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. Jan 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 and Challenges

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.
For example, we can measure the stock relationship via mutual information in a stock market that may undergo large fluctuations of stock prices. ~~This~~ Figure 1.1 represents the mutual information between the return of important financial indices over 10 years ~~and~~ over a sliding windows of 6 months. We can conclude that if we buy portfolios based on two low correlated market indices, like CAC40 and S&P, then we can ensure that even if one stock has a large decrease, the other stocks we buy would not be effected greatly. This is how mutual information helps us to maximize our wealth.

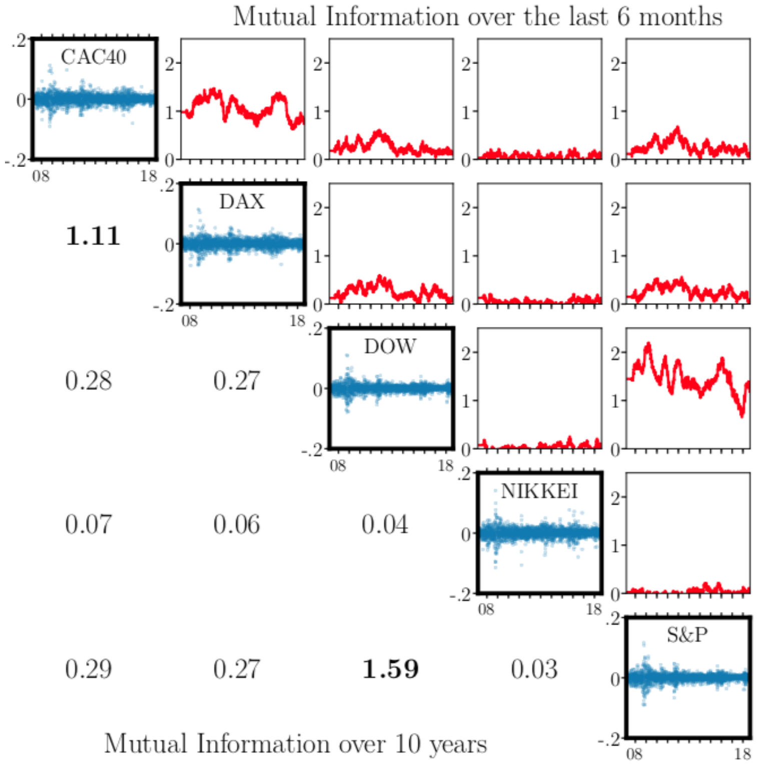


Figure 1.1.: Mutual information between the return of financial indicies

Therefore, analyzing the correlation between different attributes helps us to get close to

their relationship. Also, if the correlation structure changes brutally, we can infer that something may go wrong.
When it comes to analyzing high-dimensional data streams, many challenges have to be overcome. 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.
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. We can conclude from the Figure 1.2 that if we have n attributes, we have to calculate $n * (n-1)/2$ pairs of correlations, which means that the time complexity is $O(n^2)$.

