The Pennsylvania State University

The Graduate School

College of Information Sciences and Technology

SUPPORTING THE UNDERSTANDING OF ASSOCIATION RULE WITH VISUALIZATION

A Thesis in

Information Sciences and Technology

by

Hanqing Zhao

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

May 2017

The thesis of Hanqing Zhao was reviewed and approved* by the following:

Xiaolong (Luke) Zhang
Associate Professor in College of Information Sciences and Technology
Thesis Advisor

Guoray Cai
Associate Professor in College of Information Sciences and Technology

Frederico Fonseca
Associate Professor in College of Information Sciences and Technology

Andrea Tapia
Director of Graduate Program and Associate Professor in College of Information Sciences and Technology

*Signatures are on file in the Graduate School

ABSTRACT

Integrating advanced data-processing algorithms into visual analytic systems allows analysts to gain more insight into data and to make better decisions. However, such integration also poses new challenges, one of which is the understanding of the algorithms and their results. Sufficient knowledge about what an algorithm is about and how it works could improve the confidence of analysts on analytical results and the quality of final decisions. This thesis reports a study on the use of interactive visualization to support the understanding of the results of association rules, a popular data-mining algorithm. We developed a web-based visual analytics tool that supports the exploration of relations among items, item sets, and rules. We also conducted a qualitative study to examine the factors that affect a user's understanding of data and even the algorithm. We conducted an observational study to let users finish tasks and a semi-structured post-study interview. The findings report that most of the participants found it helpful for them to understand the algorithm better than before, but there are still barriers for. They also shared their thought comparing the visual analytic tool with pure text views. Based on the findings, we make some design recommendations for visual analytical systems targeting for similar issues. For those visual analytic system designers, we suggest that they design system supporting both overview analysis and detailed analysis, and supporting both beginners and deep users with light version and extra add on functions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I sincerely thank my advisor Dr. Xiaolong (Luke) Zhang for his guidance and support throughout the time of my study at Penn State. He always guided me on the research and gave helpful advice for my career. I was a great pleasure to work with him.

I would like to thank Dr. Cai and Dr. Fonseca as my committee members for advising my research on this thesis.

Finally, I want to say deep thanks to my parents, my friends, and my boyfriend. They accompany and share their love with me all the time, which encourage me and give me the energy to walk out of the darkness and embrace the life.

Chapter 1

Introduction

Over the recent years, the surge of data mining algorithms empowers the area of computation. Such many algorithms help data analysts better analyze the huge amount of data. The emergence of visual analytic systems also gives data scientists an environment to visually manipulate and analyze data. The integration of both two techniques gives analysts more insights in their decision-making process (Keim, 2002). Visual analytic systems help simplifies the data exploration process, where analysts can finish their tasks based on the visible interface. However, what data analysts are chasing for is not only the exploration result, but also the understanding of exploration process, especially the understanding of algorithms and data relations. They keep asking themselves how this algorithm come out with a good performance, and why these algorithms make more sense comparing to other algorithms. This phenomenon raised a problem in visual analytics research areas, which is, most visual analytics systems answer the 'what' question in novel ways, but seldom show the 'how' and 'why' questions as a detailed process. Good visual analytics system should be able to support users control of their state in exploration, and help users better control and understand algorithms (Shneiderman, 2002). A typical data mining process includes such steps; data cleaning, data selection and transformation, data mining, evaluation and presentation (Han, Pei, & Kamber, 2011). Existing visual data mining techniques are focusing on single aspects of preliminary data analysis or model evaluation, in which analysts could gain an overview on preliminary data description or mining result. What we want to achieve is to present a wider process which also involves model construction process that helps users better understand on how algorithms and the analytical process work.

The integration of both data mining algorithms and the development of visual analysis systems do help analysts generate more insights during their decision-making process. However, such integration also reveals another important potential challenges, one of which is the understanding of the algorithms and their results when using visual analytical tools. Users may fail to understand the algorithm itself when using the visual analytical tools, and on the other hand, users who deeply understand algorithms want to improve their working efficiency. Fully understand of knowledge about what the background algorithm is about and how it works will improve the confidence of analysts on analytical results and the quality of their final decisions. Moreover, besides presenting data spreadsheet and data mining result in pure text, the data exploration process is of the same importance. To help analysts better understand data relationships more efficiently and accurate, we design visual analytics tools that users can visually analyze the data through interaction. By conducting this study, we also aim to evaluate how the usability of our visual analytics tool performs.

Research Objectives

Based on the challenges stated above, we propose a research question as below: How to use interactive visualization to support the understanding of the data mining results?

To detailed describe the research question, the sub objectives of this work will solve these sub-questions.

- How to design interface and interaction to support specific data analysis process?
- How to help user better understand data analysis result using visualization?

The purpose of this study is to design and evaluate data visual analytics tools. The specific objectives are as following:

- To design visualization tool based on existing measurement,
- To understand how well visualization tools can help users better understand the relationships among data items,
- To investigate how satisfied users are with the designed tools,
- To explore design implications that we can learn from user feedback to benefit future visual analytical tool designs.

We hope that the outcomes of this research can provide a system that users can understand the relationships faster and more accurate, and users find satisfied. We also hope that our research can offer some design recommendations that inform the design of systems with similar purposes.

Research Methods

Based on the research objectives above, in our work, we conducted qualitative method to explore user's behaviors using the designed visual analytical tools. Our research is composed of two main parts:

- System design of one popular data mining algorithm;
- Usability Study aiming to evaluate how the usability of our visual analytics tool performs.

For the visual analytical tool design part, we will elaborately describe the design rationale and design implication in Chapter three.

For user study part, we designed to conduct an observational study and a post-study interview. Chapter four will present more details of our evaluation methods.

Structure Overview

This thesis is organized into five sections. We first discuss the research question and the objectives we are aiming to. In the second chapter, we collected related literature and categorized them as different parts. The third chapter is about the research design we conducted, where we elaborate the research method, system design, and data analysis methods. The fourth chapter presents the results of our study. At last, we discuss our work's contribution and implications with future work.

Chapter 2

Literature Review

In this section, we discuss the related literature from several aspects. We organize the structure by splitting our research question into three parts. The main focus of this review is the integration of data mining and visualization, association rule mining algorithm visualization, interaction in visual analytic system design to support understanding in data mining process. Besides these three main areas, we also identify different methodologies in the design of these type of research.

Integration of Data Mining and Visualization

Data mining can be identified as knowledge discovery in data sets. Analysts want to discover or extract as much useful information using automated algorithms. Data mining problems can generally be classified as supervised learning, which is predictive mining, and unsupervised learning, which is descriptive mining method (Yoo et al., 2012). Descriptive data mining infers exploratory mining, like association, clustering, and summarization, on the other hand, prediction data mining includes classification, regression, and outlier detection, etc. Typical data mining methods are related to these six types of tasks: Anomaly detection, association rule learning, clustering, classification, regression, and summarization (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

A traditional data mining workflow consists of three steps: preliminary analysis, which is the first step analysis on data; model construction, in which step analysts use different models on the data to conduct result; and the last step, model evaluation, where the model constructed before

measured by standard measurement methods and parameters (Li, 2015). Traditional data visualization usually helps in single perspective of this process, such as an overview of preliminary data analysis, or model evaluation to visualize discovered result. Such visualization may cause the effect of users' lack of understanding of data mining algorithm and data relationship. Here our work aims to expand the visual analytic tool not only on such single perspective such as pattern discovery but also on a whole process of knowledge discovery.

Visual design for Association Rule Mining Algorithm

Association rule mining algorithm aims at exploring relations between variables in high-dimensional data sets. A famous example, {beer} --> {diaper} in Walmart supermarket is concluded by using association rule method to find patterns in shopping baskets. This algorithm discovers interesting correlation relationships among market transactions, which was always used in market basket analysis to help marketing experts make decisions on designing catalog, market layout, analyzing customers behavior (Agrawal, Imielinski, & Swami, 1993; Han et al., 2011).

