ViSciPub

Visual Analysis of Scientific Publications

Programm / Ausschreibung	IKT der Zukunft, IKT der Zukunft, IKT der Zukunft China Ausschreibung 2017	**Status**	abgeschlossen
Projektstart	01.04.2018	**Projektende**	31.05.2021
Zeitraum	2018 - 2021	**Projektlaufzeit**	38 Monate
Keywords	Visual Analytics; Document Analysis; Text Visualization; Human-centered Design;		

Projektbeschreibung

Wir leben in einer Welt mit zunehmender Informationsdichte. Ein Großteil dieser Information kommt in Form von Textdaten wie etwa Nachrichten, Blogartikel oder wissenschaftliche Publikationen. Das Problem ist, dass Menschen diese Datenmengen ohne fundierte Analyse- und Kommunikationstools nicht mehr bewältigen und überschauen können. Dies birgt fatale Konsequenzen, da eine kontextbasierte Übersicht bereits für „kleine“ Themenbereiche zu einer massiven Herausforderung wird. Wichtige Information wird so übersehen, relevante Verknüpfungen bleiben verborgen, Gewichtungen im Thema werden falsch wahrgenommen.
Unser Projekt zielt auf die visuelle Analyse und Kommunikation großer Korpora wissenschaftlicher Literatur ab. Vorausgehende Arbeiten haben gezeigt, dass dies ein vielversprechender Ansatz ist um obige Herausforderungen zu bewältigen. Unser Ziel ist es die dafür notwendige Technologie einen Schritt weiter zu treiben, indem neue und innovative Konzepte holistisch erforscht und weiterentwickelt werden.
Zuerst werden wir neue Methoden für die Datenreinigung und -aufbereitung untersuchen. Hier wird der Fokus auf Schlagworten, Abstracts und Volltexten (d.h. unstrukturierte Daten) liegen, Daten deren Bereinigung in der Vergangenheit eher stiefmütterlich behandelt wurden.
Danach wird der Fokus auf neue Visualisierungsansätze für die Analyse großer Korpora wissensch. Publikationen gelegt werden. Dabei gehen wir über den State of the art visueller Dokumentenanalyse hinaus, indem wir ein breites Feld an Methoden zur Analyse (aus Statistik und Computeranalyse) einsetzen und den UserInnen eine strukturierte Möglichkeit geben, um anstelle eines einzigen, etliche Analyseverfahren anzuwenden. Hier kann das Konsortium seine reichhaltige Erfahrung mit solchen Ansätzen ausspielen.
Abschließend werden wir neuartige Formen der visuellen Kommunikation von Zusammenfassungen und Trends aus solchen Datensätzen, abzielend auf ein breites Publikum, erforschen. Etwa mit dem neuen Konzept, dass wir als „storytelling Wordles“ bezeichnen. Unsere Forschung wird dabei zwei Richtungen verfolgen. Einerseits werden wir effizientere Werkzeuge für Analysten und Designer entwickeln, und andererseits End-UserInnen beforschen, z.B. Eye-tracking Studien mit Adressaten solcher Repräsentationen. Das Ziel ist in beiden Fällen sowohl die statischen, als auch die interaktiven Formen der visuellen Kommunikation wissensch. Publikationskorpora zu verbessern.
Die Qualität der Methoden wird mittels zweier Case Studies erfasst. In der Ersten wird die Literatur aus dem Bereich der

Datenvisualisierung und angrenzender Disziplinen (IEEE VIS, ACM CHI, ACM KDD, ACM SIGGRAPH) bereinigt, analysiert und kommuniziert. Hier geht es darum einen kontextuellen Vergleich zwischen überlappenden Communities anzustellen. In der zweiten Case Study wird Literatur aus dem Bereich der mathematischen Modellbildung und Simulation evaluiert. Hier ist der Anspruch, ein sehr heterogenes und verstreutes Forschungsgebiet zu erfassen.
Um diese Ideen umsetzen zu können wurde ein schlagkräftiges Konsortium aus Wissenschaft und Industrie aus China und Österreich gebildet. Für den Industriepartner bedeutet die Arbeit an diesem brandaktuellen Thema einen gewaltigen Benefit; die Österreichisch-Chinesische Kooperation soll der Auftakt zu einer anhaltenden Zusammenarbeit zwischen den beiden Ländern in diesem Forschungsbereich darstellen.

Abstract

We live in a world of ever-growing information load. Much of this data comes as text, such as blog posts, newspaper articles, or scientific publications. Without proper analysis and communication tools, however, individuals are often not able to parse all the relevant information anymore -which has severe consequences. Getting an objective and exhaustive, yet context-driven overview over even small topics is getting a huge practical challenge. Important pieces of information might be missed, important connections might stay hidden, or the relative importance of different pieces of the topic might be completely misinterpreted.
In our work, we will focus on the visual analysis and communication of large bodies of scientific literature. Past research has shown that this is a very promising approach towards overcoming the above listed challenges. Here, we aim to take this crucial technology one step further by jointly investigating new and innovative concepts to further enrich it.
First, we will investigate novel workflows and methods for cleaning publication data for its further computational and visual analysis. Cleaning data has been shown as a huge practical issue. While there are some good examples now for cleaning structured data, cleaning unstructured data in general, and data from scientific publications in particular, has gained little attention so far. Yet, our past experience from related projects on keyword analysis has shown that this data is specifically prone to such issues.
Second, we will investigate novel ways of visually analyzing large bodies of scientific publication data, through the lens of their keywords, abstracts, and full-texts. We will go beyond state of the art of visual document analysis in that we will open doors to a boarder set of statistical and computational analysis approaches, as well as allow the user to go from using a single analysis model only, to a structured, yet usable way of leveraging multiple models through visual interfaces. The consortium will leverage their cutting-edge experience in such approaches (visual parameter space analysis).
Finally, we will create and investigate novel ways of visually communicating summaries and trends in such data to a larger audience, for instance, with a novel concept that we call “storytelling Wordles”. Here our research will include two main streams. On the one hand, supporting analysts and designers to create such communication tools easier and more effectively. Currently, there are severe problems and hurdles, for instance, when trying to manually edit such visual representations, a typical yet not well-supported process. On the other hand, we will study end-users, i.e. the recipients of such representations; for instance, with eye tracking studies. The goal is to learn and further improve both static and interactive representations for visual communication of large bodies of scientific literature.
The value of the methods will be illustrated and evaluated with two case studies. First, we will clean, analyze, and communicate literature from the area of data visualization (IEEE VIS) and related areas (ACM CHI, ACM KDD, ACM SIGGRAPH). The idea is to show how our new methods better support not only gaining insight and summarizing a single community, but also allow for contextualizing and comparing it to other, related areas. The second case study will be on scientific literature on mathematical simulation and modeling, which is particularly interesting as it is a scattered and

inhomogeneous community.
To pull off these ideas, we created a strong consortium with partners from industry and academia, from China and Austria. Working on this very timely topic will give the company partner a tremendous market benefit, while our Austrian-Chinese setup is meant to establish and foster a lasting collaboration between the two countries in the proposed area.

Projektkoordinator

- Technische Universität Wien

Projektpartner

- dwh GmbH