Graphdatenbanken für die textorientierten e-Humanities

Dissertation im Fachgebiet Informatik von Thomas Efer, M. Sc.

Graphdatenbanken · Datenmodellierung · Recherchesysteme · e-Humanities
Text Mining · Korpusexploration · Information Retrieval

Abstract *(English Version)*

In light of the recent massive digitization efforts, most of the humanities disciplines are currently undergoing a fundamental transition towards the widespread application of digital methods. In between those traditional scholarly fields and computer science exists a methodological and communicational gap, that the so-called ”e-Humanities” aim to bridge systematically, via interdisciplinary project work. With text being the most common object of study in this field, many approaches from the area of Text Mining have been adapted to problems of the disciplines. While common workflows and best practices slowly emerge, it is evident that generic solutions are no ultimate fit for many specific application scenarios. To be able to create custom-tailored digital tools, one of the central issues is to digitally represent the text, as well as its many contexts and related objects of interest in an adequate manner.

This thesis introduces a novel form of text representation that is based on Property Graph databases – an emerging technology that is used to store and query highly interconnected data sets. Based on this modeling paradigm, a new text research system called ”Kadmos” is introduced. It provides user-definable asynchronous web services and is built to allow for a flexible extension of the data model and system functionality within a prototype-driven development process. With Kadmos it is possible to easily scale up to text collections containing hundreds of millions of words on a single device and even further when using a machine cluster. It is shown how various methods of Text Mining can be implemented with and adapted for the graph representation at a very fine granularity level, allowing the creation of fitting digital tools for different aspects of scholarly work. In extended usage scenarios it is demonstrated how the graph-based modeling of domain data can be beneficial even in research scenarios that go beyond a purely text-based study.

Graphdatenbanken für die textorientierten e-Humanities

Dissertation im Fachgebiet Informatik von Thomas Efer, M. Sc.

Graphdatenbanken · Datenmodellierung · Recherchesysteme · e-Humanities
Text Mining · Korpusexploration · Information Retrieval

Abstract *(Deutsche Version)*

Vor dem Hintergrund zahlreicher Digitalisierungsinitiativen befinden sich weite Teile der Geistes- und Sozialwissenschaften derzeit in einer Transition hin zur großflächigen Anwendung digitaler Methoden. Zwischen den Fachdisziplinen und der Informatik zeigen sich große Differenzen in der Methodik und bei der gemeinsamen Kommunikation. Diese durch interdisziplinäre Projektarbeit zu überbrücken, ist das zentrale Anliegen der sogenannten „e-Humanities“. Da Text der häufigste Untersuchungsgegenstand in diesem Feld ist, wurden bereits viele Verfahren des Text Mining auf Problemstellungen der Fächer angepasst und angewendet. Während sich langsam generelle Arbeitsabläufe und Best Practices etablieren, zeigt sich, dass generische Lösungen für spezifische Teilprobleme oftmals nicht geeignet sind. Um für diese Anwendungsfälle maßgeschneiderte digitale Werkzeuge erstellen zu können, ist eines der Kernprobleme die adäquate digitale Repräsentation von Text sowie seinen vielen Kontexten und Bezügen.

In dieser Arbeit wird eine neue Form der Textrepräsentation vorgestellt, die auf Property-Graph-Datenbanken beruht – einer aktuellen Technologie für die Speicherung und Abfrage hochverknüpfter Daten. Darauf aufbauend wird das Textrecherchesystem „Kadmos“ vorgestellt, mit welchem nutzerdefinierte asynchrone Webservices erstellt werden können. Es bietet flexible Möglichkeiten zur Erweiterung des Datenmodells und der Programmfunktionalität und kann Textsammlungen mit mehreren hundert Millionen Wörtern auf einzelnen Rechnern und weitaus größere in Rechnerclustern speichern. Es wird gezeigt, wie verschiedene Text-Mining-Verfahren über diese Graphrepräsentation realisiert und an sie angepasst werden können. Die feine Granularität der Zugriffsebene erlaubt die Erstellung passender Werkzeuge für spezifische fachwissenschaftliche Anwendungen. Zusätzlich wird demonstriert, wie die graphbasierte Modellierung auch über die rein textorientierte Forschung hinaus gewinnbringend eingesetzt werden kann.