Selected Ph.D. Thesis Abstracts

This Ph.D thesis abstracts section presents theses defended in 2020 and 2021. These submissions cover a range of research topics and themes under intelligent informatics, such as proximity mining, decentralized community information systems, deep anomaly detection,and automated reasoning and cybersecurity.

DATASET PROXIMITY MINING FOR SUPPORTING SCHEMA MATCHING AND DATA LAKE GOVERNANCE

Ayman Alserafi
alserafi@essi.upc.edu
Universitat Politcnica de Catalunya (UPC), Spain, and
Universit Libre de Bruxelles (ULB), Belgium
https://hdl.handle.net/2117/345323

With the huge growth in the amount of data generated by information systems, it is common practice today to store datasets in their raw formats (i.e., without any data preprocessing or transformations) in large-scale data repositories called Data Lakes (DLs). Such repositories store datasets from heterogeneous subject-areas (covering many business topics) and with many different schemata. Therefore, it is a challenge for data scientists using the DL for data analysis to find relevant datasets for their analysis tasks without any support or data governance. The goal is to be able to extract metadata and information about datasets stored in the DL to support the data scientist in finding relevant sources. This shapes the main goal of this thesis, where we explore different techniques of data profiling, holistic schema matching and analysis recommendation to support the data scientist.

We propose a novel framework based on supervised machine learning to automatically extract metadata describing datasets, including computation of their similarities and data overlaps using holistic schema matching techniques. We use the extracted relationships between datasets in automatically categorizing them to support the data scientist in finding relevant datasets with intersection between their data. This is done via a novel metadata-driven technique called proximity mining which consumes the extracted metadata via automated data mining algorithms in order to detect related datasets and to propose relevant categories for them. We focus on flat (tabular) datasets organised as rows of data instances and columns of attributes describing the instances.

Our proposed framework uses the following four main techniques: (1) Instance-based schema matching for detecting relevant data items between heterogeneous datasets, (2) Dataset level metadata extraction and proximity mining for detecting related datasets, (3) Attribute level metadata extraction and prox- imity mining for detecting related datasets, and finally, (4) Automatic dataset categorization via supervised k-Nearest Neighbour (kNN) techniques. We implement our proposed algorithms via a prototype that shows the feasibility of this framework. We apply the prototype in an experiment on a real-world DL scenario to prove the feasibility, effectiveness and efficiency of our approach, whereby we were able to achieve high recall rates and efficiency gains while improving the computational space and time consumption by two orders of magnitude via our proposed early-pruning and pre-filtering techniques in comparison to classical instance-based schema matching techniques. This proves the effectiveness of our proposed automatic methods in the early-pruning and pre-filtering tasks for holistic schema matching and the automatic dataset categorisation, while also demonstrating improvements over human-based data analysis for the same tasks.

Scaffolding Decentralized Community Information Systems for Lifelong Learning Communities

Peter deLange
lange@dbis.rwth-aachen.de
RWTH Aachen University, Germany

Initially, the Web was developed as a decentralized system of information repositories that facilitate organizational knowledge transfer by allowing anyone to create and access content. However, Web publishing required both technical expertise and hardware infrastructure. With the rise of the Web 2.0, social networking sites and content management systems enabled all users to create Web content. But it simultaneously put the users at the mercy of the platform operators. Services could be shut down, erasing content and disrupting communities.

Decentralized community information systems radically change this dynamic by establishing participants as equal peers, which form a self-governing community. This way, a community regains control over their data, while being able to scale the infrastructure according to their needs.

In this dissertation, we followed a design science approach that provides support for communities to create and host their own, decentralized community information systems. On the one hand, we produced several artifacts to provide possible answers to the question of what properties such an infrastructure needs to fulfill. With the blockchain- based decentralized service registry, we propose a solution for making community knowledge accessible in a secure and verifiable way. On the other hand, we transfer the metaphor of educational scaffolding to the domain of service development. It is based on the idea, that a scaffold serves as a temporary supporting structure during a building's construction phase. As the construction site develops and the building gets completed, the scaffold gradually gets removed up to the point, that it is not needed anymore. With the community application editor, communities are provided with such a scaffolding environment for requirements elicitation, wireframing, modeling and coding their decentralized community applications. Once deployed on the infrastructure, those applications and development efforts remain available, even after the contributing members might have left, serving as the community's long term memory.

We demonstrated and evaluated our artifacts on a large European scale, with three longitudinal studies conducted within several communities from different areas of technology enhanced learning, such as the European voluntary service, vocational and educational training providers and in higher education mentoring scenarios. All in all, this shift from data being stored in centralized repositories to a decentralized infrastructure, hosted by community members, opens up possibilities for a more democratic and egalitarian management of community knowledge.