Association Rule Mining, as a popular data mining algorithm, attract attention from different aspects, especially in visualization area. Network visualization is widely used in association rule visualization. One famous visualization of association rules is ArulesViz (Michael Hahsler, 2017), an R package developed to visualize frequent item set and rules, which use circle size and color to encode two attributes, support and lift. Statsoft visualizes the rule using parallel coordinates alike network, which also use circle's size and color to encode confidence and item set. (Statsoft, 1995) Many other researchers also use parallel coordinates to visualize frequent itemsets, association rules and sequential patterns. (Bruzzese, Davino, & Vistocco, 2003; Kopanakis & Theodoulidis, n.d.; Yang, 2003) When focusing on high-

dimensional association rule visualization, Two-key Plot (Unwin, Hofmann, & Bernt, 2001) can show the two keys of confidence and support for large data set. Besides visualizing this algorithm as an overview, single rule visualization also emerges, such as visualized as double decker to help user understand data relationship (Hofmann, Siebes, & Wilhelm, n.d.). Lei et al, use rule matrix to interactively visualize single rule with adjusted parameters. (Lei et al., 2016)

Association rule mining visualization can also be traced from the categorical data visual analysis, which involves different types of analysis. In Friendly's book, he first detailed using SAS and hand on an experiment to visually present categorical data analysis (Friendly, 2012). Besides, in most research papers, researchers use parallel coordinates and its transformation to visualize categorical data. Such a fast ordering categorical data analysis algorithm helped visualization have a better layout (Beygelzimer, Perng, & Ma, n.d.; Ma & Hellerstein, 1999), where their algorithms help organize the original parallel coordinate clearer. Hammock plot, a modification for parallel coordinate, was invented by Schonaue to visualize categorical data(Schonlau, 2003), and his design replace coordinate polygons by rectangles to present the number. Another classic visualization, Treemap, is also modified to support categorical data visualization. CatTree gives a hierarchical categorical data visualization with interaction(Kolatch & Weinstein, 2001). In some other methods, simple statistical charts are modified. Vivacqua et al. use nested rings, an interactive visualization to present categorical data, which helps users interactively order data dimensions(Vivacqua, Cristina, & Garcia, n.d.). Shirashi and his colleges design a tool, Granular Representation, to visually explore data using cluster presentation with bar charts (Shiraishi, Misue, & Tanaka, 2009). Researchers also use algorithms to help organize this type of non-linear data, like R-Map, which maps categorical data into numerical data using their algorithms that help further clustering(Shen, Sun, Shen, & Li, n.d.) Recent researches integrated both algorithms and multi-view visualizations in interactive visual analytic systems. Fernstad developed an interactive system combining parallel coordinate, table lens and scatterplot

matrix together for an overview explorative analysis. She has a full research on categorical data visualization to support algorithm understanding and (Fernstad, 2011) Another novel contingency wheel presented by Alsallakh support visual analytics in categorical data, and he measures association based on Pearson's residuals, and use visual abstraction based on elements' frequency. His tool supports both overview and detailed item description using coordinated views (Alsallakh, Aigner, Miksch, & Groller, 2012).

Supporting Understanding of Data and Algorithm

To open the black box of computation algorithms, interactive visualization tools emerges in recent years (Muhlbacher et al., 2014). Client-driven implementations and algorithm-driven implementations are discussed with design considerations. In the section above, most visualizations are still in the black box, where the algorithms are not fully explored by users. Some of them only focus on preliminary data analysis part, like parallel coordinate and similar visual presentations(Inselberg & Dimsdale, 1991). As Li (Li, 2015) stated in his paper, model construction includes three design rationales, progressive construction, iterative prototyping, and interactive pipeline construction (Li, 2015). DimStiller gives an overview of workflow in dimensional analysis and reduction with detailed steps, in which their tool help target users understand inner workings of given parameter with visual feedback (Bremm et al., 2011). Such systems serve to give feedback for data models, and its retraining and reevaluation functions help data scientists better choose a fitting model (Seifert and Lex, 2009). Rene and Schumann developed progressive parallel coordinate and progressive tree maps to enhance visual adaption and reduce visual clusters (Rosenbaum & Schumann, n.d.). Yan et al. use visualizations like scatter plots and treemap with adjustable parameters in the decision tree model construction process (Yan et al., 2012), with such interaction, users' knowledge discovery process become

more reasonable and sensemaking. Some system only focus on model evaluation. Barlowe et al. used an automated numerical method to visualize not only dimension information but also shows outlier information, correlation information, and also significant of each features (Barlowe, Zhang, Liu, & Yang, n.d.), which give analyst enough efficient information in data mining process, but lack of further modification. Our work will focus on building a visual analytic system from a wider perspective on the data mining process.

Research Design

In traditional visualization area, research methodologies below are typically used. Common methodologies for visualization area includes: evaluating the applicability of theory to data visualization problem, comparing solution A to solution B, rapid prototyping and spiral design, proof of concept prototypes, graphics algorithms, and also interactive methods. ("Interactive Visualization: Methodologies," n.d.) This article also summarizes several evaluation techniques from different perspectives. To simplify, evaluating the applicability of data visualization problems can be categorized in two ways: quantitative methods such as giving numerical results comparing different solutions; qualitative methods using narrative or descriptive data, such as the description of the views and attitudes in a specific scenario.

Sedlmair et al. proposed a methodological framework consisting of 9 stages: learn, winnow, cast, discover, design, implement, deploy, reflect, and write. (Sedlmair, Meyer, & Munzner, 2012) Their model shows a nine-stage framework for visualization research. These stages can also be seen as a process: Find out a real-world problem; Design visualization to solve the problem; Design validation; and Reflection for design guidelines.

In our work, we will follow this traditional workflow they proposed. To be specific, we develop a proof of concept software web-based system to support detailed analysis for the chosen algorithm. To evaluate the designed software tool, a qualitative study will be conducted to collect narrative data in a specific scenario.

Chapter 3

Research Design

In this chapter, we describe the research methods used in our study. The first section is an overview of the whole study, and the second section describes system design. The detailed data collection and analysis part is elaborated in chapter 4.

Study Design

In research design part, we decide to develop a web-based prototype system to support detailed analysis of association rule mining data. To evaluate the designed software tool, a qualitative study will be conducted.

Visual analytical tool design and development

To answer the research question we come up with, a system is required to be built. We want to figure out whether the system we designed satisfies the requirement we proposed, and how its functionality could fulfill the requirements. We designed this system based on researchers understanding on association rule mining algorithm. The details of this system will be elaborated in the Design implication section.

Observational study and post-study interview

To evaluate the system we designed, we choose the qualitative strategy of doing a set of tasks and a semi-structured interview to generate insights for further design consideration.

To solve the research question we come up with, we want to figure out whether the users better understand the data and algorithms they are using. Quantitative methods can vaguely answer this question. Instead, we come up the evaluation workflow consisting such steps. First, a pilot study will be conducted, in which field experts will be recruited to ensure the tasks go smoothly. Then we will go through an observational study including pre-task questionnaires, task solving, and a post-task interview. When the task is going on, users are asked to thinking aloud. The time used in task solving will be recorded, and the users' activities during this whole process will be recorded in transcripts for further analysis.

System Design

The Proof of Concept system is designed within such steps: researchers first using a simple way to explain this algorithm, designing the system that visually encodes the concepts and workflow, using techniques to develop a system that implicates all design consideration.

Algorithm Explanation

In this sub-section. We elaborate the algorithm we chose among different algorithms, which is association rule mining algorithm. The reason we choose this algorithm is that association rule mining algorithm is a widely used algorithm to find association, correlations, and many other interesting relationships among data. As a fundamental algorithm in data mining area, it also contributes in data classification, clustering, and many other data mining tasks(Han et al., 2011). Comparing to the other data mining algorithms, it is a relatively easier algorithm to understand, in which prior knowledge is not strongly required. This algorithm contains two steps generally, which are, first, frequent itemsets generation; and second, association rule generation

from frequent itemsets (Han et al., 2011). Here we use market basket analysis example to explain key terms of this algorithms are as below:

Item: a product customer bought from the market.

