DBKDA 2023

The Fifteenth International Conference on Advances in Databases, Knowledge, and Data Applications

ISBN: 978-1-68558-056-8

March 13th - 17th, 2023

Barcelona, Spain

DBKDA 2023 Editors

Malcolm Crowe, Emeritus, University of the West of Scotland, UK

Fritz Laux, Reutlingen University, Germany

DBKDA 2023

Foreword

The Fifteenth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2023), held between March 13 – 17, 2023, continued a series of international events covering a large spectrum of topics related to advances in fundamentals on databases, evolution of relation between databases and other domains, data base technologies and content processing, as well as specifics in applications domains databases.

Advances in different technologies and domains related to databases triggered substantial improvements for content processing, information indexing, and data, process and knowledge mining. The push came from Web services, artificial intelligence, and agent technologies, as well as from the generalization of the XML adoption.

High-speed communications and computations, large storage capacities, and load-balancing for distributed databases access allow new approaches for content processing with incomplete patterns, advanced ranking algorithms and advanced indexing methods.

Evolution on e-business, ehealth and telemedicine, bioinformatics, finance and marketing, geographical positioning systems put pressure on database communities to push the 'de facto' methods to support new requirements in terms of scalability, privacy, performance, indexing, and heterogeneity of both content and technology.

We take here the opportunity to warmly thank all the members of the DBKDA 2023 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to DBKDA 2023. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the DBKDA 2023 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that DBKDA 2023 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the fields of databases, knowledge and data applications.

We are convinced that the participants found the event useful and communications very open. We also hope that Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city

DBKDA 2023 Chairs:

DBKDA 2023 Steering Committee
Fritz Laux, Reutlingen University, Germany
Andreas Schmidt, Karlsruhe Institute of Technology / University of Applied Sciences
Erik Hoel, Esri, USA
Lisa Ehrlinger, Software Competence Center Hagenberg GmbH, Austria
Peter Kieseberg, St. Pölten University of Applied Sciences, Austria

DBKDA 2023 Publicity Chairs
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

DBKDA 2023

Committee

DBKDA 2023 Steering Committee
Fritz Laux, Reutlingen University, Germany
Andreas Schmidt, Karlsruhe Institute of Technology / University of Applied Sciences
Erik Hoel, Esri, USA
Lisa Ehrlinger, Software Competence Center Hagenberg GmbH, Austria
Peter Kieseberg, St. Pölten University of Applied Sciences, Austria

DBKDA 2023 Publicity Chairs
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain

DBKDA 2023 Technical Program Committee

Taher Omran Ahmed, College of Applied Sciences, Ibri, Sultanate of Oman / Azzentan University, Libya
Julien Aligon, Institut de Recherche en Informatique de Toulouse (IRIT) | Université Toulouse 1 Capitole, France
Alaa Alomoush, University Malaysia Pahang, Malaysia
Emmanuel Andres, Hôpitaux Universitaires de Strasbourg, France
Zeyar Aung, Masdar Institute of Science and Technology, UAE
Gilbert Babin, HEC Montréal, Canada
Aruna Bansal, Indian Institute of Technology (IIT) Delhi, India
Jam Jahanzeb Khan Behan, Université libre de Bruxelles (ULB), Belgium / Universidad Politécnica de Cataluña (UPC), Spain
Flavio Bertini, University of Parma, Italy
Ali Boukehila, University of Annaba, Algeria
Zouhaier Brahmia, University of Sfax, Tunisia
Martine Cadot, LORIA, Nancy, France
Alessandro Castelnovo, Intesa Sanpaolo S.P.A / University of Milano Bicocca, Italy
Sanjay Chaudhary, Ahmedabad University, India
Yung Chang Chi, National Cheng Kung University, Taiwan
Jong Choi, Oak Ridge National Laboratory, USA
Miguel Couceiro, LORIA, France
Malcolm Crowe, University of the West of Scotland, UK
Monica De Martino, Istituto per la Matematica Applicata e Tecnologie Informatiche "Enrico Magenes" | Consiglio Nazionale delle Ricerche, Italy
Marianna Di Gregorio, University of Salerno, Italy
Anton Dignös, Free University of Bozen-Bolzano, Italy
Ivanna Dronyuk, Lviv Polytechnic National University, Ukraine
Cedric du Mouza, CNAM (Conservatoire National des Arts et Métiers), Paris, France
Lisa Ehrlinger, Software Competence Center Hagenberg GmbH, Austria
Amir Hajjam El Hassani, University of Technology of Belfort Montbeliard, France

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Contact Tracing Applications Under the European Regime

Raif Baran Tombul,
PhD Student at Universitat Autonoma de Barcelona
Barcelona, Spain
rbtombul@gmail.com

Abstract- Covid-19 pandemic obliged scholars to scrutinize new privacy concerns due to the use of digital contact tracing applications. Considering that we are living in the digital age, the type of privacy safeguards that data controllers need to take should be thoroughly investigated. Although the main goal of these applications is to tackle the spread of the pandemic in society, the privacy rights of users must also be preserved. Otherwise, serious privacy risks might appear when the pandemic is eventually over. This paper aims to contribute to this discussion by addressing the potential questions related to the privacy risks of contact tracing applications from technical and organizational measures perspectives and thus to provide a contribution to the use of privacy-preserving contact tracing applications within the European Economic Area (EEA).

Keywords- Privacy; General Data Protection Regulation; Law; Pandemic; Contact Tracing.

I. INTRODUCTION

There are many samples in the history of medicine, ranging from AIDS to Ebola, where tracing methods were conducted to determine symptomatic people and, where required, employ isolation strategies [1]. Traditional contact tracing, where a public health official interviews an infected person to determine the places and people they met, is still in place [2]. Contact tracing, identifying individuals that have been in contact with an infected person, is a key component in tackling the spread of infectious illnesses [3]. The tasks conducted by contact tracing applications could be accumulated into 3 groups: detection of contact events (proximity tests), transmission, and exposure notification [4]. Accordingly, contact tracing applications have played an important role in controlling the spread of Covid in society. However, there are some privacy concerns among users about the use of these applications, which will be reviewed in this paper. Accordingly, the European Data Protection Board (EDPB) published a guideline about contact tracing applications [5]. Additionally, the European Commission published a communication about contact tracing applications [6] to establish certain points to consider for data controllers during their use of these applications in addition to the General Data Protection Regulation (GDPR)[7]. This idea paper briefly addresses the privacy concerns stemming from the use of contact tracing applications within the EEA and mentions the importance of privacy safeguards that could play an important role in mitigating these concerns. Accordingly, in Section 2, concerns and risks about contact tracing applications will be addressed. Subsequently, in Section 3 privacy implications of the applications' architectural choice applications will be briefly analyzed. Finally, in Section 4, technical and organizational measures will be elaborated, and potential solutions will be evaluated.

II. CONCERNS AND RISKS ABOUT CONTACT TRACING APPLICATIONS

The increased use of the Internet, together with rapid advances in technology, has changed the way in which information about users is gathered, stored, and exchanged was detailed [8]. Having said that, in order to fight with pandemic efficiently, individuals should trust the privacy features of the applications, thereby downloading these applications to their mobile phones. However, mobile applications possess, as seen, both certain advantages and ambiguous aspects [9]. Applications for contact tracing can be broadly divided into two categories [10]. In the centralized system, public institutions gather data on a single server, where data matching takes place [11]. The unique codes generated by a contact event are stored on each person's device in the decentralized approach instead of being sent to a centralized server [12]. While the centralized approach assumes that individual user data which could be leaked through the application is the most notable risk, the decentralized approach assumes that the compromising of all the user data in one location is the largest risk [13]. Therefore, it is plausible to say that each method is subject to a certain amount of risks. Generally, there are two types of privacy risks to an individual when we consider exposure notification applications, these are namely identity privacy, in which situation user individuals would not desire their identity to be shared without their affirmation) and location privacy, which response to the case where the individual would not desire other people may be able to link the various locations they visited to discover location history, without their consent) [14]. Hence, citizens who live in the community and download contact tracing applications to their mobile phones due to the Covid pandemic are concerned about being tracked by data controllers that process this personal data processed via the Global Positioning System (GPS). Tracking patients with Covid-19 and activities of contact persons could cause a breach of their privacy [15]. Furthermore, processing location data has further consequences because it would enable businesses to collect the data to learn about the movements of individuals and draw conclusions about preferences and habits [16]. Although contact tracing systems do not explicitly collect or record the true identities of individual users, movement profiles based on pseudonymous tracing data make it possible to identify a large fraction of users with a high probability [17].

