THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Efficient Conversational 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.

Toward supporting automatically recommending views, there are three main challenges which we want to focus on this study are: automatically present the most important views from high dimensional datasets, support conversational model (i.e., an iterative exploration model) to discover user interest and present the most important and relevant views to the user, provide an interactive performance to deal with high dimensional datasets, 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 by considering diversity to avoid redundancy. Moreover, DiVE provides efficient pruning scheme which can reduce processing cost significantly compared to the baseline methods. However, DiVE only support single iteration and does not include user preferences to generate the recommended views. To overcome this issue, we propose efficient conversational model view recommendation ALiVE, which supports iterative data exploration and uses active learning paradigm to provide a better recommendation.

Table of Contents

1. Introduction

In the recent years with an exponential growth of available data in various domains, there has been an increasing in the number of people who try to gain insights from the dataset which is called *data enthusiast* [1]. Generally, data enthusiast as the user uses visualization tools such as Spotfire, Tableau, Google Table Fusion, Microsoft Power BI, and Qlik to perform several analytical tasks. Despite those tools provide a powerful data analysis toolbox, however, those tools mostly focus on question answering which assume that the user have a good knowledge about the dataset which this assumption is not always true. Without any prior knowledge about the data, user 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.

Such challenge motivated multiple research efforts (e.g., [2]–[8], [9]) that focused on developing views recommendation that can automatically recommend views based on some metrics that capture the utility of a recommended views. Recent case studies have shown that "a deviation-based metric" to be effective in providing the most important views [4], [5]. The main goal of deviation-based approach is to automatically provide top-k most important views (*top-k views*), which are selected from all possible visualizations. The top-k views are selected based on the most important views in the dataset. The importance of a view is defined on the basis of deviation between the queried subset of data (target view) and the reference subset of data (reference view). The reference subset can be another subset of the dataset, the rest of the dataset, or the whole dataset. The intuition behind this approach is that views that reveal substantially different trends from the reference views are likely to be important to the user.

However, although the deviation-based approach can automatically provide the top-k most important views, it often recommends similar views and leaving the analyst with a limited amount of gained insights. To address that limitation, in this work we propose ***DiVE scheme*** which posit that employing diversification techniques in the process of view recommendation allows eliminating that redundancy and provides a good and concise coverage of the possible insights to be discovered. Moreover, DiVE leverages the properties of both the importance and diversity metrics to prune a large number of query executions without compromising the quality of recommendations.

Currently, DiVE scheme recommend top-k views in single iteration and the top-k views are generated based on importance (i.e. deviation-based metric) and diversity. In order to capture the user interest, ***ALiVE scheme*** will be proposed. ALiVE scheme uses active learning approach that can minimize the user cognitive cost by selecting proper items (i.e., as few as possible) that can acquire user interest at the most. The presented views are the most important views as well as the most useful views to be used for discovering the user preferences. In each iteration, recommender engine captures the user preference and recommending views until the result is converged to the user want. Active learning paradigm has been mentioned to be an effective technique in recommender system to achieve high accuracy which only using a few labeled samples [10]. As the views recommender, ALiVE also generates user profile's exploration which can show user preferences based on her feedback. Moreover, ALiVE must be designed as an efficient scheme due to an interactive data exploration must be fast enough to enable iteration. This study [11] shows that analysts will lose effectiveness when it takes more than 500ms to get the results.

1.1. Research Problem

Given a high dimensional dataset with a high 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 and relevant views to the user.

1.2. Challenges

There are three main challenges to support automatically recommend views 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 conversational model (i.e., support multi-iteration exploration) by employing an active learning apporach to present most important views (i.e. unlabeled sample) to be labeled as relevant or irrelevant by user. Since the opportunities to get the user feedback are few due to the user may unwilling to give a label to many views, we should be sure as possible that presented views can tell us something *important* regarding her preferences. Hence, active learning can be one of the solution which is to minimize the cognitive cost of the user by selecting the unlabeled sample to be labeled by user carefully. Moreover, this proposed scheme should able to build user exploration profile that can be used for the future recommendation as well.
- **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 exploration process is needed. In order to support this challenge, the proposed schemes should have an interactive performance in each iteration as the user has the limitation 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 in single iteration.
- We propose *Active Learning View Recommendation for Visual Data Exploration (ALiVE)* for providing conversational model (i.e., iterative exploration) which follows active learning paradigm to capture user preferences and present the most important and relevant views to the user in each iteration.
- We design *ALiVE* for optimistic visualization tool to handle a high dimensional dataset, multi-users and provide an interactive speed in each iteration of exploration.

