compute the probability distribution of each subset of the dataset. They compare the probability distribution of the selected subset to another subset or whole dataset. The subset which has a high deviation/distance matrics from reference subset (another subset/whole dataset) can be defined as interesting views. On the other hand, an example work based on user-driven approach was conducted by [14]. The work which based on data-user-driven approach conducted by [15], called VizDeck. They used a data-driven approach such as statistical properties of the data to provided result views. Then the system presented all result views like in a card game metaphor, the user should keep the good views which may seem interesting to her and discard the unwanted views. By this feedback, the system learns user preferences and it can make the system able to present more suitable visualizations in the future.

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.

2.2. Result Diversification

To understand user interest and giving the recommended items in recommendation system is a non-trivial task, there have been many studies in developing algorithms that boost prediction accuracy of recommended items but it was not enough. In some case, high accuracy produces homogenous recommended items and high accuracy does not guarantee users satisfactory. Some researchers proposed the importance of diversification, they argued that diverse items mean more opportunities for users to get the satisfied items [16]. There are several works which proposing diversification in recommendation system [5], [8]. Moreover, this comprehensive survey [17] explained details about the definition of diversification, its classification, also including techniques and algorithms, and real implementation of diversification such as in database system, recommendation system, search engines and soon. There are various definitions of diversity, but most of them can be classified in one of this categories: (i) content-based diversity, means selecting results based on dissimilarity to each other [18][19]; (ii) novelty-based diversity selects the results that contain new information compared to the previous results which have been presented to the user[6]; (iii) semantic-based diversity selecting results that based on categories or topics[7]. Diversification also has been known as the NP-hard problem [20]. Several techniques have been proposed for results diversification, there are two kinds of common techniques are incremental and bounding techniques. Mostly, previous works on result diversification using incremental way, that constructs the result one by one and stop as soon as k results generated. The best known algorithm for this technique is greedy solution[8], [18], [21]. This work [22] explained and proposed a general framework of those two techniques and this work [23] proposed several natural axioms which showed that there is no diversification objective that can satisfy all the axioms simultaneously. Another work [18] presented an experimental evaluation of various common query result diversification techniques such as greedy, swap, random, motley, clustering, etc.

2.3. Active Learning

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.

3. Approach and Methodology

3.1. Diversifying View Recommendation for Visual Data Exploration

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, $I(S)$ and diversity function $f(S, D_{Cx})$ which is defined as:

$$\mathcal{F}(S) = (1-\lambda).I(S) + \lambda.f(S, D_{Cx})$$

where $0 \leq \lambda \leq 1$ is employed to control the tradeoff between Importance and diversity. The higher values of λ result in a set of more diverse views whereas lower values of λ 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.

3.2. The Improved Schemes

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.

4. Experiment and Evaluation

4.1. Current Results and Analysis

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.

4.2. Future Experiments

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.

5. Summary

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 Machine Learning.

6. Timeline

Tasks	Date
Diversifying View Recommendation for Visual Data Exploration • Read and summarize literature reviews related to a 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)	October 2017 – September 2018
Confirmation Milestone	October 2018
• Prepare the extended version of this work for the Journal submission. • Write and submit a paper to IEEE Journal TKDE *Active Learning View Recommendation for Visual Data Exploration* • 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
Extend previous works • 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

Bibliography

[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.

Appendices 1: Publication