In summary, although there are plenty of advantages generated by contact tracing applications, there are also a few vulnerabilities in terms of privacy aspects thereof. In the

 ISBN: 978-1-68558-056-8

following sections, this paper addresses these concerns by mentioning the safeguards that could be used.

III. ARCHITECTURE OF THE APPLICATIONS AND PRIVACY IMPLICATIONS THEREOF

Processing activities with centralized or decentralized protocols do have several implications for data controllers and data subjects. There is a need to understand the logic of decentralized and centralized processing. To track infected people and alert those who have come into touch with them, the centralized approach entrusts a central server with user information [18]. In contrast, the decentralized strategy relies on users' phones to keep user data and alert them, in case they are exposed to an infectious person[19]. Either choice of architecture brings advantages as well as disadvantages in terms of privacy, as already discussed in the relevant literature. However, more privacy-preserving technologies are required to mitigate the aforementioned risks rather than centralized or decentralized protocol discussion. For instance, many experts favored Bluetooth technology to prevent any sort of location-tracking-related risk. Similarly, the EPDB is in favor of the idea that the priority should be to process it without collecting localization data via Bluetooth [20]. These secure means of tracking are in line with the privacy-preserving perspective.

IV. TECHNICAL AND ORGANISATIONAL MEASURES

The EDPB recommended the adoption of both centralized and decentralized systems, provided that adequate security measures are implemented [21]. This perspective brought by the EDPB is quite useful for grasping the significance of adequate security measures implemented by data controllers. Also, as mentioned by the Commission, in general, the degree of security should match the amount and sensitivity of personal data processed [22]. Therefore, in order to control privacy and data protection risks and manage ethical concerns, this necessitates taking into consideration and combining the most efficient legal, organizational, and technical safeguards, including cutting-edge statistical and computational measures [23]. Accordingly, as per the EDPB Guidance, modern cryptographic techniques must be used to protect the data that is stored on servers and in applications, communications between the remote server and the apps [24]. EDPB also mentions the requirement of mutual authentication between servers and applications required [25]. These measures are feasible, as they have already been used for different types of digital applications by data controllers for years. However, considering the evolving nature of privacy threats, in addition to technical and organizational measures set out under article 32-1 of the GDPR and proposed by the EDPB, some tailor-made options could solidify the quality of these measures. For instance, blockchain technology, which is an open and shared database, over which no single party has control, and transactions, which include messages exchanged when two devices come into close contact, are safely recorded in blocks [26], could be useful for digital contact tracing, as proposed by Klaine and colleagues. As they mentioned, due to the fact that blockchain does not rely on a central server, this can enable global access to information while simultaneously being more resistant to harmful attacks [27]. Hence, considering that blockchain is now being used in keeping health records of patients in preserving their overall medical history without any involvement of service providers [28], it is also possible to generate a privacy-preserving, and feasible solution by implementing blockchain measures for the European contact tracing applications.

More of a generic solution to mitigate other unexpected privacy-related threats not listed in section II, hiring subject matter experts specifically devoted to implementing technical and organizational measures and designating contractual safeguards with third-party suppliers or vendors within the scope of cyber security activities could enhance the security capabilities of data controllers. In particular, considering safeguards for third-party vendors involved in any process of contact tracing applications are of massive importance to provide oversight on activities of data processors in line with article 28 of the GDPR. To this end, due to it is prevalent use and cost-efficient nature in many other fields, standard contractual clauses between controller and processor introduced by the EU Commission [29] could be an efficient safeguard for stipulating the required tailor-made safeguards that processors must implement. By this, it would be possible to generate a feasible solution for the implementation of required technical and organizational measures by third-party data processors as well, in order to mitigate any potential risk related to the involvement of third parties.

Last but not least, detailed and recurring data protection impact assessments could be an efficient way to determine privacy-related risks, regardless of the architectural design of the applications. Privacy risks associated with data regarding identifiable individuals can be mitigated in great part by using de-identification techniques in conjunction with reidentification procedures [30]. In order to have more privacy-friendly applications for any future case scenarios, all these safeguards should keep being implemented from a privacy-by-design perspective. The principle of Privacy by Design supports the idea that privacy should be deemed as a first-class citizen in the technology design and ought to be intensely inserted, as described by Besik and Freytag [31].

As a positive sign of compliance with these requirements, almost each of the data controllers in the EEA pays attention to these aforementioned risks and technical and organizational safeguards, based on their privacy policies. For instance, as a few samples of many successful ones, the Estonian application [32] properly indicates the third-party companies involved in the process, while at the same time and the Lithuanian application [33], displays the details of permissions and features the application requires. Likewise, the Italian application shared on the documentation website many important aspects related to the security of the application, such as privacy-preserving analytics, security document, and design information with the users [34].

Therefore, it is plausible to state that designing contact tracing applications with security and privacy considerations based on the potential vulnerabilities described in section II is important to diminishing any potential risks posed to data subjects.

 ISBN: 978-1-68558-056-8

V. CONCLUSIONS

Implementation of efficient technical and organizational safeguards, as well as a privacy-by-design approach, are of key importance to the success of contact tracing applications. Therefore, if efficient safeguards are put in place by data controllers of contact tracing applications, the type of architecture of applications will not have a massive impact on the level of privacy protection by merely itself, as the main goal of these applications is to block the spread of the virus throughout society, rather than tracking people movement or processing an excessive amount of their personal data. Accordingly, as a positive sign of this perspective, almost each of the data controllers within the EEA acts responsibly to comply with the GDPR requirements and other relevant guidance. For the path forward, in case such tracking applications are required again, it is diligent to implement such necessary safeguards elaborated in section IV of this paper, in addition to the existing safeguards that are already put in place by data controllers, to maintain privacy-preserving technology.

ACKNOWLEDGEMENT

Raif Baran Tombul thanks Prof. Antoni Roig Batalla for his constant support during the research.

REFERENCES

[1] T. Scantamburlo, A. Cortés, P. Dewitte, et al. “Covid-19 and tracing methodologies: A lesson for the future society”, Health Technol., Vol. 11, pp. 1051–1061, p.1052, 2021

[2] Dig Watch Website https://dig.watch/trends/contact-tracing-apps retrieved: January 2023)

[3] A. Anglemyer, et al. “Digital contact tracing technologies in epidemics: a rapid review” Cochrane Database Syst Rev., Aug 18;8(8): CD013699, p.4, 2020, doi: 10.1002/14651858.CD013699. PMID: 33502000; PMCID: PMC8241885.

[4] J. C. Nobre, L. R. Soares, B. O. R. Huaytalla, E. D. S. Júnior, and L. Z. Granville ‘On the Privacy of National Contact Tracing COVID-19 Applications: The Coronav\'irus-SUS Case’ arXiv preprint arXiv:2108.00921. p.1. 2021

[5] The European Data Protection Board, Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak, adopted on 21 April 2020

[6] Communication from the Commission Guidance on Apps supporting the fight against COVID-19 pandemic in relation to data protection 2020/C 124 I/01 available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587141168991&uri=CELEX:52020XC0417(0) (retrieved: January 2023)

[7] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

[8] C. Paine, U.D. Reips., S. Stieger, A. Joinson, and T. Buchanan, “Internet users’ perceptions of ‘privacy concerns’ and ‘privacy actions” International Journal of Human-Computer Studies 65, no. 6, pp. 526-536, 2007

[9] Amnesty Web Site: https://www.amnesty.org/en/latest/news/2020/06/bahrain-kuwait-norway-contact-tracing-apps-danger-for-privacy/ (retrieved: January 2023)

[10] Friedrich Naumann Foundation Website, https://www.freiheit.org/turkey/safety-and-privacy-time-covid-19-contract-tracing-applications (retrieved: January 2023)

[11] Friedrich Naumann Foundation Website, https://www.freiheit.org/turkey/safety-and-privacy-time-covid-19-contract-tracing-applications (retrieved: January 2023)

[12] M. Shahroz, F. Ahmad, M.S. Younis., N. Ahmad, M.N.K. M. Boulos, R. Vinuesa, and J. Qadir “COVID-19 digital contact tracing applications and techniques: A review post initial deployments”. Transportation Engineering, 5, p.100072, 2021

[13] Duke TechPolicy Sanford Article 21 February 2021 comparing centralized and decentralized contact-tracing approaches https://sites.sanford.duke.edu/techpolicy/2021/02/21/centralizedvsdecentralized/ (retrieved: January 2023)

[14] E. Mbunge “Integrating emerging technologies into COVID-19 contact tracing: Opportunities, challenges and pitfalls.” Diabetes Metab Syndr., Nov-Dec;14(6), pp. 1631-1636, 2020, doi: 10.1016/j.dsx.2020.08.029. Epub 2020 Aug 26. PMID: 32892060; PMCID: PMC7833487

[15] R.A. Kleinman and C. Merkel “Digital contact tracing for COVID-19.” CMAJ. 2020 Jun 15;192(24), pp.E653-E656, p.E654, doi: 10.1503/cmaj.200922. Epub 2020 May 27. PMID: 32461324; PMCID: PMC7828844.