1.4. Report Structures

The rest of the report is organised as the following. Section 2 explains the related work in views recommendation, diversification, and active learning paradigm. Section 3 provides 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

In order to develop views recommendation, there are two approaches can be broadly classified as data-driven approach (i.e. data characteristics oriented) and user-driven approach (i.e. user preferences oriented) as follows.

In data-driven approach, there are [12] proposed Polaris and later it is used as a formal declarative visual language for data visualization called VizQL [13]. Tableau that one of the popular data visualization tool used VizQL to automatically generates recommended chart types that match to the selected attributes, this feature is called ‘Show Me’[14]. It happens when the user starts to select the attribute of the dataset. However, this feature only recommends chart types not recommend the most important views. Another recent work of data-driven approach was Voyager [15] which used Vega-lite [16] as the backbone. Voyager uses statistical properties of the data to generate recommended views where Vega-lite is a new high-level specification language (i.e. a new grammar of interactive graphic) which using JSON object to describe the data source. Another recent work that purely using data-driven approach also was conducted by [4], [5], the authors using a statistical method called “deviation-based metric” to expose the important views as explained in the Introduction section.

Meanwhile, an example work based on user-driven approach was conducted by [9]. This study is known as behaviour driven visualization recommendation. This study focuses on user preferences, all user’s behaviour while interacting with the system (e.g. swapping, flipping, and drill-down) is extracted to get the patterns. In the background, the recommendation engine interprets the user intents based on the patterns. Another popular user-driven work also conducted by [2], called VizDeck. This work used statistical properties of the data to present views such as in a card game metaphor. Afterwards, user should keep the good views that may seem interesting to her and discard the unwanted views as her feedback. All those feedback are stored and processed as a basis for the future recommendation. To rank the recommended views, they used statistical train model that relates to particular view and used user event log as the ground truth.

2.2. Result Diversification

To understand user interest and giving the relevant recommended views is a non-trivial task. As explained in the Introduction section that high accuracy may produce homogenous recommended views and high accuracy does not guarantee users satisfactory. Due to of that, some researchers proposed the importance of diversification, they argued that diverse items mean more opportunities for users to get the satisfied items [17]. In this work, we consider diversification as an important element to be used on recommending views. There are several previous works which proposing diversification in recommendation system [18], [19]. Moreover, this comprehensive survey [20] 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 [21], 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 [22][23]; (ii) novelty-based diversity selects the results that contain new information compared to the previous results which have been presented to the user[24]; (iii) semantic-based diversity selecting results that based on categories or topics[25].

2.3. Active Learning

To discover user interest on recommender system, several items are presented to the user. The user then gives a label as relevant and irrelevant based on her preferences. However, users are reluctant to give a label to many items and not all of labeled items by user reflects same usefulness to obtain the user preferences. Hence, Active Learning [26] [27] is designed to identify a few items that can gain understanding about user’s preferences at the most. Active Learning approach selects the most important unlabeled items to be presented and request the user to give a label to the presented items. This strategies are used to improve the recommendation engine and minimize the cognitive cost for the user by selecting proper items (i.e., as few as possible) that can acquire user interest at the most.

Active learning paradigm has been used in the interactive data exploration by this work [10]. This work integrates the active learning approach and heuristics exploration technique to acquire samples. The authors used a decision tree as the classification model to identify the next promising areas of data exploration, the aims is to minimize the number of samples which want to be presented to the user.

3. Approach and Methodology

3.1. Diversifying View Recommendation for Visual Data Exploration

As described in the Introduction section, we adopted a deviation-based metric to quantify the importance of view [4], [5]. This technique used content-driven importance score to expose the quality of the individual view. However, rely on content-driven importance feature leads to a set of similar views. In order to provide full coverage of all possible interesting insights, in this work, we not only consider the quality of individual view based on importance score but also consider diversity within the set of recommended views. Toward measuring the diversity of the set of recommended views, context-driven distance is introduced in this work. Figure 1 illustrates and summarizes of content and context driven metrics.