Frequency: the frequency of buying certain product among all the market transactions.

Item set: a list of products customer buy together

Support: the probability of buying certain products among all the market transactions.

Rule: a correlation relationship. Customer buys certain items leads to the result that he will also buy certain items.

Confidence: The probability of buying certain products when some product has been bought.

In the data mining domain, association rule mining has more advanced concepts besides the concepts above. In this thesis, we only focus on the basic level of association rule mining.

Design Rationale

During the development process, three main characters are considered, which are interaction, traceable, and comparable. For interaction function, users will be able to interactively choose specific data, a certain model, and adjust parameters; for traceable function, the system can store the steps during exploration process; for comparable, it should be able to show comparison between different choices. We will develop the whole system considering these three functions primarily.

To visualize different aspects of association rule mining algorithm, we want to emphasize the important terms in the algorithm. As as we mentioned in the last section, key terms are as follows,

Item and Frequency

Item is a product customer bought from the market, and the frequency is the frequency of buying certain product among all the market transactions. In visual representation, we present each item is represented as a circle, and the circle's radius represents the item's frequency. When user wants to elaborately look through the item, he can click the item itself to detailed explore the relations related to such item.

Figure **3-1**: Part of prototype interface of designed visual analytic tool.

Itemset and Support

An Itemset is a list of products customer buy together, and the support is the probability of buying certain products among all the market transactions. In our tool, each item-set is represented as an arc. The arc connects several vertexes means that such items are in the same item-set. If there are multiple vertexes, a straight line from the middle of the arc will connect the addition vertexes. When user wants to elaborately look through the itemset, he can click the itemset itself to detailed explore and compare at side panel.

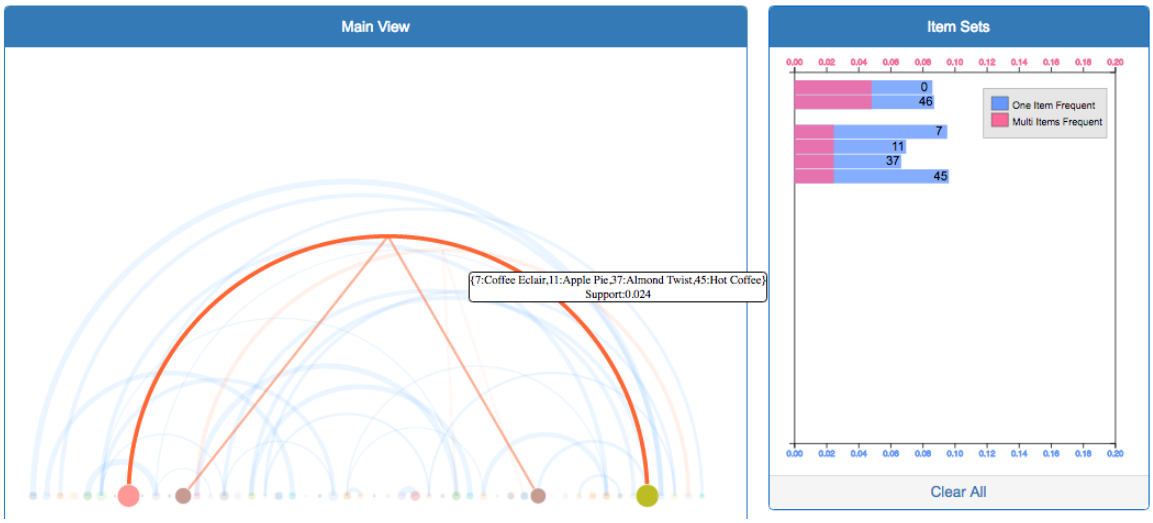

Figure **3-2**: Part of prototype interface of designed visual analytic tool.

Rule and Confidence

A Rule is a correlated relationship such as customer buys certain items leads to the result that he will also buy certain items. The probability of buying certain products when some product has been bought. It is calculated in this way: Support of the itemset divided by the Frequency of cause item. In our tool, each rule is represented as a bar below the horizontal line. The bar represents a causal relation, which means people buy several items predicts that they will probably buy the certain item. The probability is presented as the height of the bar. Lines connecting the bar are the cause, and the bar itself is the effect. When user wants to elaborately look through the rule set, he can click the rule set itself to detailed explore and compare at side panel.

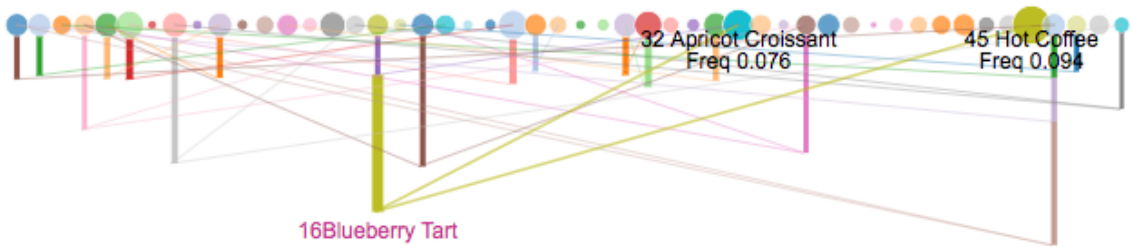

Figure **3-3**: Part of prototype interface of designed visual analytic tool.

Design Implication

Based on our previous analysis, we developed a visual analytical tool to support the understanding of association rule mining. Our system requires users to find and sort items and identify the products frequency as an overview. Users are given the ability to find interesting item set as an overview and explore itemset itself in details. Users are also able to find interesting rules through an overview and explore single rule in details. Two artifacts were designed to support such sense making process: text-based workspace below and visualization history.

The system we build is a web-based software tool, which could be run and test under a different situation. The environment we use is the technical stack with tomcat server on Linux environment. The language we choose are JavaScript, HTML and CSS, as most of the web development. We choose JavaScript based on the powerful toolkits that it has, which supports various visualization toolkits, like d3.js, echart.js, highchart.js. JavaScript is friendly to use is also because of its portability and maintenance.

As as we mentioned in the last section, key terms are visualized as follows. The figure below shows an overview of the design visual analytic tool.

