KIT
Karlsruhe Institute of Technology

Interactive Visualization of Correlations in High-Dimensional Streams

Bachelor's Thesis of

Yimin Zhang

at the Department of Informatics
Institute for Program Structures and Data Organization (IPD)

Reviewer:	Prof. A
Second reviewer:	Prof. B
Advisor:	M.Sc. C

01. Feb 2019 – xx. Month 2019

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

. . . .

(Yimin Zhang)

Abstract

A fundamental task of Data Mining is to estimate the correlation between the attributes of a data set. Knowing the relationship between a set of variables, one can infer useful knowledge about external, a priori unknown outcomes.

In contrast to static data, the data is often available as a stream, i.e., it is an infinite, ever evolving sequence of observations. Concepts learned at a certain time cannot be expected to hold in the future. Therefore, correlation estimation should be a continuous process.

Also, the data is often high-dimensional, i.e., it contains hundreds or thousands of dimensions. Besides the computational burden to estimate the correlation between many subsets, it becomes difficult for a human observer to extract knowledge from the results. The task becomes even more difficult if one considers correlations between more than two variables, because the size of the result increases exponentially.

The aim of this bachelor thesis is to develop a graphical interface for data scientists, dedicated to the visualization of correlation in user-given data streams. With this interface, available for example as a web-service, users could provide their own data sets. Then, the system's backend would estimate the correlations, and provide a visualization of the results, for example, via force-directed graphs. Users will be able to interact in several ways with the interface, by setting parameters to tune the visualization. We will evaluate the benefits of our interface via controlled user studies.

Zusammenfassung

Die Abschätzung der Korrelation von Attribute in einer Datenmenge ist einer der grundlegenden Aufgaben von Data Mining. Wenn man die Beziehung von Variablen kennt, dann kann man einige nützliche Ausgaben über zusätzliche und unbekannte Informationen folgern.

Normalerweise sind die Daten als Datenfluss verfügbar, d.h. Es ist unendlich und sogar evolutionär. Die zur-zeitigen schon erkennende Begriffe und Resultaten kann man in der Zukunft nicht mehr benutzen. Deshalb muss die Abschätzung der Korrelation ständig werden.

Ein anderes Problem ist hohe Dimension. Die Daten enthalten oft mehr als 100 oder sogar 1000 Dimensionen, sodass es ist schwierig für das Rechnen der Daten. Es ist auch schwer für ein Mensch um Daten zu analysieren. Wenn es um die Korrelation über mehr als zwei Variablen geht, wächst das Rechnen der Datenmenge exponentiell an.

Das Ziel dieser Arbeit ist die Entwicklung von einer graphischen Schnittstelle für die Daten Wissenschaftler, um die Korrelation der Daten zu visualisieren. Mit dieser Schnittstelle, als zum Beispiel Web-Service, laden die Benutzer selbst Datenmenge hoch. Danach wird das Backend von System die Korrelationen von Attribute berechnen und eine Visualisierung von Daten ausgeben, zum Beispiel, durch Force-Directed Graph. Es ist auch möglich für die Benutzer mit der Schnittstelle Parameter aufzustellen, um die Visualisierung zu verbessern. Zum Schluss bewerten wir die Vorteile und Nachteile dieser Schnittstelle durch einige Anwendungsfälle.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Correlation analysis aims at discovering and summarizing the relationship between the attributes of a data set. Knowing the relationship between a set of variables, one can infer useful knowledge about external, a priori unknown outcomes.

