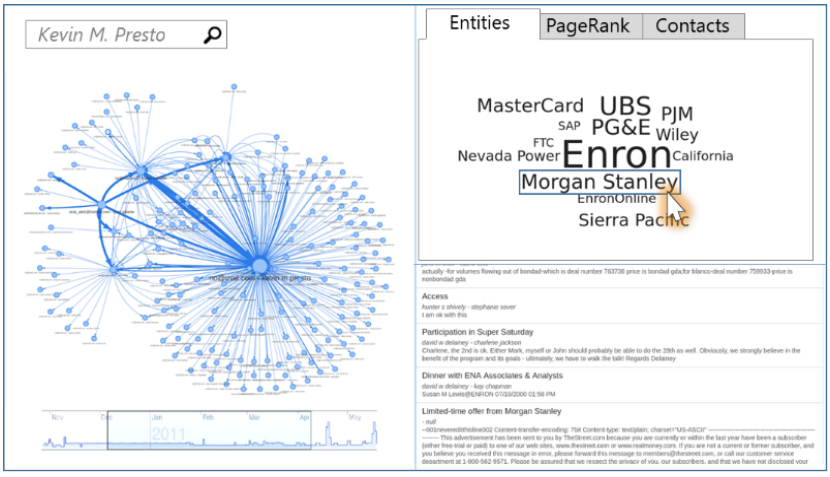
Masterproject / Data Engineering Lab
Wintersemester 2018/19
Chair: Prof. Naumann
Web Science Group

Tracking documents across large email corpora

A day in the life of an attachment

Every day, companies accumulate large amounts of heterogeneous data in the form of emails, documents like contracts and letters, or others. From text **documents and their metadata**, graphs can be extracted, where the nodes and edges are enriched with additional information. The resulting semi-structured data can be analysed by combining methods from the research areas of **Text Mining** and **Graph Analysis**.



Using metadata from a collection of emails, a social network over time can be constructed. Additional analysis of the content of respective communication can yield answers to the question: "**who knew what when**". Other methods calculate the information diffusion [2] in the social network.
This information and other inherent structures [3] extracted from heterogeneous graphs can **support the work of journalists**, auditors, or special investigators. Instead of having to read thousands of documents to get an initial sense about a dataset, extracted **information** can be **aggregated into a simplified overview** and enable more focused investigations.

Project Goal

In this Masterproject/Data Engineering Lab students will learn how to analyse large communication networks such as email data. While there is a lot of focus on emails and their metadata, attached documents are often neglected. We will develop methods to identify and analyse versions of the same document within an email corpus to see how people collaborate and attachments travel through networks.
To do so, we are first going to **classify attachments** into different categories such as contracts or forms with a **deep learning** model using text embeddings and other features. This and other information from the heterogeneous communication network, is used as context to determine if two attachments are revisions of the same document.

The **version history** of a single document is a temporal subgraph of the social network. The set of all these subgraphs is summarised into **frequent patterns.** These patterns may expose processes in a company, hierarchies, its departmental structure, or identify roles in which people are involved.
Furthermore, we will extract the **differences between revisions** and analyse their content.
We evaluate the performance of the models developed in the project by measuring how well a model can predict previously unseen data.
Journalists or auditors should be able to use this information to support and enhance their **data-driven investigations.** This allows them to track the version history of documents and also determine which kind of changes have been made by whom and when.

Methods and Datasets

For the development and evaluation of a prototype, we will make use of **real-world datasets** such as the Enron corpus [1], which is very popular among scientists, containing 600,000 emails with **1,000,000 attachments** from 150 employee inboxes. In addition, we will look at the more recent Sony Archives published by **Wikileaks**.
This project builds upon an **existing architecture** from our bachelor project “Lighthouse in the data fog” [4] which implements a **distributed processing** pipeline for files and a web interface providing overviews and a search frontend.

Contact

This project will be supervised by Dr. Ralf Krestel and Tim Repke. Thanks to the cooperation with our partners from the audit and risk assessment departments at Commerzbank, we can have our prototypes tested by domain experts and get feedback from potential users.
For questions, feel free to contact tim.repke@hpi.de.

Related Work

[1] *Introducing the Enron Corpus.* BryanKlimt, Yiming Yang. CEAS 2004
[2] *Representation Learning for Information Diffusion through Social Networks: Embedded Cascade Model.* Simon Bourigault, Sylvain Lamprier, Patrick Gallinari. WSDM 2016
[3] *Integrating Community and Role Detection in Information Networks.* Ting Chen, Lu-An Tang, Yizhou Sun, Zhengzhang Chen, Haifeng Chen, Guofei Jiang. SIAM 2016
[4] (in review) *Beacon in the Dark: A System for Interactive Exploration of Large Email Corpora.* Tim Repke, Ralf Krestel, Jakob Edding, Moritz Hartmann, Jonas Hering, Dennis Kipping, Hendrik Schmidt, Nico Scordialo, Alexander Zenner. CIKM 2018