[16] R. Raskar, et al." Comparing manual contact tracing and digital contact advice." arXiv preprint arXiv:2008.07325, p.6, 2020

[17] L. Baumgärtner, A. Dmitrienko, B. Freisleben, A. Gruler, J. Höchst, J. Kühlberg and Mira Mezini et al. "Mind the gap: Security & privacy risks of contact tracing apps." In 2020 IEEE 19th international conference on trust, security, and privacy in computing and communications (TrustCom), pp. 458-467, p.461, 2020

[18] Duke TechPolicy Sanford Article 21 February 2021 comparing centralized and decentralized contact-tracing approaches https://sites.sanford.duke.edu/techpolicy/2021/02/21/centralizedvsdecentralized/ (retrieved: January 2023)

[19] Duke TechPolicy Sanford Article 21 February 2021 comparing centralized and decentralized contact-tracing approaches https://sites.sanford.duke.edu/techpolicy/2021/02/21/centralizedvsdecentralized/ (retrieved: January 2023)

[20] P. Chakraborty, M. Subhamoy, N. Mridul, and T. Suprita “Contact Tracing in Post-Covid World: A Cryptologic Approach.” Singapore: Springer, p.31, 2020

[21] European Data Protection Board, Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak, adopted on 21 April 2020, p.9

[22] Communication from the Commission Guidance on Apps supporting the fight against COVID 19 pandemic in relation to data protection 2020/C 124 I/01 available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1587141168991&uri=CELEX:52020XC0417(08) (retrieved: January 2023)

[23] European Data Protection Board, Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak, adopted on 21 April 2020, p.9

[24] European Data Protection Board, Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak, adopted on 21 April 2020, p.9

[25] U. Gasser, M. Ienca, J. Scheibner, J. Sleigh and E. Vayena “Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid.” Lancet Digit Health, 2020 Aug;2(8), pp. e425-e434, p.431, doi: 10.1016/S2589-7500(20)30137-0. Epub 2020 Jun 29. PMID: 32835200; PMCID: PMC7324107.

[26] V.P. Klaine, L. Zhang, B. Zhou, Y. Sun, H. Xu, and M. Imran, Privacy-preserving contact tracing and public risk assessment using blockchain for COVID-19 pandemic. IEEE Internet of Things Magazine, 3(3), pp. 58-63, p.58, 2020

[27] V.P. Klaine, L. Zhang, B. Zhou, Y. Sun, H. Xu, and M. Imran, Privacy-preserving contact tracing and public risk assessment using blockchain for COVID-19 pandemic. IEEE Internet of Things Magazine, 3(3), pp. 58-63, p.58, 2020

[28] B. Aslam, A. R. Javed, C. Chakraborty, J. Nebhen., S. Raqib. and M. Rizwan, Blockchain and ANFIS empowered IoMT application for privacy preserved contact tracing in COVID-19 pandemic. Personal and ubiquitous computing, pp.1-17, 2021

 ISBN: 978-1-68558-056-8

[29] EU Commission Website, Standard Contractual Clauses https://commission.europa.eu/publications/standard-contractual-clauses-controllers-and-processors-eueea_en (retrieved: January 2023)

[30] A. Cavoukian and J. Jonas “Privacy by design in the age of big data” Information and Privacy Commissioner of Ontario, Canada, p.8, 2012

[31] S. I. Besik and J. C. Freytag. "Managing Consent in Workflows under GDPR." In ZEUS, pp. 18-25, p.18, 2020

[32] HOIA Phone Application Privacy Policy https://koodivaramu.eesti.ee/tehik/hoia/app-web/-/blob/master/content/privacy.en.md (retrieved: September 2022)

[33] Korona Stop, Privacy Policy https://koronastop.lrv.lt/uploads/documents/files/corona-stop-app/Privatumo-politika-korona-stop-en.pdf (retrieved: September 2022)

[34] Immuni Application Documentation https://github.com/immuni-app/immuni-documentation#privacy (retrieved: September 2022)

 ISBN: 978-1-68558-056-8

A Practical Automated Transformation of Entity Relationship Models to Relational Models

Gregor Grambow
Aalen University
Aalen, Germany
Email: gregor.grambow@hs-aalen.de

Simon Ruttmann
Aalen University
Aalen, Germany
Email: simon.ruttmann@studmail.htw-aalen.de

***Abstract*—Creating conceptual database schemata via Entity Relationship (ER) diagrams is a prevalent way of modeling. However, these models are not directly compatible with the relational model used in SQL databases. Over the decades, various theoretic approaches for transforming ER models to relational models have been proposed. However, this did not lead to the creation of editors capable of such transformations. Modern editors either have no transformation facilities, or do not use the ER model proposed by Peter Chen but rather provide somehow enhanced database diagrams. Thus, conceptual schemata have to be transformed manually to technical ones, which is time consuming and error-prone. To counteract this, we propose a transformation from ER models to relational ones that focuses on practical applicability and operational semantics. Further, the approach enables the original ER model as well as prevalent extensions. To prove the applicability of this approach we have created a graphical editor capable of flexibly modeling ER diagrams and automatically transforming them to relational models.**

***Index Terms*—Entity Relationship; ER Model; Relational Model; Database; Editor**

I. Introduction

Entity Relationship (ER) modeling is a prevalent option for semantic data modeling primarily applied to database schemata. This way of modeling has been used since over 40 years. It was introduced in 1976 by Peter Chen [1] and has been the focus of active research for decades. There has been a myriad of extensions to the model, like the ECR model [2], the ECR+ model [3], HERM [4], or the EER model [5]. Many of these extensions offer valuable additions to the basic ER models. Some features have been adopted but many have also been discarded. As a result of this, basic ER modeling became prevalent but with a high number of different flavors. The usage of generalization concepts, cardinality constraints or the application of the n-ary relationships as proposed by Peter Chen differs in many applications.

However, to be usable in a relational database, ER models have to be converted to relational models. Manually executed, this process can be tedious and error-prone. Thus, various approaches for standardized transformations have been proposed, like [2] [6] - [12]. Most of these approaches have two major downsides: First, they often impose certain constraints on the ER models and second, they remain rather theoretic. In many cases, after presenting their approach, the authors recommended them to be used by practitioners or in automated tools. Despite having a clear formalism, a practical implementation was never achieved, often due to the lack of operational semantics. On the other hand, there is a high number of editor tools for ER models, contained in drawing tools [13] - [16], in client software of databases [17] [18], or in modeling tools like Enterprise Architect [19]. These editors have two main issues: Each of them uses a different subset of ER concepts, some are even closer to a relational or even UML class diagram editor. In addition, the transformation aspect has been ignored almost completely. Thus, semantic data modeling for databases usually involves first drawing an ER diagram and then manually transforming it to database tables, which can be a source of numerous issues.

To counteract this, we propose an approach for automatically transforming ER models to relational models. As opposed to prior approaches we do not focus on mathematical definitions or calculus but rather on practical applicability and operational semantics. Thereby, our approach can be easily applied to editors in different programming languages. To prove this applicability we have implemented a graphical ER editor that is capable of this automatic transformation.

The rest of this paper is organized as follows: Section II discusses related approaches, while Section III defines the concrete style of ER diagram that is assumed as basis for the transformation. Section IV provides an extensive description of the transformation approach. After that, Section V shows a practical evaluation of the proposed approach followed by Section VI providing a conclusion as well as future directions.