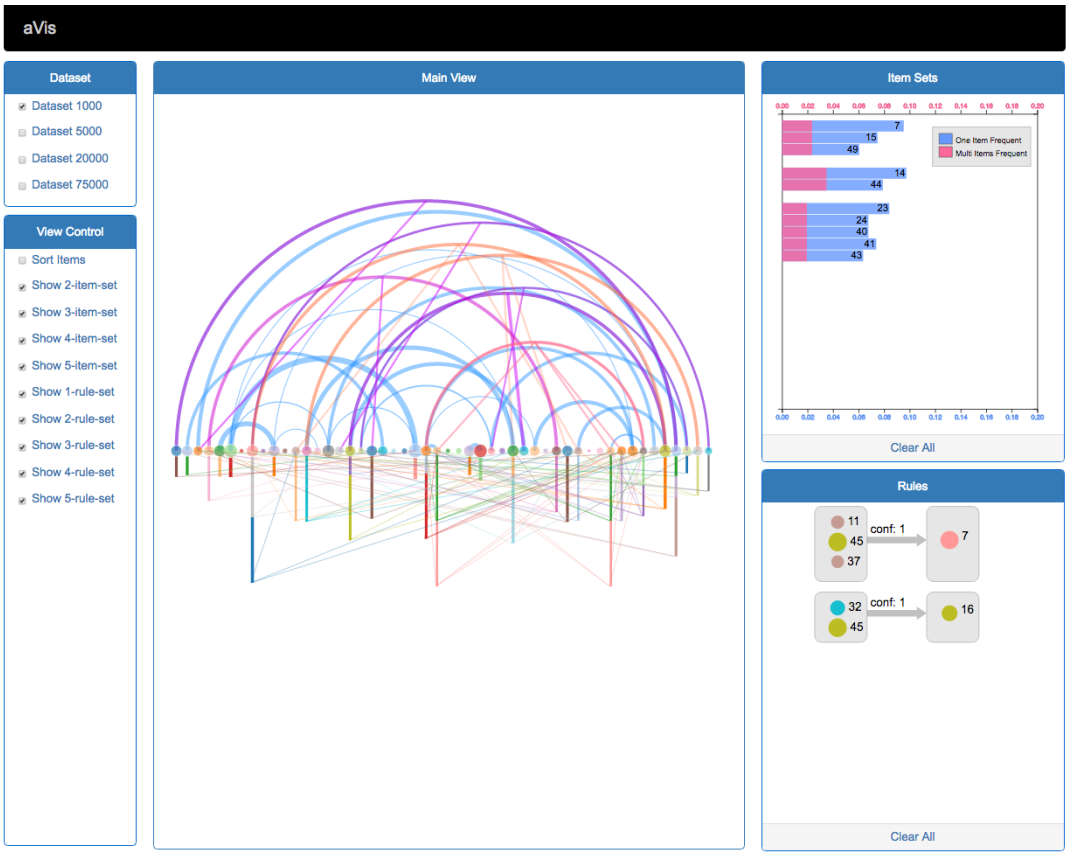

Figure **3-4**: Prototype interface of designed visual analytic tool.

Chapter 4

Evaluation

After we finish the designed visual analytic tool development, we conducted a qualitative study to evaluate how this tool is used by users in understanding data and algorithm.

Study Design

Our goal is to understand how well the designed tool can help users in performing sets of tasks related to the understanding of the results of the association rules algorithm. This study is an observational study, with some control over the tasks and tools users were given. We also had a post-study interview to gather more data.

We had one controlled factor when recruiting subjects, the prior knowledge of the algorithm. We controlled this factor because the prior knowledge can affect the understanding of the results. We are interested in how our tool can help people.

We designed a set of tasks to participants to perform. These tasks examined the different perspectives of user understanding of algorithm results, from individual items to final rules.

We presented the results to participants in two ways: a traditional, pure text-based format as the baseline, and our visual analytical approach. We asked all participants to see results in both ways, but we controlled the sequence of them: half saw the baseline first and half saw visualization first.

We had a semi-structured interview with each participant after the tasks to gather their feedback on the tasks and design.

Subjects

We recruited 8 participants for this study. They were all in the area of information sciences and technology, but with different education backgrounds. Some of them have the prior knowledge of data-mining and related algorithms and while some do not. Their knowledge level was rated from 1 (low) to 5 (high). Table 4-1 listed all participants and their backgrounds.

Table 4-1: Participants' information.

ID	Prior Knowledge Level	Education background
P1	2	Master of Science
P2	4	PhD
P3	4	Master of Science
P4	4	PhD
P5	3	Post doc
P6	2	Master of Art
P7	2	Undergraduate
P8	3	Undergraduate

Tasks

We designed following tasks.

1. to find out the what item people most likely to buy;
2. to find out an item-set and explain to the researcher the details, such as how many items are in this item-set, what these items are, and what its support is;

3. to compare two different itemsets and find out the difference between them;
4. to pick top three sets of goods that people may buy together;
5. to find out a rule/causation and explain to the researcher the details, such as which items are the cause, and which item is the effect;
6. to compare two different rule/causal relations and what are the difference between them; and
7. to pick up three best predictions.

The order of tasks was pre-determined, and cannot be counter-balanced. All these tasks were carried out in the order indicated above. This is because the latter tasks usually require the knowledge and result from the former ones. The number of tasks we list here is 7, but in the real study, we divided them into 3 groups. Q1 is task 1, Q2,3,4 are merged as task 2 since they are all about item sets, and Q5,6,7 are merged as task 3 since they are all about rules/causal relation. More details can be found in the Appendix.

Procedure

Participants were first to present the consent form and signed the form. Then they were provided information about this study and the introduction to the association rules and the visual analytic system. They had time to ask questions and practice.

As the tasks started, they were asked to complete all tasks on a laptop. During this process, they were asked to think aloud. We recorded their words and logged their activity on the computer.

Participants were assigned different orders of tools they used. Table 4-2 shows the order for each participant.

Table 4-2: The order of doing tasks by visual analytic tool and text view.

ID	Order
P1	Visual Analytic Tools First; Text View Second
P2	Visual Analytic Tools First; Text View Second
P3	Text View First; Visual Analytic Tools Second
P4	Visual Analytic Tools First; Text View Second
P5	Text View First; Visual Analytic Tools Second
P6	Visual Analytic Tools First; Text View Second
P7	Text View First; Visual Analytic Tools Second
P8	Visual Analytic Tools First; Text View Second

After completing all tasks, we conducted interviews and asked them to complete the questionnaire. The interview questions were based on previous research that studied visual analytics, and based on the performance of subject during the study. The interview questions were largely about the following issues:

- a participant's prior knowledge level;
- what aspects the participant thought the designed visual analytical tool might help to understand the relationship of data;
- what aspects the participant thought the designed visual analytical tool might make it difficult to understand the relationship of data;
- the participant's preference of text view or visualization view of association rule mining;
- the advantages and disadvantages of the designed visual analytical tool; and
- other comments

The interview questions can be found in the appendix.

Experiment Result

This section reports the experiment result when participants doing different tasks. Here we list all their activities below. Based on their activity logs, we can better analyze the result to gain more insights.

P1

When doing tasks by visual analytical tool: For task 1: P1 finished task 1 easily and correctly, which means he can find out the most bought items, and he used the function of sorting; P1 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, and compare two item sets to researcher. But P1 found it a little difficult when asked to find the top three most bought item sets and got wrong. P1 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to the researcher, and picked up top three predictions.

When doing tasks by looking through text view: P1 finished task 1 easily but got wrong, which means he can find out the most bought item but made a mistake. P1 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, compare two item sets to researcher, and find the top three most bought item sets, but P1 also got wrong on finding the top three most bought item sets. P1 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

P2

When doing tasks by visual analytical tool: For task 1: P2 finished task 1 easily and correctly, which means he can find out the most bought items, and he used the function of sorting; P2 also

finished task 2 easily when asked the questions of finding one item set, explain it to researchers, and compare two item sets to researcher. But P2 found it a little difficult when asked to find the top three most bought item sets and got wrong. P2 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

When doing tasks by looking through text view: P2 finished task 1 easily but got wrong, which means he can find out the most bought item but made a mistake. P2 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, compare two item sets to researcher, and find the top three most bought item sets, but P2 also got wrong on finding the top three most bought item sets. P2 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

P3

When doing tasks by visual analytical tool: For task 1: P3finished task 1 easily and correctly, which means he can find out the most bought items, and he used the function of sorting; P3also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, and compare two item sets to researcher. But P3 found it a little difficult when asked to find the top three most bought item sets and got wrong. P3 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

When doing tasks by looking through text view: P3 finished task 1 easily and correctly. P3 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, compare two item sets to researcher, and find the top three most bought item sets easily and correctly. P3 finished task 3 easily but with his own thought, which means he can

easily find out a rule, explain it to researchers, compare two rules to researcher, but picked up top three predictions on the items with high support and high confidence.