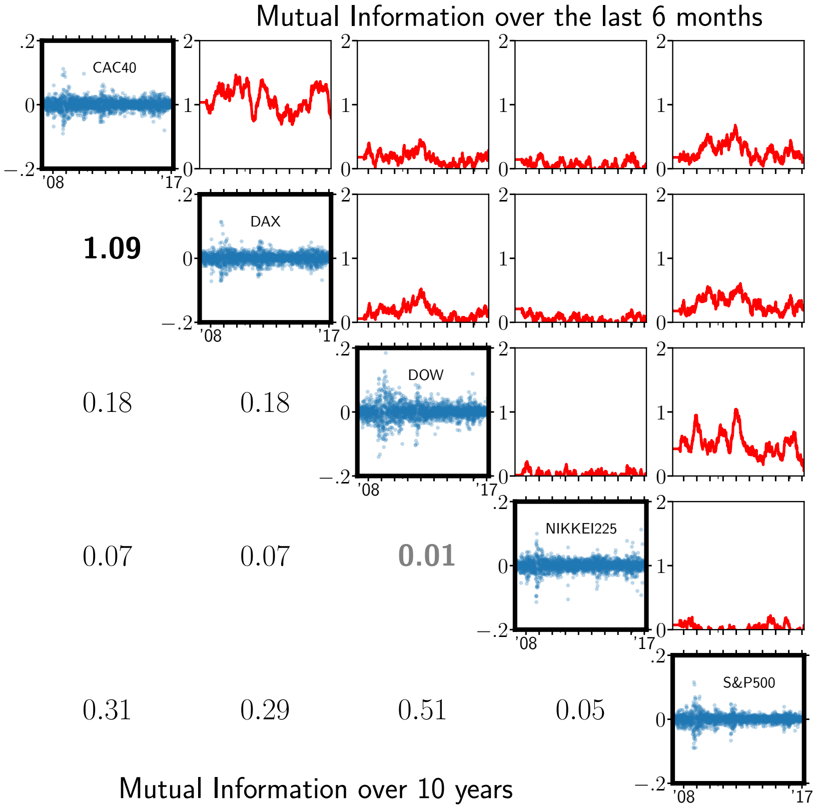


Figure 1.1: Mutual information between the return of financial indicies

For example, we can measure the stock relationship via mutual information in a stock market that may undergo large fluctuations of stock prices. Figure 1.1 represents the Mutual Information between the return of important financial indices over 10 years over a sliding windows of 6 months. Such a correlation monitoring system could be of great help to a financial analysis.
As an instance, investing in low correlated market indices, such as CAC40 and NIKKEI, may be desirable minimize the risks of a portfolio: If one of the stock decreases, the other stock is unlikely to be affected too.This is how mutual information helps us to maximize our wealth.
Therefore, analyzing the correlation between different attributes helps us to understand their relationship. In general, the correlations between attributes remain same or change gradually. If the correlation structure changes brutally, which often indicates a sudden peak or valley, we can infer that one of the attributes may have an enormous change or the relationship between them may differ thoroughly. Figure 1.2 shows a simulation of "gradual" and "abrupt" changes for correlations.

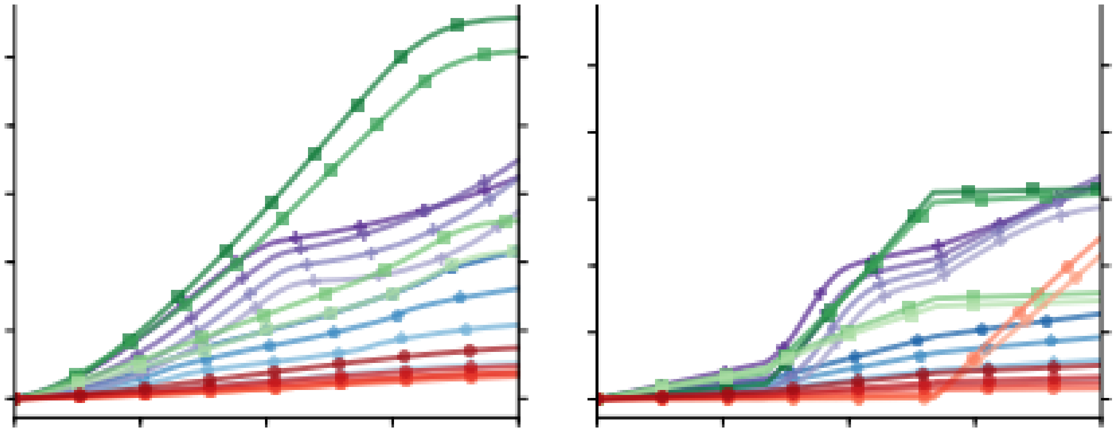


Figure 1.2: "Gradual" changes Vs "abrupt" changes for correlations

1.2 Challenges

The challenges of analyzing high-dimensional data streams is twofold: the evolving nature of streams and the high-dimensionality.