II. Related Work

In this section, we cover two types of related work: Scientific approaches presented for transforming ER models into relational models and practical ER editor tools.
In the scientific community, the case of transforming ER models has been extensively discussed over the decades. Most proposed approaches date back to the 1980s and 1990s. As mentioned, one big issue concerning general applicability is the high number of different ER variants. The basic variant proposed by Peter Chen [1] includes n-ary relationships and weak entities but no generalization concepts. Most of the extensions [2] - [5] to that model focused on adding structures for generalization. Due to that, many different transformation approaches take different variants of the ER model as basis.

 ISBN: 978-1-68558-056-8

Most attention was paid to the original version of the ER model [6] - [9] and extended ER models with generalization structures [2] [10] - [12]. All of them have in common that they remain rather theoretic and do not consider operational issues [20]. Further, to the best of our knowledge, none of these approaches lead to the development of a prevalent ER editor tool.

The fact that most research regarding that topic was carried out decades ago lets us assume, that the transformation approaches have been adopted and are now prevalent in ER editors. Therefore, the second part of this section deals with contemporary ER editors. However, before investigating the transformation capabilities, another issue has to be dealt with: There is a high number of ER editors that do not use the ER model as proposed by Peter Chen. Many of them focus on simplified binary relationships. In addition, they do not cover the most prevalent extension to Chen's model: generalization. Such diagrams are often nearer to a relational database diagram than to a real ER diagram. One category featuring such diagrams is drawing tools like Lucid Chart [13], Draw.io [14], or Visual Paradigm [15]. Some of them even contain concepts like stored procedures or triggers and can thus not be considered ER editors. Furthermore, none of them provides a transformation approach. Another category providing such diagrams is database client software, e.g., from PostgreSQL [17] or MySQL [18]. These editors provide visual modeling that can be directly used as database tables. However, the diagrams are rather close to the relational approach and not to Chen's model. The number of editors covering the latter is rather limited. One with a good set of concepts is the Enterprise Architect [19], which covers most elements considered prevalent in ER models as of today: n-ary relationships, generalization, and multi-valued or composite attributes. However, there is no transformation approach in place. Only one editor features an ER model like proposed by Chen and also a transformation approach: ERD+ [16]. But that editor only features a rather limited set of modeling elements. It does not support n-ary relationships and the use of generalization is rather restrictive. In addition, the translation of weak types is only possible on the most trivial level. This also applies to the translation of multi-valued and composite attributes. The combination of attribute types is not supported at all. In summary, it can be said that the tool can only be used for simple, not extensive ER Models.

As ER modeling stays relevant, there are also contemporary approaches dealing with this model. However, most of these approaches don't deal with transforming ER models to relational ones. Examples include approaches for creating ER models from text using natural language processing [21]-[23] or applying ER modeling for creating specific models, e.g., for ontologies [24], Kanban systems [25] or software structures [26]. A small number of approaches deals with the relational transformation, but they either provide only very basic transformations with no practical application [27] or no novel transformations at all [28].

All in all, it can be stated that the proposed transformation approaches did not make it into applicable tools, mostly due to the lack of coverage of operational issues. On the other hand, modern editors seem to focus on simplified binary variants that are closer to relational tables than to Chen's ER model.

III. Definition of the ER Model

As described in sections I and II, there is a number of variations of ER models. However, in order to develop algorithms for transforming the ER model to the relational model, it is mandatory that modeling capabilities are known. Due to this, the ER model used in this work will be defined to ensure an unambiguous transformation.

When defining the modeling capabilities, one goal is to provide the modeler as much freedom as possible. The ER model defined in this paper is heavily based on the model presented by Kemper and Eickler [29]. This model includes the most prevalent modeling components such as entities, n-ary relations and attributes. It also contains existence-dependent types and covers generalization in form of IsA-Structures. The model is additionally extended by the attribute types "Multi-valued attribute" and "Compound attribute" according to Vossen [30].

An entity is the most basic ER component, covered within this paper. It does not have a direct connection with another entity, must have at least one identifying attribute and may be referenced by any number of attributes.

The relation is derived analogously to the entities from the basic principles of the ER model. Within this work, n-ary relations are supported. Every relation must connect at least two entity types and may be referenced by any number of attributes. To increase the modeling capabilities, the ER model also allows reflexive relations from one entity to itself.

The third component commonly used in ER models are attributes. In the context of this work, a distinction is made between three types of attributes. Regular single-valued attributes, multi-valued attributes and identifying attributes. To extend the capabilities of the model, regular and multi-valued attributes can be referenced by further regular and multi-valued attributes. This means that each regular and multi-valued attribute can act as a composite of other attributes. The attributes that form a composite are called composite attributes. Attributes, which are part of a composite attribute are also implicitly given the possibility to act as an composite attribute, which allows the multiple application of composition and multi-valuedness. With the multiple application of multi-valuedness and composition, complex attribute structures can be formed. These structures can form cyclic dependencies and ambiguities between attributes when they reference each other directly or indirectly via further attributes. Nevertheless, it must be guaranteed that each attribute can be uniquely assigned to exactly one entity or relationship. Each attribute can be assigned unambiguously to one entity or relationship if there is a direct connection or exactly one path to an entity or relationship via further attributes. The uniqueness requirement implies that attribute structures must have the form of trees with the corresponding entity or relation as root.

 ISBN: 978-1-68558-056-8

The transformation presented in this paper also supports existence-dependent types, which are also referred to as weak types. Weak types depend on a parent type for identification. In the ER model, this dependency is modeled by the use of a relationship between the parent entity and the existence-dependent entity, called a weak relation. It should be noted that in this ER model, the parent does not have to be a strong entity. Therefore it is possible that a parent entity is also a weak entity, which in turn depends on another parent entity. Due to this multiple application of existence dependence, restrictions are required to uniquely assign a weak entity its dependent type. These are conceptionally analogous to the multiple application of the composition and multi-valuedness of attributes.

The ER model and transformation also incorporates generalization. Overall, generalization can be divided into a number of different types, characteristics, and constraints. In the context of this paper, generalization in the ER model is implemented exclusively by means of IsA structures, omitting the notation and transfer of specific properties of generalization. IsA structures associate multiple entities with each other. Each entity of the IsA structure acts as a subtype or supertype entity. An IsA structure references exactly one supertype entity and any number of subtype entities. The subtype entities of the IsA structure inherit all attributes of the supertype entity.

When defining the rules for IsA structures, a multiple-inheritance of the attributes of a supertype entity to a subtype entity must be excluded. To avoid this, a restriction to tree-structures, as with attributes, could be made but would be too restrictive. Instead the restriction is made based on the three following sets.

A) The ”subtype set” of an entity contains the entity itself and all entities, which inherit from the entity.
B) The ”supertype set” of an entity includes the entity itself and all other entities, which the entity inherits from.
C) The ”influenced type set” of an entity includes the subtype set and additionally for each entity in the subtype set the supertype set.

The sets defined above on the basis of an expression of several IsA structures are shown in Figure 1. The sets here start from the entity highlighted in blue. The entity within the area shown in red is part of the supertype set. The entities within the green area shown are part of the subtype set. The influenced type set contains all entities, which are highlighted in purple.

If an entity is connected as a subtype of an IsA structure the influenced type set of the entity must be disjoint with the supertype set of the supertype of the IsA structure. In contrast, when an entity is connected as the supertype of an IsA structure, the influenced type set of all subtypes of the IsA structure must be disjoint with the supertype set of the entity.

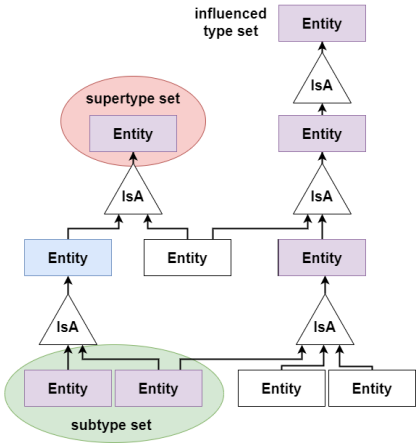

Fig. 1. Subtype-, supertype and influenced type set

In summary, the ER model discussed in this paper is extended by the following concepts:

- Multivalence of attributes
- Composition of attributes
- Reflexive, unary relationships
- N-ary relationships
- Existence dependency
- Generalization

IV. TRANSFORMATION

The implementable transformation of ER models is realized in several steps, which are executed sequentially.