P4:

When doing tasks by visual analytical tool: For task 1: P4 finished task 1 easily and correctly, which means he can find out the most bought items, and he used the function of sorting; P4 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, and compare two item sets to researcher. P4 used filter function to filter out only two-items item sets and he is the only one answer this question correctly. P4 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

When doing tasks by looking through text view: P4 finished task 1 easily but got wrong, which means he can find out the most bought item but made a mistake. P4 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, compare two item sets to researcher, and find the top three most bought item sets, but P4 also got wrong on finding the top three most bought item sets. P4 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

P5

When doing tasks by visual analytical tool: For task 1: P5 finished task 1 easily and correctly, which means he can find out the most bought items, and he used the function of sorting; P5 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, and compare two item sets to researcher. But P5 found it a little difficult when asked to find the top three most bought item sets and got wrong. P5 finished task 3 easily correctly, which means

he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

When doing tasks by looking through text view: P5 finished task 1 easily but got wrong, which means he can find out the most bought item but made a mistake. P5 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, compare two item sets to researcher, but when asked to find the top three most bought item sets, P5 got wrong. P5 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

P6

When doing tasks by visual analytical tool: For task 1: P6 finished task 1 easily and correctly, which means he can find out the most bought items, and he used the function of sorting; P6 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, and compare two item sets to researcher. But P6 found it a little difficult when asked to find the top three most bought item sets and got wrong. P6 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

When doing tasks by looking through text view: P6 finished task 1 easily but got wrong, which means he can find out the most bought item but made a mistake. P6 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, compare two item sets to researcher, and find the top three most bought item sets easily and correctly. P6 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

P7

When doing tasks by visual analytical tool: For task 1: P7 finished task 1 easily and correctly, which means he can find out the most bought items, and he used the function of sorting; P7 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, and compare two item sets to researcher. But P7 found it a little difficult when asked to find the top three most bought item sets and got wrong. P7 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

When doing tasks by looking through text view: P7 finished task 1 easily, which means he can find out the most bought item easily and correctly. P7 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, compare two item sets to researcher, and find the top three most bought item sets, but P7 also got wrong on finding the top three most bought item sets. P7 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

P8

When doing tasks by visual analytical tool: For task 1: P8 finished task 1 easily and correctly, which means he can find out the most bought items, and he used the function of sorting; P8 also finished task 2 easily when asked the questions of finding one item set, explain it to researchers, and compare two item sets to researcher, and P8 found it easy when asked to find the top three most bought item sets but got wrong. P8 finished task 3 easily correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

When doing tasks by looking through text view: P8 finished task 1 easily and correctly, which means he can find out the most bought item. P8 also finished task 2 easily and correctly when asked the questions of finding one item set, explain it to researchers, compare two item sets to researcher, and find the top three most bought item sets. P8 finished task 3 easily and correctly, which means he can easily find out a rule, explain it to researchers, compare two rules to researcher, and picked up top three predictions.

Table 4-3: Participants' result of doing tasks by using our designed visual analytic tool.

Pid	**Task 1**	**Task 2**	**Task 3**
P1	Finish the task easily	Difficult to find the item set with the highest support and got wrong	Finish the task easily
P2	Finish the task easily	Difficult to find the item set with the highest support and got wrong	Wrongly find the top three item sets
P3	Finish the task easily	Difficult to find the item set with the highest support and got wrong	Finish the task easily
P4	Finish the task easily	Finish the task easily	Finish the task easily
P5	Finish the task easily	Difficult to find the item set with the highest support and got wrong	Finish the task easily
P6	Finish the task easily	Difficult to find the item set with the highest support and got wrong	Finish the task easily
P7	Finish the task easily	Difficult to find the item set with the highest support and got wrong	Finish the task easily
P8	Finish the task easily	Finish the task easily	Finish the task easily

Table 4-4: Participants' result of doing tasks by looking though text view.

Pid	**Task 1**	**Task 2**	**Task 3**
P1	Finish the task easily but got wrong	Finish the task easily but got wrong	Finish the task easily
P2	Finish the task easily but got wrong	Finish the task easily but got wrong	Finish the task easily
P3	Finish the task easily	Finish the task easily	Finish the task easily but have different opinions
P4	Finish the task easily	Finish the task easily	Finish the task easily
P5	Finish the task easily but got wrong	Finish the task with difficult and got wrong	Finish the task easily
P6	Finish the task easily	Finish the task easily	Finish the task easily
P7	Finish the task easily	Finish the task easily but got wrong	Finish the task easily
P8	Finish the task easily	Finish the task easily	Finish the task easily

From the result above, we found out such patterns:

1. When doing task 1, visualization help users finish the task 1 easily. Users can easily find out the item people mostly like to buy among all the items by visualization. Users can finish task 1 easily by looking though the text view but some of them have the wrong result.
2. When doing task 2, people can easily identify item set, compare item sets and their support easily by both text view and visualization. However, when finding out the top

three mostly bought item sets, most of the users find it difficult to do this task by visualization. Some of the users easily finish the task by text view but got the wrong result.

3. When doing task 3, people can easily identify rule set, compare rule sets and their confidence easily by both text view and visualization. When finding out the top three mostly confident rule sets, most of users find it easy to do this task by both visualizations. Some of the users got the wrong result, or even their own understanding.

Interview result

After finishing all the tasks using both visual analytic tool and text view, we interviewed each participant for 10 minutes. They provided their background level based on their own though from 1 to 5 (1-low, 5-high), and they talked about the PROs and CONs of both text view and visualization with comments. We discuss their though in the section below.

Discussion

Overall from the task result and interview result, we find both PROs and CONs of the designed visual analytical tool and text view.

Visual analytical tool

PRO:

Overall, users have a good impression on the tool we designed and developed. Based on their feedback, we find that our tool helps them better understand association rule and data relations. Visualization of association rule gives them an overview of data points. Users speak highly of the tools we designed. To be specific, it has such advantages.

- Visualization is easy to understand and to use.

[P5] "I think it is easy for me to use. (I prefer) clearly this one(visualization). I think this prototype is pretty good."

[P7] "This data picture is more clear for me, more straightforward for me"

- Visualization provides users with an overview to take a look. It helps them have a big picture of all the data.

[P2] "If I want to an overview, I will probably, choose the main view(visualization), but if I want, really detailed information, I probably go to the text view."

[P2] "I will see the main view give a really good overview of item set or the rule set, so I can easily see those rules, also the item set panel, rule panel gives individual views of different items,

[P4] "I think that arc helps a lot. I think, apparently you will show, it shows some patterns, and easy to locate, some of them can dig into it, so you have a big picture of how they related, you have a quick grasp of the data you have... um yes"

- Visualization helps users better understand the terms in association rule mining.

[P2] "The visualization is definitely more efficient than just give you the numbers, like the support, what's the support, what's the confidence"

[P2] "In the design, you can use…can see the height of the bar, you can easily see the support (confidence), OK, you get more confident on those rules, and arc is like, very clear like how to find three items, four items in the item set, and also the rule panels likes, if you are interesting in one of the rules when you click on bar, the other panel is also helpful if you want to go deeper to see what's like the items in the item set. It gives you a

clear way to see more than one rule, also provide clear information of individual item set"

[P6] "Yes. I think the thickness, is really good, and this one, it makes more direct, and also, the size of the circle, especially you can sort it, it makes much more easier to find out the most bought item, and I think the bars are also good, but I just can't find out which one is longer"

[P8] "I will see the links between the items shows a clearly, an item set contains, and the size of the circle shows the frequency of the item"

- Users get insights when using visualization tool. Some user thinks about other terms in a different research area, while some user thinks of solving them problem in their own way.

[P3]"Let me see, so this means we can combine support and confidence together to make comparison. I think you shouldn't focus on one..um one measurement, one metric alone. So you have to take it together as a whole. So I think the support is big, and the confidence is also big, that makes a confidential rule."

[P4] "I think top three (bought items) are all in two-items (two-item item sets)"

[P6] "You know what? That reminds me two concepts when I was studying Economics, which are Complement good and Substitute good. Complement goods means that two goods have to be sold together."

CON:

The visualization uses different visual elements to encode different dimensions. Users have to remember all the information of projection of data to visual representation. When users

want to find out more information of association rule, the function of visualization is limited for them to have deeper analysis. To be specific, we summarize such feedbacks of disadvantages.

- Users have to remember metaphor/projection from data term to graph element

[P6] "I almost forgot, I mean the projection of terms on the graph, because when I see the circle, it takes me time to remind myself what is this represent for, and the bars, the arcs, the thickness of arcs, and these small lines."