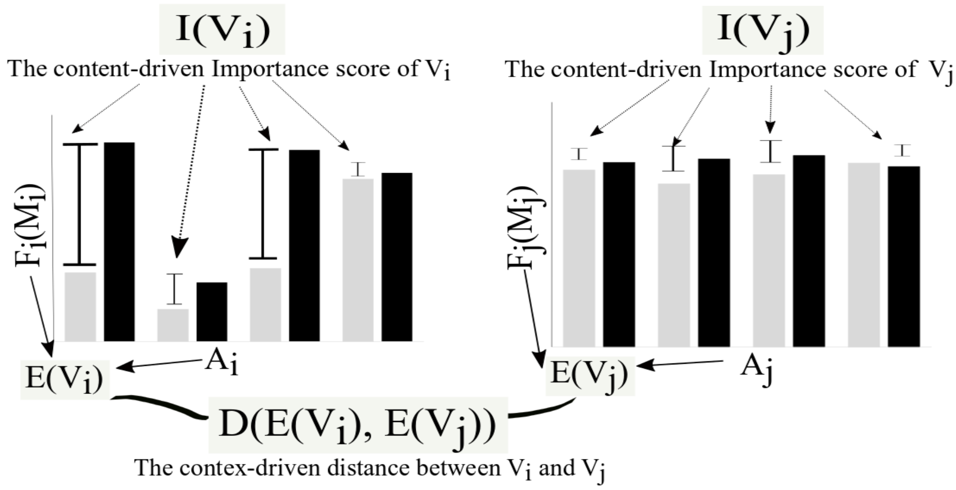

I(Vi)
The content-driven Importance score of Vi
I(Vj)
The content-driven Importance score of Vj
Fi(Mi)
Fj(Mj)
E(Vi)
Ai
E(Vj)
Aj
D(E(Vi), E(Vj))
The contex-driven distance between Vi and Vj

Figure 1. Content vs. Context of views

To consider both importance and diversity, we propose the hybrid objective function that integrates both 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:

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

where $0 \leq \lambda \leq 1$ is employed to control the tradeoff between importance and diversity. The higher values of λ produces 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. The details of our schemes as follows:

- Linear-Importance and Greedy-Diversity: Recommends top-k views on the basis of only importance and only diversity respectively. These both schemes are the baseline.
- DiVE-Greedy: Recommends top-k views on the basis of hybrid objective function using greedy algorithm.
- DiVE-iSwap: Selects top-k views on the basis of hybrid objective function using swap algorithm initialized by Linear-Importance. Due to the initial set S is generated by Linear-Importance, this scheme cannot escape from executing all query views.
- DiVE-dSwap: Selects top-k views on the basis of hybrid objective function using swap algorithm initialized by Greedy-Diversity.
- Static-pruning: DiVE-Greedy and DiVE-dSwap methods using static pruning technique (i.e., use static maximum bound on pruning) to reduce the number of view queries.
- Adaptive-pruning: DiVE-Greedy and DiVE-dSwap using adaptive pruning method (i.e., use adaptive maximum bound on pruning). The maximum bound will be updated while number of sample views are executed. We rely on non-parametric predictive interval models (*PI*) to determine its value with a certain level of confidence without any assumption on the population [28].

3.2. The Improved Schemes

There are two main improvement plans for DiVE scheme: 1) diversification function; 2) rectifying error bound on adaptive pruning method. Firstly, the current results that we have are using MaxSum diversification. MaxSum uses average score of diversity of the set S which is by computing the total sum of all distances then dividing by $k * (k - 1)$ while another diversification function such as MaxMin uses the maximum of minimum score of distance in the set S. Hence, the range diversity score from those both approaches are different that may result different performance especially for the pruning. Secondly, as shown in Figure 4 that by using PI = 80, DiVE can prune more queries but there are some errors in terms of effectiveness because of an error in maximum bound. If there is a way to use PI = 80 without reducing the quality of recommended views which is by rectifying error bound, it will be impressive.

Although DiVE has better performance compared to the baseline methods, however, DiVE only support one iteration to generate recommended views. For instance, given a high dimensional dataset D_B (e.g., Flight dataset). To generate all possible views from D_B, all subsets must be known, the subset is described as '**`WHERE`**' clause, for example: '`carrier = US`', '`carrier=AA`', '`origin=PHX`', '`destination=BNA`' and etc. From each subset high number of views are generated depends on the number of attributes, measures, and aggregate functions that is used. In case of DiVE scheme, each subset in D_B will be compared to whole dataset then all views from all subsets are ranked and top-k views are selected as the recommended views. Hence, DiVE can be categorized as data-driven schemes and does not consider user preferences.