A) Creating a data model structure of ER diagrams
B) Transformation of attributes
C) Transformation of IsA structures
D) Transformation of weak types
E) Transformation of relationships
F) Cascading of primary keys for attributes

These steps will be explained in more detail below.

A. Structural data model of ER models

In order to execute algorithmic approaches for translating the ER model into the relational model, a basic data model is required on which they can operate.

Any ER diagram essentially consists of elements such as entities, relationships, attributes, IsA structures, and associations between those elements. Therefore, the structure can be expressed directly as a graph. Furthermore, information about the cardinality between a relation and an entity can be stored in the edges of the graph. In case of a IsA structure, the edges also contain information about whether the connected node is a supertype or subtype.

In Section III, it was explained that attributes are always expressed in the form of trees. This makes it possible to further restrict the graph. Since all entities and relationships can have attributes, each of these elements acts as the root of a tree. Each attribute in the ER model is therefore represented as a node in the tree. The edges within the tree structure do not hold any additional information.

 ISBN: 978-1-68558-056-8

B. Transformation of attributes

The goal of this translation is to express any attribute structure by means of relations. From the previous subsection, it is known that all attributes are held within a tree. Therefore, only these trees have to be considered in this translation. If a tree consisting exclusively of single-valued attributes is considered, this can be solved trivially by creating a relation for the tree's root entity or relationship and adding each attribute, which is a leaf of the tree to that relation. Algorithmically, this can be done by traversing the tree in post-order and checking whether the current node is a leaf. However, if a tree contains multi-valued attributes, these cannot be added to the tree's root relation as a relation expresses a fixed sized schema and multi-valued attributes can take on any number of values. In this case, a standalone relation must be created for that attribute and a reference between it and the tree's root relation has to be created. In the case of composite attribute structures, attention must be paid between which relations a reference is created. It is important to note, that the referenced relation does not necessarily have to be the tree's root relation. As it is quite possible that the relation of a multi-valued attribute references a relation of another multi-valued attribute. The transformation of composite attribute structures can be executed by creating a relation for each attribute at the beginning. If the tree is then traversed in post-order, and the current node represents a multi-valued attribute a reference can be created between the current node's relation and the node's parent relation. For single-valued attributes, the relation can be merged with the relation of its parent node. Within the merge, all attributes and references of the child relation are transferred to the relation of the parent node. This procedure can be extended by skipping composite attributes, which consist of only one additional multi-valued attribute.

Listing 1 shows the algorithm for translating attributes in pseudo code. It is executed for each entity and each relation in the ER graph. The algorithm is explained below.

Listing 1. Transformation of attributes

```
1  Function TransformAttributeTree (Parent)
2      For Each Child in TreeNode //Execute post-order
3          TransformAttributeTree (Child)
4      End For
5      If TreeNode is Multivalued Attribute Then
6          TreeNode.Table <- marked
7      End If
8
9      If TreeNode is Leaf Then //Recursion resolution
10         Return
11     End If
12
13     For Each Child in TreeNode
14         If Child.Table is marked Then
15             If TreeNode.Children.Size = 1 Then
16                 //Handling of "forwarding" attributes
17                 MergeTable(TreeNode.Table, Child.Table)
18                 TreeNode.Table <- marked
19             Else
20               TreeNode.Table.References.Add(Child.Table)
21             End If
22         Else
23             MergeTable(TreeNode.Table, Child.Table)
24         End If
25     End For
26 End Function
27
28 Function MergeTable(ParentTable, ChildTable)
29     ParentTable.Columns.AddAll(ChildTable.Columns)
30     ParentTable.References.AddAll(ChildTable.References)
31     Delete ChildTable
32 End Function
```

The initial situation of the algorithm, shown in Listing 1, is that the ER graph has been created. In addition, a relation is created for each attribute. As within this algorithm, relations which correspond to a single-valued attribute are successively unified. The relations which were created for multi-valued attributes are preserved. For these relations only the above mentioned references are created. In the shown algorithm the postorder traversal takes place in the lines 2 to 4, as well as 9 to 11. For the handling of multi-valued attributes, these are marked in each call in lines 5 to 7. This takes place before the recursion resolution, in order to seize also multi-valued attributes, which are leaves of the attribute tree. Otherwise, those would not be marked and would be merged in the following lines. In lines 13 to 25, each direct child attribute is handled for an attribute. If it is a marked attribute, a reference to the child attribute is created in line 20. In the special case that the attribute has only one multi-valued attribute as a child, lines 15 to 18 are executed and a "skipping" takes place, regardless of whether the current attribute is a multi-valued or single-valued attribute. If, on the other hand, it is a single-value attribute (line 23), the relation of the child attribute can be resolved by merging it with the relation of the current attribute. The merging itself is done by adding all columns and references of the child table to the parent table (line 29 to 31).

Figure 2 illustrates this algorithmic process. The arrows indicate, how the algorithm will process the attributes. The cross next to the relations shows that these have been resolved. The resolution takes place in the same call as the arrows shown in the same color. The result is shown on the bottom right. Note that the attribute value D consists here of many attribute values F and one attribute value E. A direct reference between the entity relation and the relation F could not represent this situation. Also note that the attribute B is merged within the algorithm. This is permissible because entity I is associated with exactly one value for the attribute G.

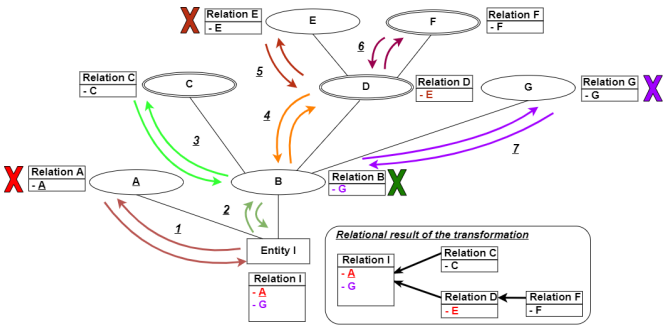

Fig. 2. Transformation of complex attribute structures

To complete the translation of the attributes into the relational model, the references between the created relations must

be mapped in the form of foreign key dependencies. For this purpose, the remaining relations can be traversed in pre-order and primary keys can be added to the relations. These also act as foreign keys to the primary keys of the referenced relation. The cascading of the foreign keys is not executed directly, as it is only ensured that the root entity or relation contains all primary keys due to its identifying attributes. However, as new primary keys may be added if the entity is a weak entity or part of an IsA structure the immediate execution could lead to invalid references between the relations. Therefore the execution of the cascading will take place at the end of the whole transformation process as Step F.

C. Transformation of IsA-Structures

The transformation of IsA structures is realized by means of foreign key dependencies between the subtypes and supertypes of the IsA structure. It should be noted that the relations of the entities must exist and the primary keys must be located in them. Because of this, the transformation of the attributes must take place before the transformation of IsA structures. Each subtype of an IsA structure inherits all primary keys of the supertype. In addition, these inherited primary keys refer to the upper type as foreign keys. If entities are part of several IsA structures, "higher level" IsA structures have to be translated first to ensure that entities at lower levels receive all primary keys. To illustrate the translation order Figure 3 shows an entity-relationship model on the left and the relational model on the right. The red highlighted IsA structure can only be translated after the blue and green IsA structures. The blue one, on the other hand, only after the green one above it.

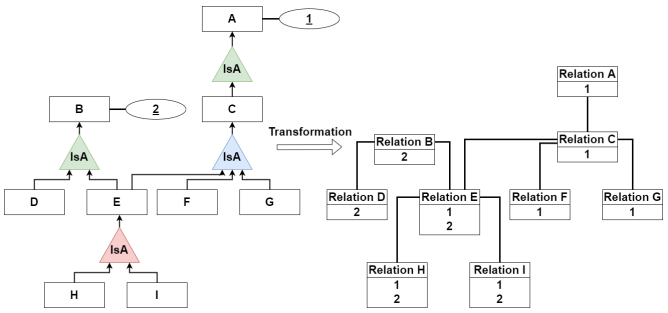

Fig. 3. Transformation order of IsA-Structures

Listing 2 shows the algorithm for transforming an IsA structure. Here, the processing order of the structures is maintained by traversing all IsA structures up to N times, where N is the number of IsA structures.

