Exploration Interface for Jointly Visualised Text and Graph Data

Tim Repke
tim.repke@hpi.uni-potsdam.de
Hasso Plattner Institute
University of Potsdam, Germany

Ralf Krestel
ralf.krestel@hpi.uni-potsdam.de
Hasso Plattner Institute
University of Potsdam, Germany

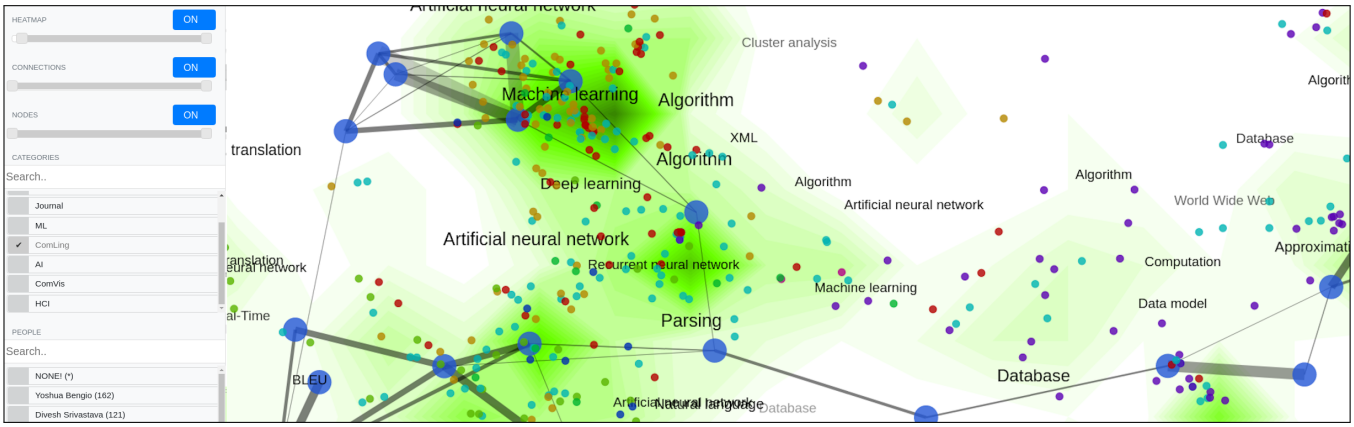

Figure 1: Screenshot of the *MODiR* interface prototype showing an excerpt of a citation network.

ABSTRACT

Many large text collections exhibit graph structures, either inherent to the content itself or encoded in the metadata of the individual documents. Example graphs extracted from document collections are co-author networks, citation networks, or named-entity-cooccurrence networks. Furthermore, social networks can be extracted from email corpora, tweets, or social media. When it comes to visualising these large corpora, traditionally either the textual content or the network graph are used. We propose to incorporate both, text and graph, to not only visualise the semantic information encoded in the documents' content but also the relationships expressed by the inherent network structure in a two-dimensional landscape. We illustrate the effectiveness of our approach with an exploration interface for different real world datasets.

1 INTRODUCTION

Substantial amounts of data is produced in our modern information society each day. A large portion of it comes from the communication on social media platforms, within chat applications, or via emails. This data has a dual characteristics: *text* and *graph*. The metadata provides an inherent graph structure given by the social network between correspondents and the exchanged messages constitute the textual content. In addition, there are many other datasets that exhibit these two facets. Some of them are found in bibliometrics, for example in collections of research publications as co-author and citation networks. In data exploration scenarios, the goal of getting an overview of the datasets at hand is insurmountable with current tools. The sheer amount of data prohibits simple visualisations of networks or meaningful keyword-driven summaries of the textual content. Data-driven journalism [2] often has to deal with leaked, unstructured, very heterogeneous data, e.g. in the context of the Panama Papers, where journalists needed to untangle and order huge amounts of information, search entities, and visualise found patterns [1]. Similar datasets are of interest in the context of computational forensics [3]. Auditing firms and law enforcement need to sift through huge amounts of data to gather evidence of criminal activity, often involving communication networks and documents [4]. Users investigating such data want to be able to quickly gain an overview of its entirety, since the large amount of heterogeneous data renders experts' investigations by hand infeasible. Computer-aided exploration tools can support their work to identify irregularities, inappropriate content, or suspicious patterns. Current tools[1] lack sufficient semantic support.

We propose *MODiR*, a scalable multi-objective dimensionality reduction algorithm, and show how it can be used to generate an overview of entire text datasets with inherent network information in a single interactive visualisation. Special graph databases enable the efficient storage of large relationship networks and provide interfaces to query or analyse the data. However, without prior knowledge, it is practically impossible to gain an overview or

IUI '20 Companion, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7513-9/20/03.
https://doi.org/10.1145/3379336.3381470

[1] e.g. https://www.nuix.com/ or https://linkurio.us/

quick insights into global network structures. Although traditional node-link visualisations of a graph can provide this overview, all semantic information from associated textual content is lost completely. Technically, our goal is to combine a network layouts with dimensionality reduction of high-dimensional semantic embedding spaces. Giving an overview over latent structures and topics in one visualisation may significantly improve the exploration of a corpus by users unfamiliar with the domain and terminology. Figure 1 contains a screenshot of our interactive prototype implementation combining graph and content information.