[P7] "This one, the support I can't find the top one without the (projection) information"

- The overview is too informative that users find it difficult to adapt.

[P1] "I think too much like lines, like circles, they are just overlap, right?"

[P3] "But you know, this bar has too (much) different colors, so it's a little confusing"

[P6] "When they intersect with each other, it makes me frustrating."

- Our visual encoding way is difficult to let users compare different elements with high accuracy

[P8] "It's just sometimes it's hard to find which arc is wider, which bar is larger, so many arcs together, sometimes it stuck when clicking"

- The visual analytical tool is delicate to manipulate, users have to carefully interact

[P7] "And I think when those arc, all the staff is a little hard to find, pick the top line, they are just too delicate"

- The generalization of visual analytic tool

[P5] "I have designed some prototypes, but not the visualization, just the interaction prototype. The most important problem is, as far as I think, you may develop a software just for some specific users, but can you generalize it? For example, this one is maybe adaptive to your tasks here. But if I were a Walmart manager, I want to do more than just this, can I use it? Maybe not, not yet. Could I extend this software based on my new requirement? To customize, to generalize my concern?"

- Lack of add-on function such as querying, filtering, sorting

[P6] "I think maybe you can add more sorting function, like sorting arcs, sorting any attributes"

Text view

PRO:

Comparing the visualization, text view of data result has advantages as well. It provides detailed information when users want to go deeper to see the original data.

- Detailed information provided in the text view

[P2] "If I want to an overview, I will probably, choose the main view, but if I want, really detailed information, I probably go to the text view."

[*P2] but if I really want to explore, like go to deeper, I will probably go to the bottom one, the one with text information. In the beginning, it not easy to get what everything is, but with text view, I can quickly navigate the rules with the highest one"*

CON:

However, text view also reveals disadvantages among the user's feedback. Here are some of the user's opinions on the other side of text view.

- The number of index confuse user with the number of item

[P3] "The index is number, and items are also number, that's a little confusing, you have to explain it separately that it's entry, it's index, you can't, you know, confuse people in the first sight of this page"

[P4] "When I see rules and item sets, there were numbers on it, I thought they were number indicate some meaning, but I eventually find out they are just index"

[P4] "Maybe the better way to show the name, um.. the index is ok, maybe confuse with other number, a lot of numbers… and actually index has no meaning, maybe you can use a, b, c, d…you can even use Greek numbers, or use some logo, it's fine."

- Easy to ignore

[P6] "I just skip it (data points), I didn't even recognize them as a part of your whole data"

- Not enough information

[P2] "I think it better have an instruction for text view, or for example, conf number, not everyone know what it means"

- Difficult to locate

[P8] "Cause it (Text view) is not sorted, it's complicated to find which item is larger than the other one."

Visualization VS Text view

When asked about the preference of which one is better. Users have their feedback aiming at different situation. It seems that users tend to use visualization as an overview and use text view as detailed analysis; users tend to use visualization for smaller data set and use text view for larger data set; users prefer using both at the time when they analyze data.

[P2] "I will see the main view give a really good overview of item set or the rule set, so I can easily see those rules, also the item set panel, rule panel gives individual views of different items, but if I really want to explore, like go to deeper, I will probably go to the bottom one, the one with text information. In the beginning, it not easy to get what everything is, but with text view, I can quickly navigate the rules with the highest one"

[P4] "When you want to look at the absolute number, they try to, then this one (Text view) gives you numbers that you can quickly scan it, but this one (visualization) gives you, like a shape and edges. This one (visualization) can easily locate some relationship, but if you really want to know what is the absolute number, when you have already have a question in your mind, actually seeing this (Text view) is quicker. When you have a bunch of data, first look I want to look at this (visualization), so I can ask some questions cause I have some ideas. So after I look at this (visualization), data have some meanings for me, they are just rough datas, and then I see this (text view), um to answer my questions"

[P8] "I will prefer the other one (visualization) (when the data set is larger)"

Based on the interview result and task finish result. We list some design recommendations as follow:

1. When visualizing the data points, both overview and detailed view are important. Multi-view visualization tool helps show both overview and detailed view;
2. The visualization encoding should be considered carefully. Not only the space utilization is important, but the user's visual acceptance is important as well;
3. Filtering, sorting, querying functions can help users better analyze data;
4. Information board, or functions such as help button can help users review the functions of different uses;
5. Visualization and text view should be integrated together for better use.

Chapter 5

Conclusion

In this research study, we focus on algorithm-oriented visualization that integrating data mining techniques with interactive visualizations. The main propose of this research study is to help users better understand not only data but also algorithms during an interactive visual mining process. We use research method of proof of concept system design and qualitative strategies to evaluate the outcome. As most visual analytics systems focus on preliminary data analysis and model evaluation, the process of interactive model construction, which is the most important part of data mining process, is rarely studied. Our system is built to show an overview of the whole mining progress so that this topic is worth studying in this area. We hope that the outcomes we conduct will help data analysts not only understand the whole data analysis process, but also release their workload in the future.

Contribution and Limitation

This thesis reports the study of using visualization help users better understand data relation of association rule mining. In this study, we designed a novel visualization tool and used the qualitative method to gain some insights of designing such visual analytics tool. Our work gives design recommendations for the developers to design further tools.

Besides the contribution, our work still has some limits. 1) For system design part, we design our visual analytic system based on researcher's own understanding, lacking the user's aspect of thinking; Another consideration is that we only focus on part of the terms in Association Rule Mining algorithm, in which we lack some of the visual design element to let users better

analyze data sets. 2) For the qualitative part, we analyze interviews of all the participants from different ranges. However, our study only focus on some of the people, but the pool of participants was limited to a relative higher education institution. If we seek more participants from another environment, we may gain more findings from other aspects.

Future Work

Our future work will focus on such aspects below: 1) For system design, we would modify our system based on not only our own understanding, but also the recommendation from the interviews; 2) For the qualitative part, we would have a wider range of participants and conduct preliminary study before development in the future.

REFERENCE

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(May), 207–216. https://doi.org/10.1145/170036.170072

Alsallakh, B., Aigner, W., Miksch, S., & Groller, M. E. (2012). Reinventing the contingency wheel: Scalable visual analytics of large categorical data. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2849–2858. https://doi.org/10.1109/TVCG.2012.254

Barlowe, S., Zhang, T., Liu, Y., & Yang, J. (n.d.). Multivariate Visual Explanation for High Dimensional Datasets, 147–154.

Beygelzimer, A., Perng, C., & Ma, S. (n.d.). Fast Ordering of Large Categorical Datasets for Better Visualization, 239–244.

Bremm, S., Landesberger, T. Von, Bernard, J., Schreck, T., Von Landesberger, T., Hess, M., … Keim, D. (2011). DimStiller: Workflows for dimensional analysis and reduction. *Information Visualization, IEEE Symposium on*, *15*(6), 257–274. https://doi.org/http://doi.ieeecomputersociety.org/10.1109/INFVIS.2004.3

Bruzzese, D., Davino, C., & Vistocco, D. (2003). Parallel Coordinates for Interactive Exploration of Association Rules, 0–0. Retrieved from https://www.iris.unina.it/handle/11588/346788#.WNhTaxLyvVo

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, March 15). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. https://doi.org/10.1609/aimag.v17i3.1230

Fernstad, S. J. (2011). *Algorithmically Guided Information Visualization: Explorative Approaches for High Dimensional, Mixed and Categorical Data*. Retrieved from

http://liu.diva-portal.org/smash/record.jsf?pid=diva2:445884

Friendly, M. (2012). Visualizing Categorical Data: Data, Stories, and Pictures. *Mosaic A Journal For The Interdisciplinary Study Of Literature*, 1–9. Retrieved from papers2://publication/uuid/D6901171-8BDA-4D99-A6C5-A25D0A9672BD

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Retrieved from https://books.google.com/books?hl=en&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=Han,+J.,+Pei,+J.,+%26+Kamber,+M.+(2011).+Data+mining:+concepts+and+techniques.+Elsevier.&ots=tyMrYXlHZ3&sig=w-eUPpmooNRXZpSpxiSczfD6NPw

Hofmann, H., Siebes, A. P. J. M., & Wilhelm, A. F. X. (n.d.). Visualizing Association Rules with Interactive Mosaic Plots £. Retrieved from http://ai2-s2-pdfs.s3.amazonaws.com/5369/2bc97691ab74b04dfa1aec1266f2d5e688b0.pdf

Inselberg, A., & Dimsdale, B. (1991). Parallel coordinates. *Human-Machine Interactive Systems*. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4684-5883-1_9

Interactive Visualization: Methodologies. (n.d.). Retrieved March 2, 2016, from http://ccom.unh.edu/vislab/VisCourse/Methodology.html

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1–8. https://doi.org/10.1109/2945.981847

Kolatch, E., & Weinstein, B. (2001). CatTrees: Dynamic Visualization of Categorical Data Using Treemaps.