If the supertype of the selected IsA structure inherits from further IsA structures, and these have not yet been translated, the current pass is skipped (line 6 to 10). Specifically, line 6 determines all IsA structures that must already be transformed in order to transform the current IsA structure. In Figure 3, this corresponds to the green and blue highlighted IsA structure in the case of the currently treated red IsA structure. Following this, line 7 checks whether all have already been transformed. If at least one IsA structure has not been transformed, the current call is skipped. This also applies if the selected IsA structure has already been translated (line 2 to 4). If the IsA structure can be transformed in this call, the actual transformation takes place by creating the foreign key dependencies in line 12 to 16.

Listing 2. Transformation of IsA-Structures

```
1  Function TransformIsAStructure(IsAStruct)
2    If IsAStruct is transformed Then
3      Return
4    End If
5
6    UpperLayerIsAs <- GetInheritedIsAs(IsAStruct.SuperType)
7    UnhandledUpperLayerIsAs <- UpperLayerIsAs !transformed
8    If UnhandledUpperLayerIsAs not empty Then
9      Return
10   End If
11
12   For Each SubType in IsAStruct.Subtypes
13     For Each PrimaryKey in SuperType
14       AddForeignKeyAsPrimaryKey(Supertype, Subtype)
15     End For
16   End For
17   IsAStruct <- isTransformed
18 End Function
```

D. Transformation of weak types

For the translation of weak types to the relational model, it is a prerequisite that relations exist for all entities and relationships and that all attributes are already contained in them. Therefore, the transformation of attributes and IsA structures must be performed beforehand. IsA structures must be translated before, since a strong entity, on which a weak entity depends, can receive further primary keys during the translation of IsA-Structures.

Equivalent to the transformation of IsA-Structures, the translation order has to be considered. The translation has to start from weak entities, which have a connection to a strong entity or an already translated weak entity by means of a weak relationship. The relation of the weak relationship always has to be merged with the relation of the dependent entity. During the translation, the relation of the entity to be translated keeps a reference to the strong or already translated entity. This reference can then be used to create the foreign key dependencies.

The translation process of an ER diagram is shown in Figure 4. The figure starts immediately after the execution of the algorithms for the translation of attributes and IsA structures. The first rectangle shows an example ER model, which is transformed over a series of steps. The second step is the starting point of the algorithm, where all elements occur as a relation, conditioned by the previously executed attribute algorithm. The green highlighted elements represent the weak relationships in the ER model, which are required to be transformed. According to the mentioned translation order, the elements to be translated are determined and transformed in each step. In Figure 4, the blue elements are translated first, followed by the red elements.

Algorithmically, the merging of relations and reference creation from Figure 4 is shown in Listing 3. The compliance with the order is done, analogous to the translation of IsA

 ISBN: 978-1-68558-056-8

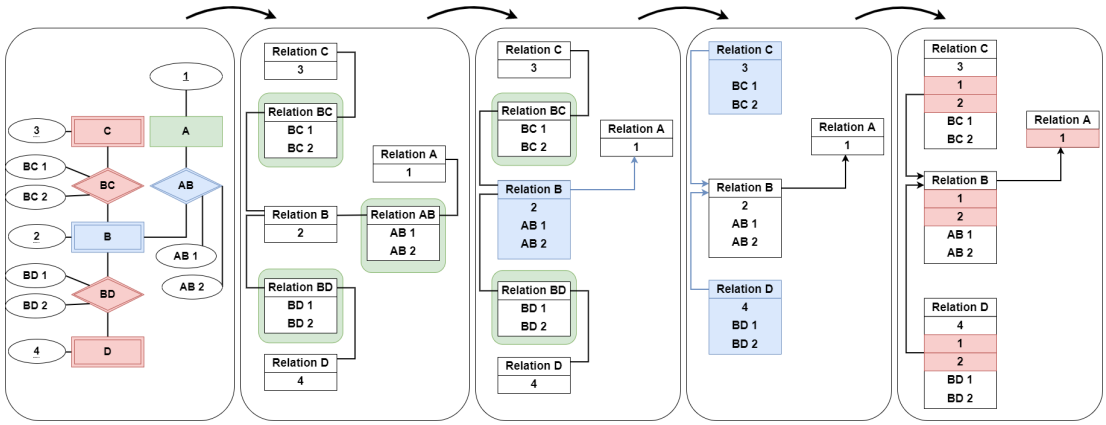

Fig. 4. Transformation of weak types

structures, by checking all weak entities up to N times, where N equals the number of weak entities in the graph.

Listing 3. Transformation of weak types

```
1  Function TransformWeakEntity(WeakEnt)
2      If WeakEnt is transformed Then
3          Return
4      End If
5
6      WeakRelations <- GetConnectedWeakRelations(WeakEnt)
7      For Each WeakRel in WeakRelations
8
9          OtherEnt <- GetOtherEntity(WeakEnt, WeakRel)
10         If OtherEnt is no StrongEntity or
11                     is not transformed WeakEntity Then
12             Continue
13         End If
14         WeakEnt.Transformed <- true
15         WeakEnt.References.Add(OtherEntity)
16         MergeTables(WeakEnt, WeakRel)
17         Return
18     End For
19 End Function
```

In contrast to the IsA structure algorithm, the transformation algorithm resolves all connected weak relationships (line 6) and immediately tries to transform them (line 7 to 18). If weak relationships are connected to a strong entity or an already transformed weak entity (line 9 to 13), the current weak entity can determine its existence-dependent type and therefore can be transformed (line 15 to 16).

Note, that the cardinalities of the weak relationship are not to be considered for the basic transformation, since these can only be 1:1 and N:1 towards the identifying type. The given algorithm realizes both functionalities by means of a foreign key dependency.

E. Transformation of relationships

Transformation of regular relationships requires prior execution of all previous algorithms, since all relations for entities, weak types and relationships require to have the complete primary keys.

Generally, there are three cases to consider when translating.

If a relationship connects two entities and the cardinality is N:M, the primary keys of the two entities are added to the relation of the relationship. These then reference the primary keys of the entity relations as foreign keys. The translation of N-ary relations is done regardless of their cardinality. The transformation of these is analogous to the translation of binary N:M relations. In this case, the relation of the relationship receives the primary keys of all connected entities. Each of this primary keys refers to the primary key of the corresponding entity in the form of a foreign key.

If the cardinality of the relationship is 1:N, the relationship is resolved by merging the relation of the relationship with the entities relation on the N side. In addition, this merged relation receives all primary keys of the opposite entity as normal attributes. These act as foreign keys on the opposite entity. In the third case, the relationship cardinality is 1:1. The translation is to be performed analogously to 1:N relationships, and the entity that receives the foreign keys and relation attributes must be specified for this purpose. Since this work, the Min-Max notation is used, the optionality has to be considered to avoid zero values. If one of the cardinalities describes an optionality, the attributes of the relationship and foreign keys are added to the entity on the other side. If both or none of the cardinalities describe an optionality, then the optionality is arbitrary.

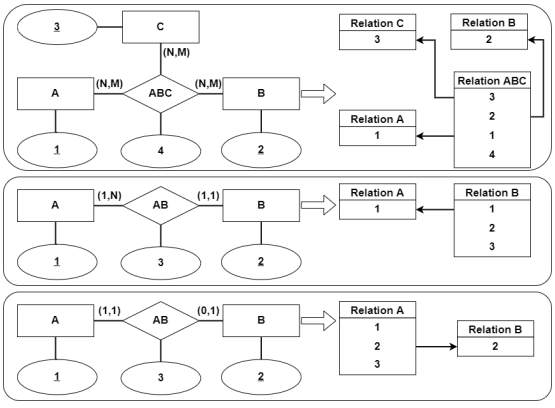

Fig. 5. Transformation of relationships

Figure 5 shows the above cases. On the left side is the ER model. On the right side is the relational model resulting

from the translation of the ER model. The first box in Figure 5 shows the transformation of N:M relationships, while the second displays a 1:N transformation. The last one shows the transformation of an optional 1:1 relationship.

V. Evaluation

Since the main focus of this paper is on practical application, also a practical evaluation process was chosen. To evaluate the algorithms given in this work a web application has been developed and put into operation, which implements all algorithms presented in this work. The application also contains a graphical editor.