1.2.1 The evolving nature of streams

In contrast to static data, the data is often available as a stream, i.e., it is an infinite, ever evolving sequence of observations. As the concepts learned at a certain time cannot be expected to hold in the future, correlation analysis should be a continuous process.

1.2.2 The high-dimensionality

Also, the data is often high-dimensional, i.e., it contains hundreds or thousands of dimensions. In the case of streams with many dimensions, it is difficult to extract actionable insights from the correlation matrix, as the number of pairs of attributes increases quadratically and the coefficients evolve over time in unforeseen ways. For the pairwise correlation analysis of any data steam with n components, one need to compute the correlation between $\dfrac{n * (n-1)}{2}$ pairs, i.e., with $n = 5$, one needs to compute 10 pairs, with $n = 10$, one needs to compute 45 pairs, it becomes difficult and impossible to understand the result of the correlation analysis. The visualization of different number of pairs of attributes shown in Figure 1.3 illustrates this, which also indicates $O(n^2)$ of the time complexity.

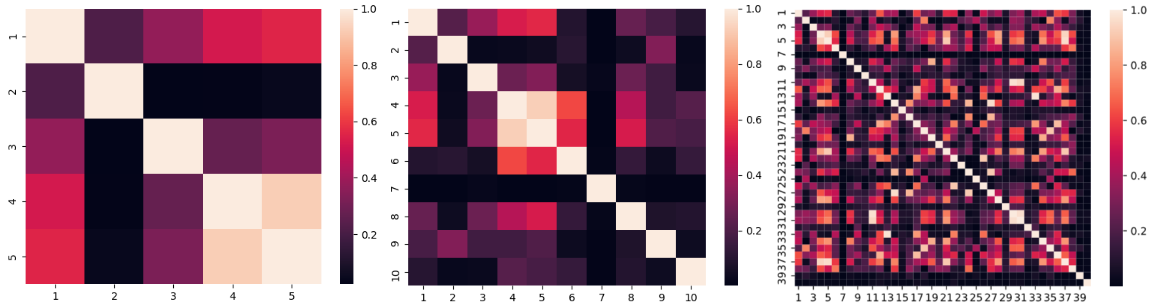


Figure 1.3: Visualization of different number of pairs of attributes

1.3 Goal of the thesis

The goal of this thesis is to propose and evaluate new tools for the interactive visualization of correlation in high-dimensional streams. In this thesis, we are going to answer the following three questions:

- What visualization methods are the most appropriate to visualize correlation?
- How can one visualize the evolution of correlation results?
- What are the desirable features of a correlation monitoring interface?

To answer these questions, we first search for some interactive visualization methods and then develop an interface to visualize correlations. This interface can be available in a browser and should be user friendly, which means that users provide their own data sets and the system's back-end calculates the correlations and then provides the visualization of the results, for example, via force-directed graphs. Users are also able to interact in several ways with this interface, such as setting parameters to tune the visualization. At last, we evaluate the benefits of our interface systematically via controlled user studies and discuss the advantages and disadvantages of each visualization method. Based on the results, we discover the appropriate visualization methods of correlations in high-dimensional streams and the desirable features of the interactive interface.