Kopanakis, I., & Theodoulidis, B. (n.d.). VISUAL DATA MINING & MODELING TECHNIQUES 1.2 Overview of Visual Data Mining. Retrieved from http://s3.amazonaws.com/academia.edu.documents/30796201/VDM_Techniques_KDD2001.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1490575494&Signature=iKfcTuYGWqTr1OEHp1HdOvFFXz0%3D&response-content-disposition=inline%3Bfilename%3DVisual_Data_Mining_and_Modeling_Techniqu.pdf

Lei, H., Xie, C., Shang, P., Zhang, F., Chen, W., & Peng, Q. (2016). Visual Analysis of User-Driven Association Rule Mining. In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction - VINCI '16* (pp. 96–103). New York, New York, USA: ACM Press. https://doi.org/10.1145/2968220.2968222

Li, K. (2015). On Integrating Information Visualization Techniques into Data Mining: A Review. *arXiv Preprint arXiv:1508.06576*, *7250*(December), 2–5. Retrieved from http://arxiv.org/abs/1503.0202

Ma, S., & Hellerstein, J. L. (1999). Ordering categorical data to improve visualization. *Infovis-99*, (1), 1–4. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.7388&rep=rep1&type=pdf

Michael Hahsler, S. (2017). *ArulesViz*. Retrieved from http://lyle.smu.edu/IDA/arules/

Muhlbacher, T., Piringer, H., Gratzl, S., Sedlmair, M., & Streit, M. (2014). Opening the black box: Strategies for increased user involvement in existing algorithm implementations. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1643–1652. https://doi.org/10.1109/TVCG.2014.2346578

Rosenbaum, R., & Schumann, H. (n.d.). Chances and Limits of Progression in Visualization.

Schonlau, M. (2003). Visualizing Categorical Data Arising in the Health Sciences Using Hammock Plots. *In Proceedings of the Section on Statistical Graphics, American Statistical Association; 2003, CD-ROM*, 1–7.

Sedlmair, M., Meyer, M., & Munzner, T. (2012). Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2431–40. https://doi.org/10.1109/TVCG.2012.213

Seifert, C., & Lex, E. (2009). A visualization to investigate and give feedback to classifiers. *Proceedings of the European Conference* .... Retrieved from http://mics.fim.uni-passau.de/wp-content/papercite-data/pdf/seifert2009.pdf

Shen, Z., Sun, J., Shen, Y., & Li, M. (n.d.). R-Map : Mapping Categorical Data for Clustering and Visualization Based on Reference Sets, 992–998.

Shiraishi, K., Misue, K., & Tanaka, J. (2009). A Tool for Analyzing Categorical Data Visually with, 342–351.

Shneiderman, B. (2002). Inventing Discovery Tools: Combining Information Visualization with Data Mining. *Information Visualization*, *1*(1), 5–12. https://doi.org/10.1057/palgrave.ivs.9500006

Statsoft, I. N. C. (1995). Statistica for Windows. Tulsa, OK, 74104.

Unwin, A., Hofmann, H., & Bernt, K. (2001). The TwoKey Plot for Multiple Association Rules Control (pp. 472–483). https://doi.org/10.1007/3-540-44794-6_39

Vivacqua, A. S., Cristina, A., & Garcia, B. (n.d.). NRV : USING NESTED RINGS TO INTERACT WITH CATEGORICAL DATA.

Yan, X., Qiao, M., Li, J., Simpson, T. W., Stump, G. M., & Zhang, X. L. (2012). A Work-Centered Visual Analytics Model to Support Engineering Design with Interactive Visualization and Data-Mining. *2012 45th Hawaii International Conference on System Sciences*, 1845–1854. https://doi.org/10.1109/HICSS.2012.87

Yang, L. (2003). Visualizing Frequent Itemsets, Association Rules, and Sequential Patterns in Parallel Coordinates (pp. 21–30). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44839-X_3

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, *36*(4), 2431–2448. https://doi.org/10.1007/s10916-011-9710-5

Appendix A

Task sheets

Task sheets Supporting the understanding of association rules with visualization

You are asked to finish tasks using our tool on researchers' PC/laptop within 60 minutes at most. During this process, you are asked to think aloud. Researchers will record your audio and your activity. Suppose you were a shopping mall analyst, here is a tool to help you interactively mining the data points. From the data points, you may find patterns help you sell products better. Before mining the data set, here are some basic introductions.

Item: a product customer bought from the market.
e.g. An apple

Frequent: the frequency of buying certain product among all the market transactions.
e.g. Among all the 100 transactions, apple is bought 50 times,
the frequency of this item is 0.5 (50/100=0.5)

Item set: a list of products customer buy together
e.g. Apple, banana, and milk

Support: the probability of buying certain products among all the market transactions.
e.g. Among all the 100 transactions, apple and banana is bought together for 10 times, the support of this item set is 0.1 (10/100=0.1)

Rule: a causal relationship. Customer buy certain items cause the result that he will also buy certain items. Note that the cause can be multiple items, while the effect is only one item.
e.g. People who buy apple will also buy banana.

Confidence: The probability of buying certain products when some product has been bought. It is calculated in this way: Support of the item set / Frequency of cause item
E.g. Among all the 100 transactions,
-Apple is bought 50 times, the frequency of this item is 0.5. (50/100=0.5)
-Banana is bought 30 times, the frequency of this item is 0.3. (30/100=0.3)
-Apple and banana is bought together for 10 times
The support of this item set is 0.1. (10/100=0.1)
-Rule: people buy apple will also buy banana.
The confidence of this rule is 0.2: (support of item set 0.1)/(support of apple 0.5)→(0.1/0.5=0.2)
-Rule: people buy banana will also buy apple. The confidence of this rule is 0.33: (support of item set 0.1)/(support of banana 0.3)→(0.1/0.3=0.33)

You will do these tasks as follow, during the time you are doing your tasks, please think aloud,

- Task 1:
  - As shown on the screen, each item is represented as a circle, and the circle's radius represents the item's frequency. could you find out the what item people most likely to buy?
- Task 2:
  - As shown on the screen, each item-set is represented as an arc. The arc connects several vertexes means that such items are in the same item-set. Could you find out an item-set? How many items are in this item-set? What are these items? What is its support?
  - When you click on an arc, an item-set will also show up on the right panel as bar charts. Could you compare two different item-sets? What are the difference between them?
  - Could you pick top three sets of goods that people may buy together?
- Task 3:
  - As shown on the screen, each rule is represented as a bar below the horizontal line. The bar represents a causal relation, which means people buy several items predicts that they will probably buy certain item. The probability is presented as the height of the bar. Lines connecting the bar are the cause, and the bar itself is the effect. Could you find out a rule/causation? Which items are the cause, and which item is the effect?
  - When you click on a bar, a rule will also show up on the right panel as a figure with arrow. Could you compare two different rule/causal relations? What are the difference between them?
  - Could you pick up three best predictions?