Within the editor the user is able to create extensive and complex ER models, which are only restricted by a few limiting rules, required to allow an unambiguous logical assignment of the ER components. Based on the user-generated diagram, the presented algorithms are used to transform the diagram into a relational model without further user intervention.

To enforce correctly modeled ER diagrams, a validation process was implemented. This validation mechanism is designed in a user-friendly and proactive manner to ensure that a user is able to model fast and easy. Therefore, the validation offers maximum flexibility by only restricting actions which would necessarily lead to a violation of a rule and therefore lead into an mandatory reverse action. Utilizing the presented algorithms and the validation procedure, the graphical editor is able to translate any model that can be modeled with it into the relational model. Furthermore, the editor visually presents the created relational model to the user.

To increase the practical applicability, an SQL generator was implemented within the application, which works on the basis of the generated relational model and generates SQL schema definitions. The generated SQL is in PostgreSQL dialect.

The graphical editor is shown in Figure 6. At first, a left side bar can be seen. Elements on this bar can be dragged into the drawing area to the right to create new elements. The use of drag & drop is intended to make it possible to create elements quickly and intuitively. The drawing area itself can be expanded endlessly to the right and bottom. If an element is selected, a side bar becomes visible on the right side, which provides additional options. These are, for example, assigning a name, deleting the element or creating a link to another element. In the case of Figure 6, a relation was selected which further enables the option to add new associations or edit the cardinalities to existing entities.

To further enhance the practical applicability, a save and load function has been implemented. Using this functionality by clicking buttons on the right top of the editor, a model can be saved in the form of a text file and at any time be loaded into the editor. By using the button at the center bottom the model can be transformed into the relational model, which will be, without the need of any further user interaction, transformed by the implemented algorithms and visually presented to the user. Furthermore, it is possible to freely switch between the conceptual and relational view using the tab bar in the upper left. In order to generate SQL code from the relational model, data types can be entered in the columns of the relational model.

By means of these results, it is shown that the given algorithms are implementable and are capable of performing a transformation of ER models into relational models. We have also conducted a preliminary evaluation regarding the correctness of the transformation by testing the editor with a predefined set of ER models containing different combinations of modeling elements. The evaluation was successful as the editor correctly transformed all supported concepts. A more comprehensive evaluation will be part of our future work. To enable a broader evaluation and application of the editor, we made it available open source [31].

VI. Conclusion

Despite its age, ER modeling is still the most prevalent way of creating conceptual data base schemata. Since its advent in the 1970s, various extensions have been proposed. Due to this, many different flavors are currently used in modern editors. To be applicable as technical database schema, the ER models have to be transformed into relational models. This can be a complicated and error-prone task. Therefore, various standardized transformation approaches have been proposed over the decades. However, these approaches remained rather theoretic and did not include operational semantics. Thus, no tool support was established utilizing them and the transformation process remained manually to a large extend.

Despite this issue, modeling support for ER models was achieved. To date, a high number of editors is available in different flavors. There are diagram tools offering ER diagram creation, database clients with ER schema creation options, or other editors like UML editors incorporating database modeling. While some of them only provide diagrams, others enable the direct application to relational databases. However, there is still no practical automated transformation of ER diagrams to relational ones. Editors offering this do not enable the modeling of real ER diagrams but rather enhanced DB diagrams or omit important prevalent concepts like n-ary relationships or generalization.

To tackle this issue, we proposed an approach for transforming ER models to relational ones with a strong focus on applicability and operational issues. The ER model incorporates the most prevalent and necessary concepts [29] as n-ary relationships, multi-valued and composite attributes, or generalization. All of these can be correctly transformed in any meaningful combination enabling great flexibility for the input models.

To prove the applicability of the proposed approach, we have implemented a graphical ER editor capable of creating diagrams containing all mentioned concepts as well as a an automated transformation to relational models. As practical applicability was our focus, we also added a validation mechanism to the editor that guarantees the creation of correct and transformable ER models while providing the user as much flexibility as possible and a good user experience. All in all we have shown a transformation approach that can easily

 ISBN: 978-1-68558-056-8

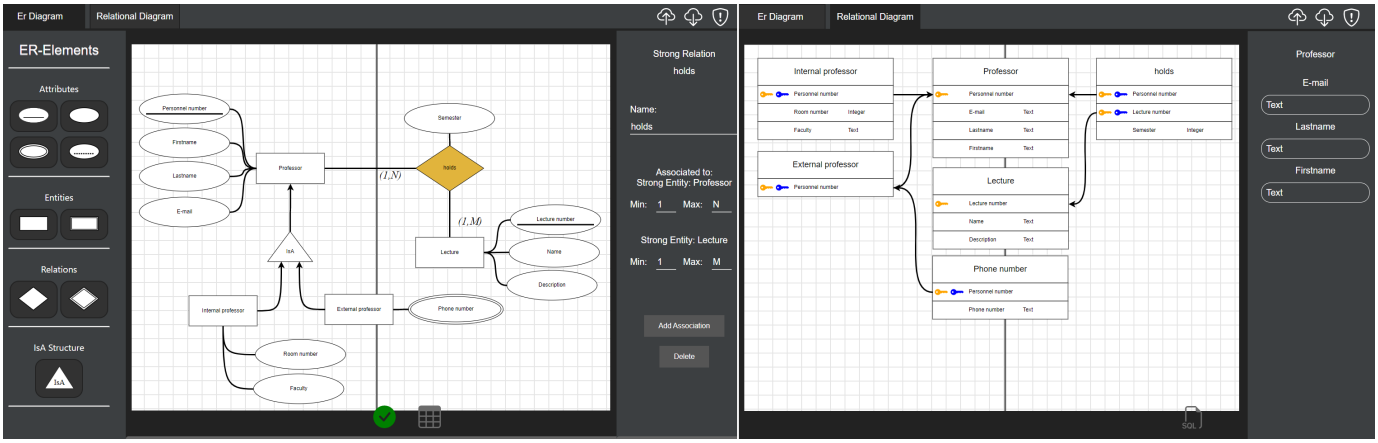

Fig. 6. ER Modeling Tool

be implemented in editors. This can aid future conceptual database modeling, spare time, and reduce errors resulting from manual transformation.

Our future work will focus on promoting the transformation approach and the editor. At first we will carry out studies investigating the correctness of the approach and the usability of the editor. Further, we will add additional features to both of them. This includes additional and optional concepts to the ER models, like composition structures as well as other diagram types and transformations.

REFERENCES

[1] P. Chen, "The entity-relationship model—toward a unified view of data," ACM Trans. on DB Sys. vol. 1, no. 1, pp. 9-36, 1976.

[2] R. Elmasri and S.B. Navathe, Fundamentals of Database Systems. Benjamin/Cummings, Redwood City, 1994.

[3] S. Spaccapietra and C. Parent, "ERC + : an object based entity-relationship approach," Conceptual Modeling DBs and Case: An Integrated View of Information Systems Development, John Wiley, 1992.

[4] B. Thalheim, "Extending the entity-relationship model for a high level, theory-based database design," 1st Int. East West Database Workshop, pp. 161-184, 1990.

[5] T.J. Teorey, "Database Modeling and Design," The Entity-Relationship Approach, Morgan-Kaufmann, San Francisco, 1990.

[6] J. Makowsky, V. Markowitz, and N. Rotics, "Entity-relationship consistency for relational schemas," Int. Conf. on Database Theory, pp. 306-322, 1986.

[7] A. D'Atri and D, Saccà, "Equivalence and Mapping of Database Schemes," VLDB '84, pp. 187–195, 1984.

[8] E. Wong and R. Katz, "Logical design and schema conversion for relational and DBTG databases," Int. Conf. on ER Approach to Sys. Analysis and Design, pp. 311-321, 1979.

[9] S. Jajodia, P.A. Ng, and F.N. Springsteel, "The Problem of Equivalence for Entity-Relationship Diagrams," IEEE Trans. on SE vol. 9, no. 5, pp. 617 - 630, 1983.

[10] T. Teorey, D. Yang, and J. Fry, "A logical design methodology for relational databases using the extended entity-relationship model," ACM Computing Surveys (CSUR) vol. 18, no. 2, pp. 197–222, 1986.