In order to achieve our goal, ALiVE is proposed as the conversational data exploration model which supports iterative data exploration and pay attention to user preferences. Beside consider importance score and diversity such as in DiVE, ALiVE uses active learning paradigm to present views in each iteration. Figure 2 shows an illustrative of active learning approach, but it is oversimplified due to only for demonstration purpose. For instance, given D_B (e.g., Flight dataset), the subsets are described in the subset A, subset B, subset C, subset D, as shown in Figure 2. The number of views in each subset should be same however because of the pruning, some views may already pruned. Let assume that the number of views which has high importance scores in each subset as shown in the Figure 2.

Figure 2 (left) shows two types of view, white colour views are unlabelled views whereas the brown view is the labelled view. The question is, in the next iteration, which view that need to be presented to the user in order to be labelled by her. If we rely on diversity, the candidate view can be in subset B, C, and D which the farthest from the labelled view in the subset A. The diversity function is blind, it only considers the most distant point from the current point. However, while using active learning approach, the first priority view that need to be selected is from subset D (green colour). The subset A, B, or C will not be the first priority because the subset A already has labelled view, whereas subset B there are only a few views, furthermore the subset C which only has two views. Meanwhile, to know whether user interested in to the subset D is the most *"important"* because this subset has many views.

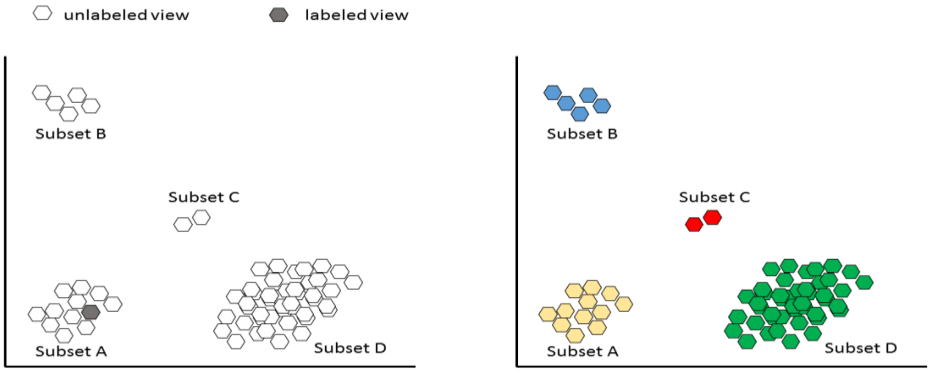


Figure 2. Active Learning illustrative example

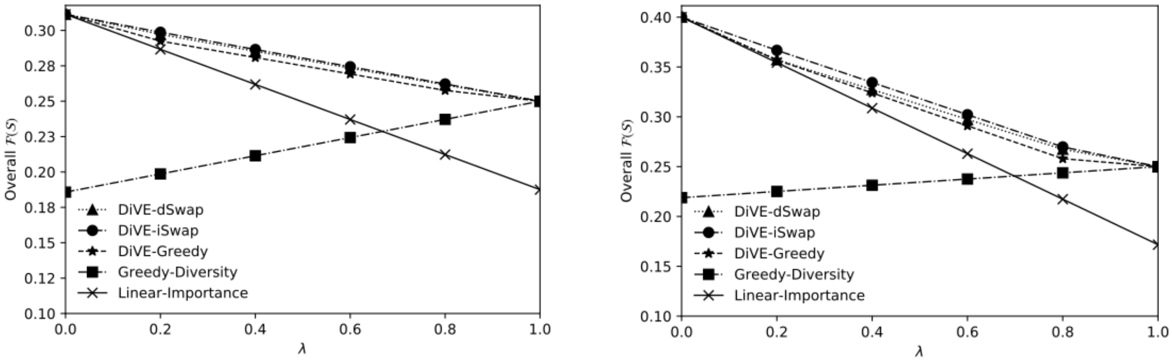
4. Experiment and Evaluation

4.1. Current Results and Analysis

Parameter	Range (default)
datasets	Heart disease, Flights
diversity weight ratio	3 (Attribute) : 2(Measure) : 1(Agg. Func)
tradeoff weight λ	0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0
result set (size of k)	5, 15, 25, 35
Prediction interval %	80, 85, 90, 95, 97, 98

Table 1: Parameters testbed in the experiments

We have conducted our experiments over the following datasets: **1) Heart Disease Dataset**[1]: This dataset is comprised of 9 dimensional attributes and 5 measure attributes, resulting in a total of 180 possible views. **2) Airline (Flights) Dataset**[2]: This dataset is comprised of 7 dimensional attributes and 4 measure attributes for a total of 112 views. While its dimensionality is lower than the heart disease data, it is a relatively large dataset of almost one million tuples, which helps in evaluating the incurred query processing time. The details of the parameters that used in our experiment testbed is shown in Table 1. For each experiment, the performance measures are averaged over a query workload of ten random queries submitted to select ten different subsets of data. We evaluate the performance of nine schemes as mentioned in the Methodology section.