1.4 Thesis outline

The process of investigation can be split into three parts, that is performing a literature review, developing a graphical interface and evaluating this interface, which also leads to the three parts of this thesis, the visualization methods, the design and implementation of interface, and the evaluation of this interface.
In Chapter 2, the first part of the thesis, we introduce some state-of-art methods for correlation visualization. Section 2.1 introduces the correlation matrix, which is the standard tool of performing the correlation analysis. In Section 2.2, Section 2.3 and Section 2.4, some other examples of data visualization are discussed.
Chapter 4 is the main part of this thesis, which consists the design of the interface in Section 4.1 and the implementation of this interface in Section 4.2.
We introduce the evaluation of the developed interface in Chapter 5, which is the last part of this thesis. Chapter 5 contains the experimental settings of controlled user studies for evaluation in Section 5.1 and the result in Section 5.2.
In the end, we summarize the whole thesis.

2 Visualization Methods

In this chapter, we first introduce the correlation matrix in Section 2.1 and then introduce some useful interactive visualization methods for correlations in Section 2.2, Section 2.3 and in Section 2.4.

2.1 Correlation Matrix

The most familiar measure of dependence between two quantities is the Pearson product-moment correlation coefficient [6], known as "Pearson's correlation coefficient". It is obtained by dividing the covariance of the two variables by the product of their standard deviations. The population correlation coefficient $\rho_{X,Y}$ between two random variables X and Y and standard deviations σ_X and σ_Y is defined as

$$\rho_{X,Y} = \mathrm{corr}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{2.1}$$

where cov means covariance, and corr is a widely used alternative notation for the correlation coefficient. The Pearson correlation is defined only if both of the standard deviations are finite and non-zero.

The standard tool of correlation analysis is the computation of a correlation matrix, which is used to investigate the dependence between multiple variables at the same time.

2.2 Heatmap

A common visualization is the heatmap[7], which is originated in 2D displays of the values in a data matrix. The Figure 2.1 is a heatmap of a correlation matrix, in which the variables with strong correlation (high values) are printed in light colour and those with low correlation are in dark colour.

2.3 Force-Directed Graph

Also, Force-Directed Graph[4] is a useful visualization, which assigns forces among the set of edges and the set of nodes of a graph drawing. The purpose of it is to position the nodes of a graph in two-dimensional or three-dimensional space so that all the edges are of more or less equal length and there are as few crossing edges as possible. In such a

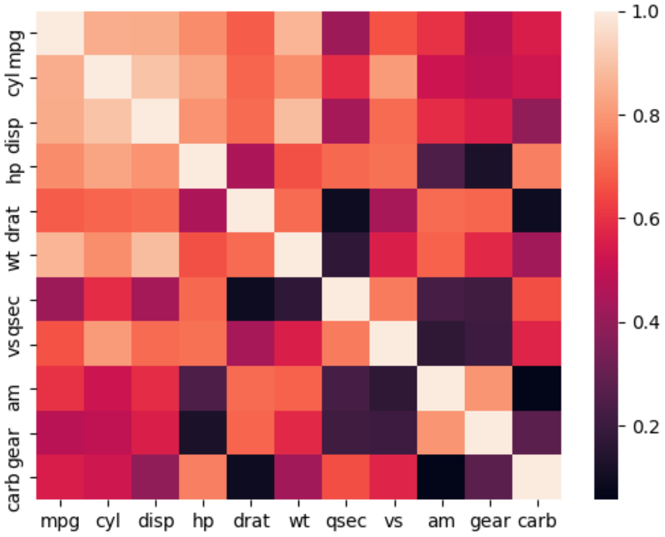


Figure 2.1: Correlation Matrix of a data set in heatmap

simulation, the forces are applied to the nodes, pulling them closer together or pushing them further apart. This can be used to simulate the relationship of different attributes throughout the time, in which the force is the representation of correlation matrix. The Figure 2.2 shows an example of Force-Directed Graph, which actually uses the same data set in Figure 2.1.

2.4 Bar Graph

A bar chart[3] or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a line graph. One axis of the chart shows the specific categories being compared, and the other axis represents a measured value. Bar charts provide a visual presentation of categorical data. Categorical data is a grouping of data into discrete groups, such as months of the year, age group, shoe sizes, and animals. These categories are usually qualitative. The Figure 2.3 is an example of bar graph.