2 VISUALISING A LARGE TEXT CORPUS WITH NETWORK INFORMATION

We developed *MODiR* to visualise large text corpora exhibiting inherent graph structures. For example, named entity relationships can be extracted from a text corpus resulting in a graph, where nodes and edges can be represented by the context they were extracted from. Other datasets allow the analysis of collaborations, such as in co-authorship or citation networks based on academic articles. Interactions between people can be analysed from social media platforms through tweets, posts, or chats, etc.

Social networks are commonly visualised as node-link drawings, where people are shown as circles of varying size based on a weight metric and connections between them as lines connecting the circles. The layout of nodes should visually convey the inherent structure of the network graph. Beside the network, we also visualise associated textual content (documents, such as posts, emails, papers, etc.). Similar to Cartograph [5], we base the visualisation on a *document landscape*. Salient structures of the text corpus become visible in the form of more densely populated regions.To align the network and documents, the graph layout is adjusted to place the circle for a node near the documents associated with it. We focus on integrating all three principles into a single joint visualisation.

We derive our approach from state-of-the-art methods for drawing either the network layer or the document landscape. Documents are assumed to be in the form of high-dimensional vectors and entities are linked among one another and to the documents. *MODiR* is a multi-objective dimensionality reduction algorithm with three objectives that formalise the requirements stated above.

Objective (1): Similar Documents Are Near One Another. To achieve this, we use high-dimensional document embeddings as the vector distances can be interpreted as semantic similarity. Similar to tSNE, this objective aims to replicate the pairwise distances in the high-dimensional space on the two-dimensional canvas. We use a architecture similar to word2vec skip-gram with neighbourhoods to optimise this objective.

Objective (2): Dissimilar Documents Are Apart From One Another. The optimal solution to the previously defined objective would be to project all documents onto the same point on the two-dimensional canvas. In order to counteract that, we introduce negative examples for each pair of context documents.

Objective 3: Connected Entities Are Near One Another And Their Documents. Attracts documents that are related through entities. This has two implicit effects: An entity gets closer to its documents as they are attracted to it without having to explicitly compute this position.Also, related entities are implicitly attracted to one another. We do not consider a repulsing objective as the first two objectives provide enough counteracting force.

3 EXPLORING DATA LANDSCAPES

The motivation for this paper is to visualise social networks along with their respective text documents, for exploring and understanding large heterogeneous datasets. To demonstrate the effectiveness of our multi-objective layout algorithm and the interface prototype as shown in Figure 1, we use datasets from different domains, namely the Enron email corpus, scholarly publications and the co-author network, and entity networks co-mentioned in news articles.The characteristics of the networks differ greatly as the ratio between documents, nodes, and edges shows. In an email corpus, a larger number of documents is attributed to fewer nodes and the distribution has a high variance (some people write few emails, some a lot). In the academic corpora on the other hand, the number of documents per author is relatively low and similar throughout. Especially different is the news corpus, that contains one entity that is linked to all other entities and to all documents.

Users exploring such data, e.g. journalists investigating leaked data or young scientists starting research in an unfamiliar field, need to be able to interact with the visualisation. Our prototype allows users to explore the generated landscape as a digital map with zooming and panning. The user can select from categories or entities to shift the focus, which highlights characterising keywords and adjusts a heatmap based on the density of points to only consider related documents. We extract region-specific keywords and place them on top of the landscape. This way, the meaning of an area becomes clear and supports fast navigation.

4 CONCLUSIONS

In this paper, we demonstrated a method to jointly visualise heterogeneous text and network data with all its aspects on a single canvas. Therefore, we identified three principles that should be balanced by a visualisation algorithm. Our novel multi-objective dimensionality reduction algorithm provides a layout for the data. From that we generate landscapes which consist of a base-layer, the semantic *document landscape* and a *graph layer* onto which the inherent network is drawn. Apart from an implementation of the algorithm we demonstrate its results for a number of real world datasets in a prototype for an interactive exploration interface.

REFERENCES

[1] Marie-Anne Chabin. 2017. Panama papers: a case study for records management? *Brazilian Journal of Information Science: Research Trends* 11, 4 (2017), 10–13.
[2] Mark Coddington. 2015. Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting. *Digital Journalism* 3, 3 (2015), 331–348.
[3] Katrin Franke and Sargur N Srihari. 2007. Computational forensics: Towards hybrid-intelligent crime investigation. In *IAS*. IEEE, New York City, USA, 383–386.
[4] Mukundan Karthik, Mariappan Marikkannan, and Arputharaj Kannan. 2008. An intelligent system for semantic information retrieval information from textual web documents. In *IWCF*. Springer-Verlag, Heidelberg, Germany, 135–146.
[5] Shilad Sen, Anja Beth Swoap, Qisheng Li, Brooke Boatman, Ilse Dippenaar, Rebecca Gold, Monica Ngo, Sarah Pujol, Bret Jackson, and Brent Hecht. 2017. Cartograph: Unlocking spatial visualization through semantic enhancement. In *IUI*. ACM, Geneva, CH, 179–190.