You will do these tasks as follow, during the time you are doing your tasks, please think aloud,

- Task 1:
  - As shown on the screen, each item is represented as a number, could you find out the what item people most likely to buy?
- Task 2:
  - As shown on the screen, each item-set is represented as an entry. One entry means that such items are in the same item-set. Could you find out an item-set? How many items are in this item-set? What are these items? What is its support?
  - Could you compare two different item-sets? What are the difference between them?
  - Could you pick top three sets of goods that people may buy together?
- Task 3:
  - As shown on the screen, each rule is represented as an entry. The arrow represents a causal relation, which means people buy several items predicts that they will probably buy certain item. The probability is presented using conf number. Could you find out a rule/causation? Which items are the cause, and which item is the effect?
  - Could you compare two different rule/causal relations? What are the difference between them?
  - Could you pick up three best predictions?

Appendix B

Interview Questions

Interview Questions

-Do you have any prior knowledge? Have you learned anything about data mining? Could you measure your prior knowledge level between 1-5(1-low, 5-high)?

-During the time you doing these tasks, do you have any troubles? If yes, what troubles you?

-During the time you doing these tasks, what aspects do you think this tool will help you understand the relationship of data? What aspects do you think will confuse you and make it difficult to understand?

-When you were doing task *, I saw that you do it smoothly/tough. Why?

-Suppose you were a market analyst, which one do you prefer? Text view or visualization view? Why?

-Do you have any suggestions to improve this tool?

Appendix C

Text View of Association Rule Mining

```
Dataset:  data/1000/1000-out1.csv  MinSup:  0.01  MinConf:  0.5;
================================================================;
1 :   13,     9      support= 0.01;
2 :   17,     29     support= 0.018;
3 :   17,     47     support= 0.026;
4 :   42,     5      support= 0.01;
5 :   23,     24     support= 0.033;
6 :   45,     9      support= 0.01;
7 :   14,     5      support= 0.013;
8 :   4,      9      support= 0.049;
9 :   2,      46     support= 0.039;
10 :   40,    43     support= 0.021;
11 :   37,    45     support= 0.025;
12 :   23,    33     support= 0.01;
13 :   27,    28     support= 0.053;
14 :   23,    42     support= 0.011;
15 :   31,    5      support= 0.01;
16 :   16,    45     support= 0.033;
17 :   22,    5      support= 0.058;
18 :   1,     19     support= 0.04;
19 :   33,    42     support= 0.038;
20 :   14,    44     support= 0.034;
21 :   16,    32     support= 0.04;
22 :   49,    7      support= 0.024;
23 :   0,     46     support= 0.047;
24 :   12,    31     support= 0.044;
25 :   14,    22     support= 0.012;
26 :   29,    47     support= 0.018;
27 :   46,    0,     2      support= 0.038;
28 :   47,    29,    17     support= 0.018;
29 :   18,    3,     35     support= 0.038;
30 :   15,    7,     49     support= 0.023;
31 :   32,    45,    16     support= 0.032;
32 :   31,    12,    36,    48     support= 0.031;
33 :   11,    45,    37,    7      support= 0.024;
34 :   24,    23,    43,    40,    41     support= 0.019;
35 :   24,    41,    43,    40,    23     support= 0.019;
Skyline Itemsets:  35;
Rule 1 :       17   --> 47         [sup= 0.026  conf= 0.509803921569 ];
Rule 2 :       24   --> 23         [sup= 0.033  conf= 0.5 ];
Rule 3 :       4    --> 9   [sup= 0.049  conf= 0.538461538462 ];
Rule 4 :       9    --> 4   [sup= 0.049  conf= 0.544444444444 ];
Rule 5 :       2    --> 46         [sup= 0.039  conf= 0.541666666667 ];
Rule 6 :       27   --> 28         [sup= 0.053  conf= 0.588888888889 ];
Rule 7 :       28   --> 27         [sup= 0.053  conf= 0.519607843137 ];
Rule 8 :       22   --> 5   [sup= 0.058  conf= 0.537037037037 ];
Rule 9 :       5    --> 22         [sup= 0.058  conf= 0.563106796117 ];
Rule 10 :      19   --> 1   [sup= 0.04  conf= 0.526315789474 ];
Rule 11 :      32   --> 16         [sup= 0.04  conf= 0.526315789474 ];
Rule 12 :      0    --> 46         [sup= 0.047  conf= 0.559523809524 ];
Rule 13 :      46   --> 0   [sup= 0.047  conf= 0.552941176471 ];
Rule 14 :      12   --> 31         [sup= 0.044  conf= 0.556962025316 ];
Rule 15 :      0,   2      --> 46         [sup= 0.038  conf= 0.95 ];
Rule 16 :      18,  3      --> 35         [sup= 0.038  conf= 0.926829268293 ];
Rule 17 :      18,  35     --> 3   [sup= 0.038  conf= 0.826086956522 ];
Rule 18 :      3,   35     --> 18         [sup= 0.038  conf= 0.974358974359 ];
Rule 19 :      15,  7      --> 49         [sup= 0.023  conf= 0.741935483871 ];
Rule 20 :      15,  49     --> 7   [sup= 0.023  conf= 0.958333333333 ];
Rule 21 :      32,  45     --> 16         [sup= 0.032  conf= 1.0 ];
Rule 22 :      31,  12,    36     --> 48         [sup= 0.031  conf= 0.775 ];
Rule 23 :      31,  12,    48     --> 36         [sup= 0.031  conf= 1.0 ];
Rule 24 :      31,  36,    48     --> 12         [sup= 0.031  conf= 1.0 ];
Rule 25 :      12,  36,    48     --> 31         [sup= 0.031  conf= 1.0 ];
```

```
Rule 26 :       11,  45,    37     --> 7   [sup= 0.024  conf= 1.0 ];
Rule 27 :       11,  45,    7      --> 37         [sup= 0.024  conf= 1.0 ];
Rule 28 :       11,  37,    7      --> 45         [sup= 0.024  conf=
0.888888888889 ];
Rule 29 :       45,  37,    7      --> 11         [sup= 0.024  conf= 1.0 ];
Rule 30 :       24,  23,    43,    41     --> 40         [sup= 0.019  conf= 1.0 ];
Rule 31 :       24,  23,    40,    41     --> 43         [sup= 0.019  conf=
0.678571428571 ];
Rule 32 :       24,  43,    40,    41     --> 23         [sup= 0.019  conf= 1.0 ];
Rule 33 :       23,  43,    40,    41     --> 24         [sup= 0.019  conf= 1.0 ];
Rule 34 :       24,  41,    43,    23     --> 40         [sup= 0.019  conf= 1.0 ];
Rule 35 :       24,  41,    40,    23     --> 43         [sup= 0.019  conf=
0.678571428571 ];
Rule 36 :       24,  43,    40,    23     --> 41         [sup= 0.019  conf= 1.0 ];
Rule 37 :       41,  43,    40,    23     --> 24         [sup= 0.019  conf= 1.0 ];

Freq:

[(0, 0.084), (1, 0.085), (2, 0.072), (3, 0.078), (4, 0.091), (5, 0.103), (6,
0.034), (7, 0.093), (8, 0.037), (9, 0.09
), (10, 0.041), (11, 0.068), (12, 0.079), (13, 0.056), (14, 0.095), (15,
0.073), (16, 0.081), (17, 0.051), (18, 0.084
), (19, 0.076), (20, 0.04), (21, 0.044), (22, 0.108), (23, 0.082), (24, 0.066),
(25, 0.038), (26, 0.047), (27, 0.09),
 (28, 0.102), (29, 0.061), (30, 0.049), (31, 0.091), (32, 0.076), (33, 0.078),
(34, 0.042), (35, 0.075), (36, 0.084),
 (37, 0.065), (38, 0.026), (39, 0.055), (40, 0.066), (41, 0.072), (42, 0.082),
(43, 0.062), (44, 0.077), (45, 0.094),
 (46, 0.085), (47, 0.074), (48, 0.077), (49, 0.059)]
```