Heart disease dataset *Flights dataset*

Figure 3. Impact of λ on F(S), k = 5

Figure 3 shows how the performance of each scheme in terms of overall objective function $F\ (S)$ is effected as the value of λ varies from 0 to 1. It is clearly seen in Figure 3, that for the lower values of λ the highest objective function value is achieved by Linear-Importance method. To the contrary, the Greedy-Diversity method achieves highest values of $F\ (S)$ as the λ approaches 1.

[1] http://archive.ics.uci.edu/ml/datasets/heart+Disease

[2] http://stat-computing.org/dataexpo/2009/the-data.html

Hence, there is a crossover between Linear-Importance and Greedy-Diversity. However, our proposed schemes based on the hybrid objective function have stable performance for all values of λ and are able to achieve F (S) values higher than those achieved by Linear-Importance and Greedy-Diversity.

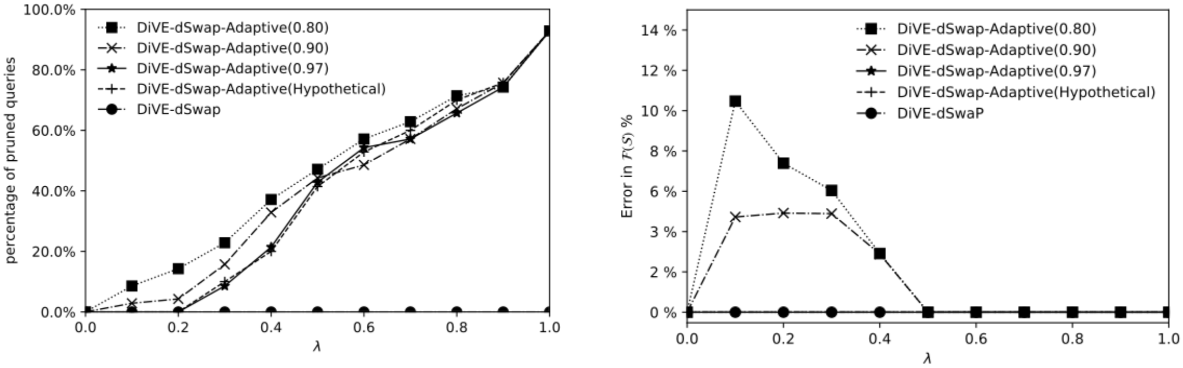


Figure 4. Pruning performance of DiVE-dSwap-Adaptive

Figure 4 (left) shows the performance of DiVE-dSwap-Adaptive with different values of PI. The interesting observation is the fact that DiVE-dSwap-Adative is able to prune 15% queries for λ values as low as 0.2. For higher values of λ the percentage of pruned queries is between 60% and 90%. The highest number of queries are pruned for PI = 0.8. Further we evaluate the effectiveness of methods with adaptive pruning in terms of the F (S) values. Figure 4 (right) shows the loss in F (S) in comparison to the F (S) values achieved by Hypothetical methods. The loss for DiVE-dSwap-Adaptive is 0% for PI = 0.97, with a larger sample size the accuracy of approximated importance score is higher. For a smaller sample size of PI = 0.80 there is 0% loss while λ = 0 because at the moment there are no pruned queries. However, there is a maximum loss of 10% at λ = 0.1. The loss in F (S) values decrease as λ increases as the impact of importance score becomes smaller in the hybrid objective function. Meanwhile, starting $\lambda \geq 0.5$ the loss is 0%.

