Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Bachelorarbeit

Bachelorarbeit

Interactive Visualization of Correlations in High-Dimensional Streams

A fundamental task of Data Mining is to estimate the dependency 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. Also, in *Predictive Maintenance* for example, it has turned out to be very useful to look at the evolution of those dependencies. Abrupt changes in the correlation of a set of attributes often are the sign of a shift in the underlying process and can help identify failures, intrusions or equipment breakdowns.

However, the data is often available as a *stream*: in contrast to static data, streams are infinite and always evolving, so concepts learned at a certain time cannot be expected to hold in the future. This means that correlation/dependency 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 pairs/subsets, its 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 topic of this bachelor thesis is the design of interactive visualization methods to monitor the evolution of correlations in high-dimensional streams. In particular, the following aspects are of interest:

- What visualization methods are the most appropriate to visualize correlation matrices?
- How can one visualize the evolution of correlation matrices over time?
- How can one integrate multivariate correlation results in the visualization?
- All in all, what are the desirable features of a correlation monitoring interface? For example, how can users upload data und configure the desired outcome intuitively?

This results in the following tasks:

- Literature review about the visualization of correlation matrices / sparse matrices.
- Development of a front-end, e.g., an interface in a browser available as a web-service, for the visualization of correlation in user-given data streams.
- Evaluation of this front-end via controlled user studies.

Throughout this work, the student will get a deeper understanding of correlation analysis and become familiar with high-dimensional data streams. The student will train highly valuable Data Science skills, such as visualization methods and interviewing skills, and get the opportunity to build software that will be used extensively in research. For this thesis, intermediate knowledge of programming and interactive visualization libraries, such as D3.js, are helpful, but not necessary, provided the student is eager to learn.

Ansprechpartner

Edouard Fouché, M. Sc. edouard.fouche@kit.edu +49 721 608-47337 Raum: 342

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34