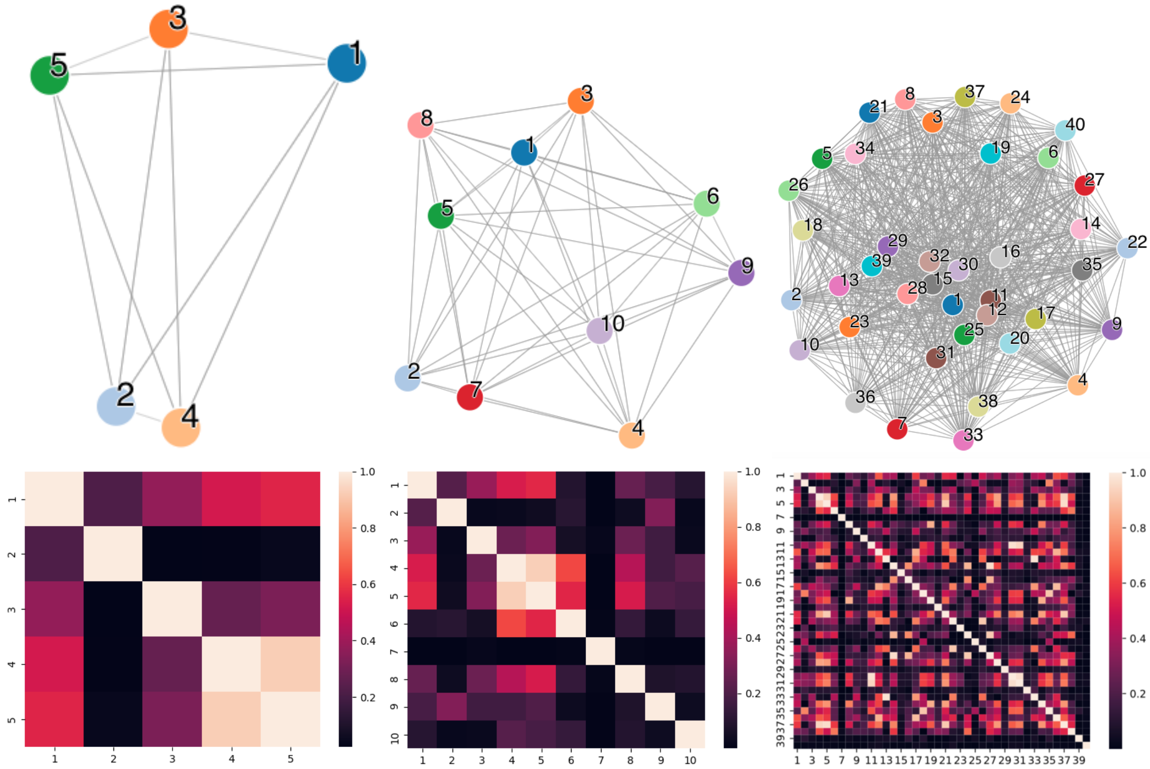


Figure 1.2.: Visualization of different number of pairs of attributes

1.2. Goal of the thesis

The goal of this thesis is to do an investigation on interactive correlation visualization. In this thesis, we are going to answer the following three questions. They are: What visualization methods are the most appropriate to visualize correlation ~~matrices~~? How can one visualize the evolution of correlation ~~results in the visualization?~~ What are the desirable features of a correlation monitoring interface? And the process can be split into three parts, that is performing a literature review, developing a graphical interface and evaluating this interface. 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 backend would calculate the correlations and then provide the visualization of the results, for example, via force-directed graphs. Users will also be able to interact in several ways with this interface, such as setting parameters to tune the visualization. At last, we would evaluate the benefits of our interface systematically via controlled user studies.

1.3. Structure of the thesis

This thesis could be divided into three parts, the technology part, design and implement of interface, and evaluation of this interface.
In the first part about the technology chapter 2, some basic knowledge about visualization of correlations would be introduced. section 2.1 introduced the correlation matrix, which is the standard tool of performing the correlation analysis. In section 2.2, some examples of data visualization would be discussed. And the coding tool to perform these visualizations would be introduced in section 2.4.
chapter 4 is the main part of this thesis, which consists the design of the interface in section 4.1, the implement of this interface in section 4.2 and the test of this developed interface in section 4.3.
The last part of this thesis is the evaluation of the developed interface, which would be discussed in the chapter 5.chapter 5 is consisted of the design of controlled user studies in section 5.1, the performing of controlled user studies in section 5.2 and the result in section 5.3.
Finally, there comes the summary of the whole thesis.

2. Technology

In this chapter, some basic technology about interactive visualization of correlations would be intorduced. First, we would introduce the correaltion matrix in section 2.1.Then we would introduce some useful visualizations in section 2.2. To perform the visualization of correlation matrix in our interface, we use D3.js, which would be discussed in section 2.4.

2.1. Correlation Matrix

The most familiar measure of dependence between two quantities is the Pearson product-moment correlation coefficient [**correlation**], or "Pearson's correlation coeffi~~cient", commonly called simply "the 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 nonzero.

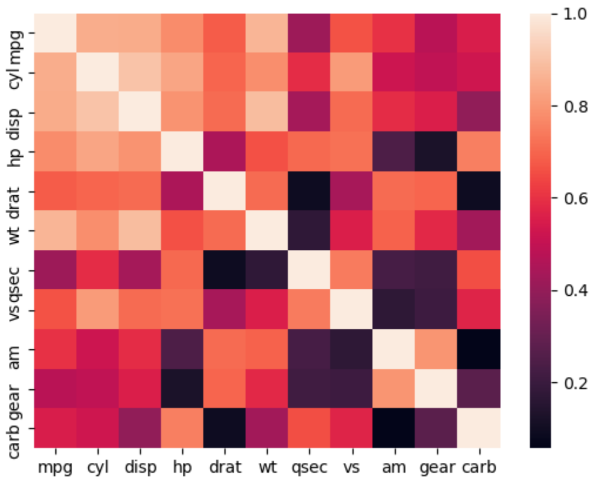


Figure 2.1.: Correlation Matrix of a data set

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. A common visualization is the heatmap[**heatmap**], 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 are printed in light color and those with low correlation are in dark color.