DEEP ANOMALY DETECTION IN DISTRIBUTED SOFTWARE SYSTEMS

Sasho Nedelkoski
nedelkoski@tu-berlin.de
Distributed and Operating Systems, der Technischen
Universität Berlin, Berlin, Germany

ARTIFICIAL Intelligence for IT Operations (AIOps) combines big data and machine learning to replace a broad range of IT Operations tasks. The task of anomaly detection has a prominent position in ensuring the required reliability and safe operation in distributed software systems. However, the frequent software and hardware updates, system heterogeneity, and massive amount of data create a challenging environment. The detection of anomalies in these systems predominantly relies on metric, log, and trace data. Each of them provides a different view of the internal states of the systems. By induction, improving the detection in every data source increases the overall anomaly detection performance in the system.

This thesis provides the following contributions. (1) We present a method based on variational inference and recurrent neural network to address the detection of anomalies in system metric data that possibly exhibit multiple modes of normal operation. (2) We propose a novel log parsing through language modelling that enables learning of log representations for downstream anomaly detection. We identify the learning of log representations as a major challenge toward a robust anomaly detection. Therefore, we additionally design a method that learns log representations by distinguishing between normal data from the system of interest and easily accessible anomaly samples obtained through the internet. (3) We describe a self-supervised anomaly detection task that utilizes the entire trace information to robustly detect anomalies that propagate through system components. (4) In a rule-based approach, we combine the presented methods for a multi-view anomaly detection.

The methods presented in this thesis were implemented in prototypes and evaluated on various datasets including production data from a cloud provider. They provided (1) an F1 score of 0.85 on metric data, (2) parsing accuracy of 99% and F1 score improvement of 0.25 in log anomaly detection, (3) increase in F1 score of 7% in trace anomaly detection over the state of the art, and (4) broadened spectrum of detected anomalies. The results were peer-reviewed and published at renowned international conferences.

HYPOTHESIS GENERATION VIA AUTOMATED REASONING WITH APPLICATIONS TO CYBERSECURITY

Jose N. Paredes
jose.paredes@cs.uns.edu.ar
Universidad Nacional del Sur (UNS), Bahia Blanca,
Argentina

IN recent years, a wide variety of malicious behaviors have taken root in social platforms such as fake news, hate speech, malware diffusion, among others. This kind of behavior leads to a set of related problems, which has produced unforeseen consequences in many arenas; motivated by this situation, this thesis focuses on the study of automated hypothesis generation systems to address such issues. As a first contribution, two basic approaches are considered for the detection of a specific type of malicious behavior that we call adversarial deduplication. In the first approach, the generation of hypotheses is based on the use of well-defined logical rules, though the essence of its operation is supported by results obtained from applying previously-deployed machine learning techniques (that is, a part of the symbols necessary for the logical machinery are yielded by ML tools). In the second approach, ML techniques are used with a more central role; specifically, we use classifiers to tackle the problem, and hypothesis generation is carried out by simpler rules that are fired when the output of the classifiers exceeds a certain threshold.

Given that the initial proposal focuses on a specific problem (and therefore suffers some of the same limitations as ad-hoc approaches in the literature), our ultimate aim is to develop more robust and general systems. In particular, it is crucial to be able to handle situations not only single problems, but rather consider their multiplicity and take advantage of the relationships that may exist between them. In order to make progress in this direction, the main contribution is the presentation of the NETDER (Network Diffusion and Existential Rules) architecture to reason about malicious behavior on social platforms, which, in principle, seeks to serve as a guide for the implementation of software tools in such domains. NETDER includes four main modules: Data Ingestion (handles issues such as data cleaning, inconsistency, data analytics, among others, as well as other higher-level issues such as trust and uncertainty management); Ontological Reasoning (manages the knowledge base, both for background knowledge as well as for the network, and provides inference services); Network Diffusion (handles the evolution of the network in the form of diffusion processes, and checks conditions for the Ontological Reasoning Module); and Query Answering (handles the coordination of the two previous modules in order to answer the specific queries that users issue to the system).

After presenting the architecture, we study of the computational cost of query answering in its different instantiations, given that this process fundamentally drives the generation of hypotheses. Our analysis yields an interesting set of results that range from polynomial time tractability to the possibility that termination is not guaranteed, depending on the features that are made available to the model. Additionally, we develop a use case to illustrate how the approach can be applied in a cybersecurity domain to reason about products that are at risk of attack based on Darknet forum posts.

The final contribution is an experimental evaluation of the NETDER architecture. Given the difficulty of obtaining adequate datasets with ground truth (which is necessary to carry out performance evaluations), it was necessary to develop a general testbed designed with the purpose of generating social networks with complete traces of posting activities, potentially involving all kinds of malicious content such as fake news, malicious actors, botnets, links to malware, hate speech, etc. Our results constitute an important step towards achieving the ultimate goal, which is to develop automated robust hypothesis generation systems that can be used to address malicious behavior in social platforms.