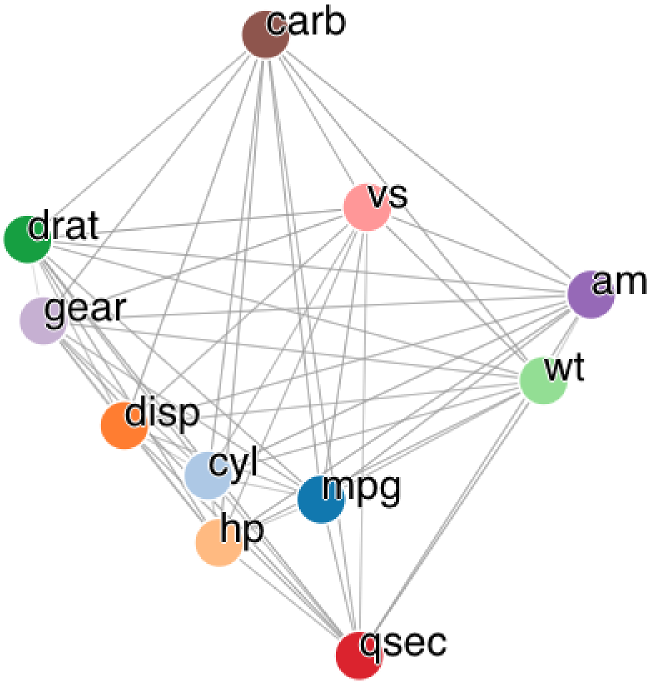


Figure 2.2: Force-Directed Graph

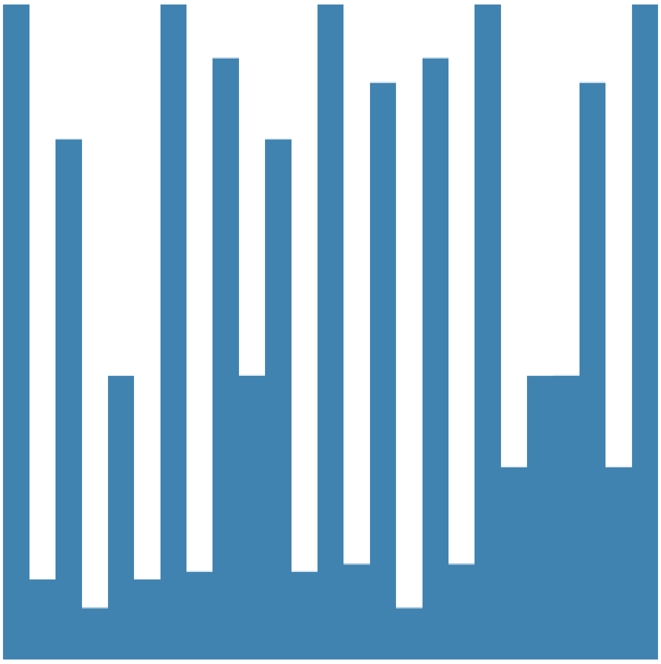


Figure 2.3: Bar Graph

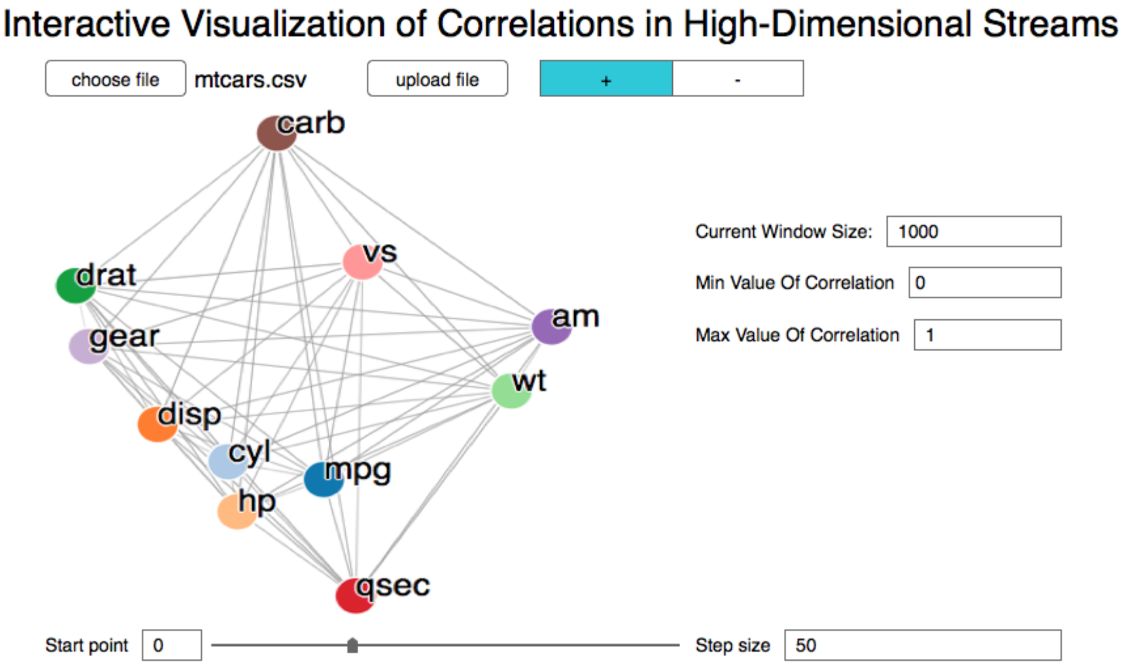
3 Related Work

4 Interface

To research questions, we develop an interface in a browser available as a web-service. In Section 4.1, we describe the mock-up of this interface and its available functions. Ans we introduce the details of implementation in Section 4.2.

4.1 Design

Figure 4.1 is the mock-up of this interface. Users can upload their data sets as a csv file. After the calculation in the back-end, the visualization of data correlation is shown in the website, for example, via a force-directed graph. In the mock-up, a sliding window with start point and step size is supposed to be used to represent the continuous process of data throughout the time. Also, we have some simple user settings, such as changing the window size, setting the minimum and maximum of correlation to be visualized.



Interactive Visualization of Correlations in High-Dimensional Streams

mtcars.csv

Current Window Size: 1000

Min Value Of Correlation 0

Max Value Of Correlation 1

Start point 0

Step size 50

Figure 4.1: Mockup of the interface

4.2 Implementation

4.2.1 D3.js