2.2. Force-Directed Graph

Also, Force-Directed Graph[**force**] 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 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.

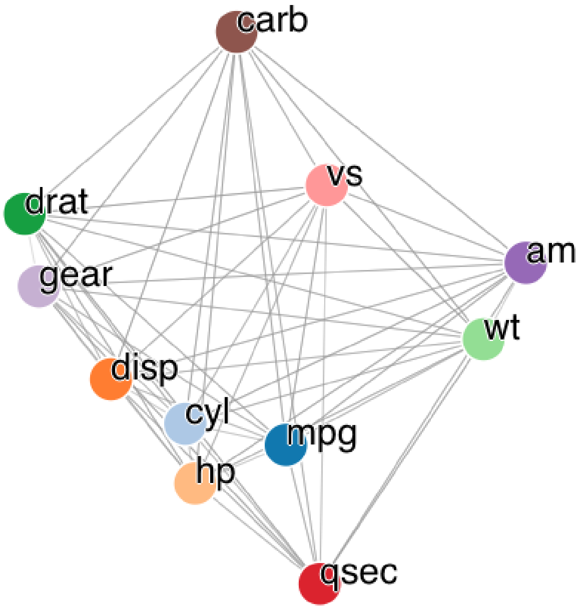


Figure 2.2.: Force-Directed Graph

2.3. Bar Chart

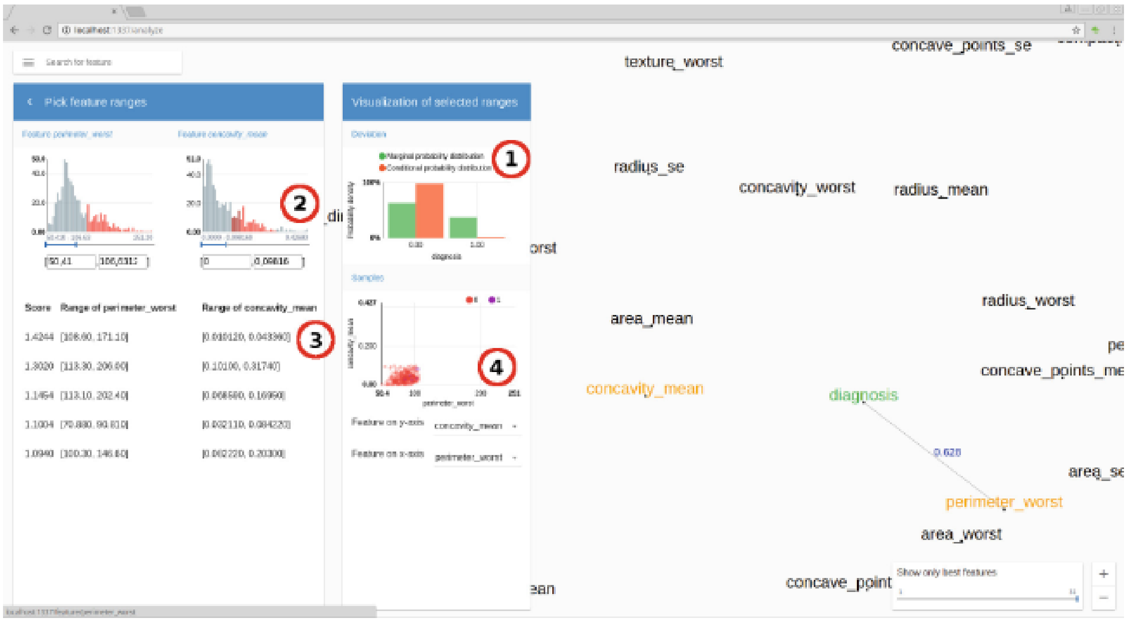
A bar chart[**bar**] or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a line graph. One axis of the 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.

2.4. D3.js

D3.js (Data-Driven Documents)[**D3**] is a 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. In our project, we mainly use D3.js to implement the visulization of correlations in high-dimensional data streams.

3. Related Work

With the exponentially increasing amount of acquired multivariate data, several multidimensional visualization techniques have been proposed during the last decades. Parallel coordinates and scatterplot matrices are widely used to visualize multi-dimensional data sets. But these visualization techniques are insufficient when the number of dimensions grows, as we have mentioned in section 1.1. To solve this problem, different approaches to preselect the best views or dimensions have been proposed in the last years.
[**Matrics**] Georgia Albuquerque et al.presented three new methods to explore multivariate data sets: a parallel coordinates matrix, in analogy to the well-known scatterplot matrix, a class-based scatterplot matrix that aims at finding good projections for each class pair, and an importance aware algorithm to sort the dimensions of scatterplot and parallel coordinates matrices.
Also, some interactive visualization frameworks are developed in the recent years.