4.2. Future Experiments

The future experiments will follow two major directions. First, we will do comprehensive experimental studies for the extended version of DiVE. The main improvement for DiVE as explained in the previous section which is experiment using another diversification and improve the adaptive pruning scheme to be more efficient. Secondly, we will extend DiVE to support multi-iteration data exploration to discover user preferences. This scheme is called ALiVE (Active Learning View Recommendation for Visual Data Exploration) due to it utilizes active learning paradigm to discover the user preferences. We will design ALiVE to be an efficient scheme that support an interactive speed, multi-iteration, multi-users, and high accuracy in recommending views.

5. Timeline

During the past one year, one paper has been published. In the next about three years, further research will be conducted and several publications will be submitted on the relevant topics as follows:

Year	Year 1				Year 2				Year 3				Year 4			
Quarter	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
PhD Milestones																
Confirmation					🟩											
Mid-Candidature Review									🟩							
Thesis Review													🟩			
Thesis Submission																🟩
Publications																
Conferences			🟧							🟧						
Journals						🟧								🟧		

Tasks	Date
Diversifying View Recommendation for Visual Data Exploration • Read and summarize literature reviews related to view recommendation especially data-driven approach. • Find the gap from the previous proposed approaches. • Propose an idea that can improve the quality of the recommended views. • Propose diversification to avoid redundancy while recommending views. • Propose new objective function for recommending views which are based on importance and diversity and conducting the experiments. • Propose a technique to reduce the cost while generating recommended views. • Write and submit a paper to CIKM 2018 (May 23, 2018)	Oct. 2017 – Sept. 2018
Confirmation Milestone	**Oct. 2018**
• Prepare the extended version of this work for the Journal submission. • Write and submit a paper to IEEE Journal. (February 2019) *Active Learning View Recommendation for Visual Data Exploration* • Read and summarize more literature reviews related to user-driven approach on view recommendation. • Conducting the experiments using active learning approach to build an iterative exploration model. • Write and submit a paper to ICDE 2020 (October 2019)	Oct. 2018 – Sept. 2019
Mid-candidature Review Milestone	**Oct. 2019**
Optimistic Visualization Tool (Extend previous works) • Focus on improving effiency of our iterative exploration model.	Oct. 2019 – Sept. 2020
Thesis Review Milestone	**Oct. 2020**
• Write and submit a paper to IEEE Journal. (January 2021). • Thesis write up.	Oct. 2020 – Sept. 2021
Thesis submission	**Aug. 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] D. Gotz and Z. Wen, “Behavior-driven visualization recommendation,” in *Proceedingsc of the 13th international conference on Intelligent user interfaces - IUI ’09*, 2008, p. 315.

[10] 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.

[11] Z. Liu and J. Heer, “The Effects of Interactive Latency on Exploratory Visual Analysis,” *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2122–2131, Dec. 2014.

[12] C. Stolte, D. Tang, and P. Hanrahan, “Polaris: a system for query, analysis, and visualization of\nmultidimensional relational databases,” *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 1, pp. 1–14, 2002.

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

[14] Tableau, “Show Me.” [Online]. Available: http://onlinehelp.tableau.com/current/pro/desktop/en-us/buildauto_showme.html. [Accessed: 19-Dec-2017].

[15] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, “Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations,” *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 649–658, 2016.

[16] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer, “Vega-Lite: A Grammar of Interactive Graphics,” *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 341–350, 2017.

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

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

[19] 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.

[20] 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.

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

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

[23] H. A. Khan, M. A. Sharaf, and A. Albarrak, “DivIDE: Efficient Diversification for Interactive Data Exploration,” *Proc. 26th Int. Conf. Sci. Stat. Database Manag.*, p. 15:1--15:12, 2014.

[24] 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.

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

[26] N. Rubens, D. Kaplan, and M. Sugiyama, “Active Learning in Recommender Systems,” in *Recommender Systems Handbook*, Boston, MA: Springer US, 2011, pp. 735–767.

[27] M. Elahi, F. Ricci, and N. Rubens, “Active Learning in Collaborative Filtering Recommender Systems,” Springer, Cham, 2014, pp. 113–124.

[28] Y. Hu, S. Sundara, and J. Srinivasan, “Estimating aggregates in time-constrained approximate queries in Oracle,” in *Proceedings of the 12th International Conference on Extending Database Technology Advances in Database Technology - EDBT ’09*, 2009, p. 1104.

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

Appendices 1: Submitted Publication

During almost one year of PhD study, one paper has been submitted.

Title: *DiVE: Diversifying View Recommendation for Visual Data Exploration*
Submit: Conference on Information and Knowledge Management (CIKM) 2018
Acceptance Notification: August 6th 2018