D3.js (Data-Driven Documents)[5] is a data-driven JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of the widely implemented SVG, HTML5, and CSS standards and allows great control over the final visual result. SVG stands for Scalable Vector Graphics which is technically an XML based markup language. It is a great tool to display icons, logos, illustrations or charts, which is supported in all major browsers and requires no third-party lib because of owning the DOM interface;. In our project, we mainly use D3.js to implement the visualization of correlations in high-dimensional data streams.

5 Evaluation Via User Studies

...

5.1 Experimental Settings

...

5.2 Result

...

6 Conclusion

...

Bibliography

[1] Albuquerque, G., Eisemann, M., Lehmann, D. J., Theisel, H., and Magnor, M. Quality-Based Visualization Matrices. *In Proceedings of the Vision, Modeling and Visualization* (2009), 341–350.

[2] B, L. K., Riekenbrauck, N., Thevessen, D., Pappik, M., Stebner, A., Kunze, J., Meissner, A., Shekar, A. K., and Emmanuel, M. Machine Learning and Knowledge Discovery in Databases. 404–408.

[3] Kelley, W. M.; Donnelly, R. A. The humongous book of statistics problems. *New York* (2009).

[4] Kobourov, S. G. Spring embedders and force-directed graph drawing algorithms. *eprint arXiv:1201.3011* (2012).

[5] Murray, S. Interactive data visualization for the web, an introduction to designing with d3, 2013.

[6] Rodgers, J. L.; Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *The American Statistician* (1988).

[7] Wilkinson, Leland; Friendly, M. The history of the cluster heat map. *The American Statistician* (2009).