Score	Range of perimeter_worst	Range of concavity_mean
1.4244	[108.60, 171.10]	[0.010120, 0.043360]
1.3020	[113.30, 206.60]	[0.10100, 0.31740]
1.1454	[113.10, 202.40]	[0.068500, 0.16950]
1.1004	[70.880, 90.810]	[0.002110, 0.084220]
1.0940	[100.30, 146.80]	[0.062220, 0.20300]

Figure 3.1.: Features drawn using a force-directed graph (right), with the target highlighted in green. An analysis view of two features (left) for inspecting the correlations. (Color figure online)

[**FEXUM**]Louis Kirsch et al. developed an interactive Framework for Exploring and Understanding Multivariate Correlations (FEXUM) to simultaneously visualize all feature correlations to the target and pairwise correlations. This visualization provides a layout in which a smaller distance between two features denotes a greater redundancy. In Figure 3.1, nodes represent features and weighted edges represent distances.

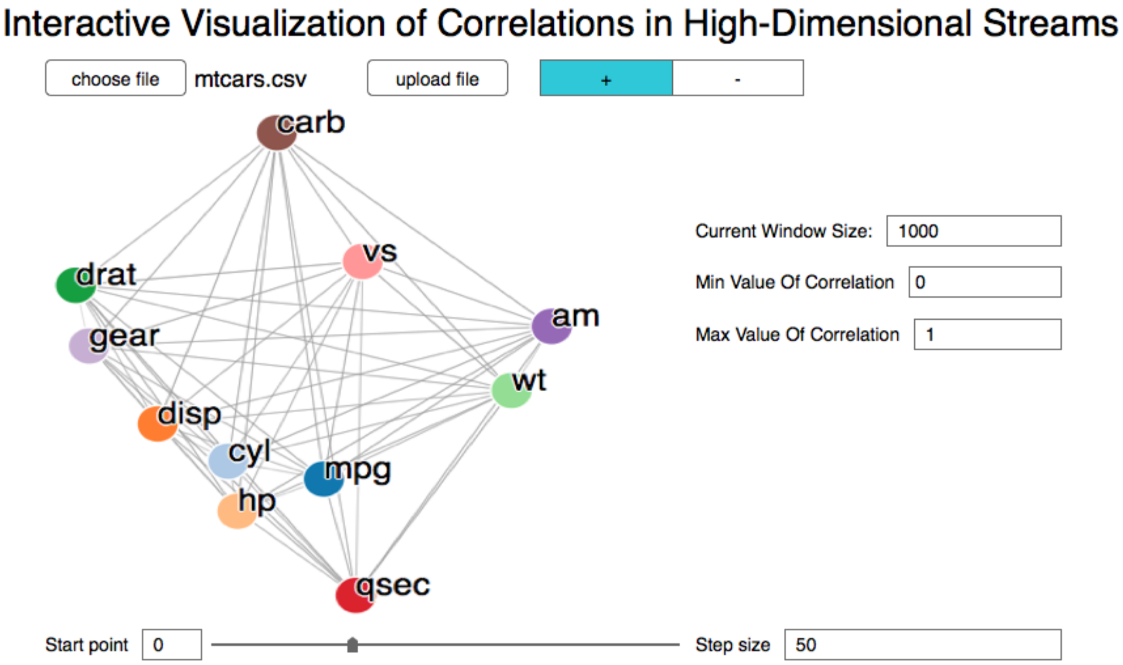
What we are going to do is not to choose a target attribute to focus on, but to represent the whole correlations of a data set. Also, visualizing a concise but useful summary of correlations in the stream over time is our goal. To make it more user-friendly, attributes with strong correlation would be closer to each other in our visualization.

4. Interface

To ~~investigate the~~ research questions ~~we have presented in the section 1.2,~~ we developed an interface in a browser available as a web-service. In section 4.1, we would declare the implement idea of this interface and its available functions. The details of implement would be presented in section 4.2 and the section 4.3 is written to show the test run of the interface.

4.1. Design of the interface

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 would be 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 would 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. Implement of the interface

…

4.3. Test of the interface

…

5. Evaluation

...

5.1. Design of the evaluation

...

5.2. Performing the evaluation

...

5.3. Result

...

6. Conclusion

. . .

This is the SDQ thesis template. For more information on the formatting of theses at SDQ, please refer to `https://sdqweb.ipd.kit.edu/wiki/Ausarbeitungshinweise` or to your advisor.

6.1. Example: Citation

A. Appendix

chap:appendix

A.1. First Appendix Section

Figure A.1.: A figure

…