[11] T. Ling, "A Normal Form For Entity-Relationship Diagrams," 4th Int. Conf. on the ER approach, pp. 24-35, 1985.

[12] H. Sakai, "Entity-relationship approach to the conceptual schema design," SIGMOD '80, pp. 1-8, 1980.

[13] Lucidchart. Last visited: 2023.01.24. [Online]. Available: https://www.lucidchart.com

[14] Draw.io. Last visited: 2023.01.24. [Online]. Available: https://app.diagrams.net/

[15] VisualParadigm. Last visited: 2023.01.24. [Online]. Available: https://online.visual-paradigm.com

[16] ERDPlus. Last visited: 2023.01.24. [Online]. Available: https://erdplus.com/

[17] pgAdmin. Last visited: 2023.01.24. [Online]. Available: https://www.pgadmin.org/

[18] MySQL Workbench. Last visited: 2023.01.24. [Online]. Available: https://www.mysql.com/products/workbench/

[19] Enterprise Architect. Last visited: 2023.01.24. [Online]. Available: https://www.sparxsystems.de/

[20] C. Fahrner and G. Vossen, "A survey of database design transformations based on the entity-relationship model," Data and Knowledge Engineering vol. 15, no. 3, pp. 213–250, 1995.

[21] M. Kasra Habib, "On the Automated Entity-Relationship and Schema Design by Natural Language Processing," Int. J. of Engineering and Science vol. 8, no. 11, pp. 42–48, 2019.

[22] P. G. T. H. Kashmira and S. Sumathipala, "Generating Entity Relationship Diagram from Requirement Specification based on NLP," 3rd Int. Conf. on Information Technology Research, pp. 1-4, 2018.

[23] S. Ghosh, P. Mukherjee, B. Chakraborty, and R. Bashar, "Automated Generation of E-R Diagram from a Given Text in Natural Language," Int. Conf. on Machine Learning and Data Engineering, pp. 91-96, 2018.

[24] M. Ahsan Raza, M. Rahmah, S. Raza, A. Noraziah, and R. Abd. Hamid, "A Methodology for Engineering Domain Ontology using Entity Relationship Model," Int. J. of Advanced Computer Science and Applications vol. 10, no. 8, pp. 326–332, 2019.

[25] K. Ľachová and P. Trebuňa, "Modelling of Electronic Kanban System by Using of Entity Relationship Diagrams," Int. Scientific Journal about Logistics vol. 6, no. 3, pp. 63–66, 2019.

[26] A. Ramírez-Noriega, Y. Martínez-Ramírez, J. Chávez Lizárraga, K. Vázquez Niebla, J. Soto, "A software tool to generate a Model-View-Controller architecture based on the Entity-Relationship Model," 8th Int. Conf. in Software Engineering Research and Innovation, pp. 57-63, 2020.

[27] Y. Liu, X. Zeng, K. Zhang, and Y. Zou, "Transforming Entity-Relationship Diagrams to Relational Schemas Using a Graph Grammar Formalism," IEEE Int. Conf. on Progress in Informatics and Computing, pp. 327-331, 2018.

[28] L. Yang and L. Cao, "The Effect of MySQL Workbench in Teaching Entity-Relationship Diagram (ERD) to Relational Schema Mapping" Int. J. Modern Education and Computer Science vol. 7, pp. 1–12, 2016.

[29] A. Kemper and A. Eickler, Datenbanksysteme. Eine Einführung [English: DBMS. An Introduction] Oldenbourg, 2015.

[30] G. Vossen, Datenmodelle, Datenbanksprachen und Datenbankmanagementsysteme [English: Data Models, Database Languages and DBMS] 5. Aufl. Oldenbourg, 2005.

[31] Editor Implementation. Last visited: 2023.01.24. [Online]. Available: https://github.com/SimonRuttmann/ERModellingTool

 ISBN: 978-1-68558-056-8

Sustained Growth of Football Teams with Academy Training

- Proposal of Shapley-based Measurement -

Seiji Matsuhashi
Faculty of Economics
Gakushuin University
Tokyo, Japan
e-mail: 22122004ATgakushuin.ac.jp

Yukari Shirota
Faculty of Economics
Gakushuin University
Tokyo, Japan
e-mail: yukari.shirotaATgakushuin.ac.jp

***Abstract*—In this paper, the dominant factor for sustainable growth in football teams is described. Using the data of Japan-League teams, we conducted machine learning based regression and its interpretation by Shapley values revealed the Academy development was significant. In the financially large-scaled strongest teams, the Academy development is important to sustain the high scores/ranking. In small or medium sized teams, Academy development is another approach for growth to the upper league under limited budget. To measure the Academy development level, Matsuhashi's Measure based on Shapley values is proposed and it is proven that Matsuhashi's Measure has the highest correlation with the top strongest teams' winning points.**

Keywords-football; academy training; Shapley values; SHAP; effective growing; Matsuhashi's Measure.

I. INTRODUCTION

This study is a regression-based corporate evaluation analysis, and the subject of the analysis is football teams in the Japan-League (hereafter J-League), which will soon celebrate its 30th anniversary since 1993. The main evaluation indicator for football clubs is how effectively they perform under a limited budget. This is a common goal to every professional sports organization.

Investing a large amount of money for a high performance is a straightforward and instant approach. The first objective of this paper, however, is to explore another approach by which a small/medium sized team can effectively grow with a limited budget. The first author, Matsuhashi, has been interested in data analysis of the J.-League for many years. He has found some cases in which utilizing young players from the academy were important to the league performance for the teams. These clubs had small budgets in early years. However, owing to the results of such young players, the club increased its ranking, promoted to J1 League, and gradually expanded its scale. Then, in order to measure the Academy development level, Matsuhashi's Measure was defined in our previous work [1].

The second objective of the paper is to explore sustainability of the already large-scaled teams. In general, it is difficult to maintain high-performance in a company. In J-League, it is difficult for the strongest teams to keep the top positions. The paper showed that one of the driving forces is Academy development. In the paper, we will prove that Matsuhashi's Measure is useful to measure the large-scaled football team's sustainability.

We use machine learning regression analysis and Shapley values for interpretation of the result. The next section describes the data we used. Section 3 describes the analysis method using Shapley values. Section 4 explains Matsuhashi's Measure (MM) and shows the high correlation between the MM values and sustained high performance large teams. In Section 5, discussion concerning Matsuhashi's Measure is conducted. Finally, we conclude this paper.

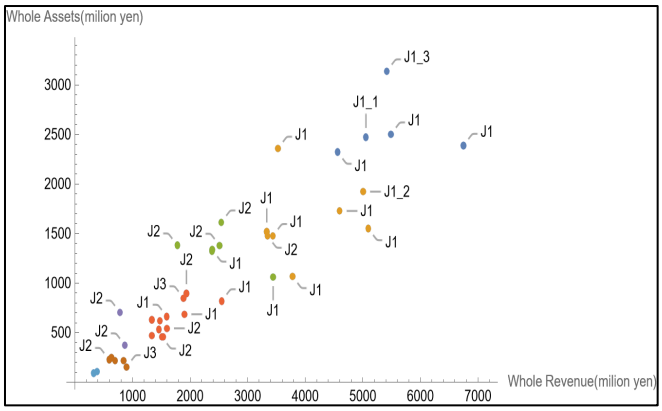

Figure 1. 10-year average operating revenues and total assets of the J-League clubs through 2021.

II. DATA AND METHOD

In this section, we explain the data used and the regression analysis method.

The cost of strengthening a professional soccer team is enormous in every country. The dominant part of the cost is the personnel cost for star players, but there is a disparity in financial scale of teams. Figure 1 shows a scatterplot between 10-year average operating revenues and total assets of the clubs from J-League. J1 is the top category, followed by J2 and J3. As shown in Figure 1, there is the tendency that the larger financial sized team, the higher ranking it has.

On the other hand, however, some medium-sized teams belong to J1. We found that they are committed to Academy development. Therefore, the following hypotheses were formulated for the growth pattern of small and medium-sized clubs.

 ISBN: 978-1-68558-056-8

Hypothesis: Smaller clubs with smaller budgets can take an approach to achieve higher performance by training young players in Academies.

We will conduct analysis to find such clubs. The analysis method is regression. The target variable is the annual ranking in the league converted into a score of 100 points, with 100 being the highest. As for the choice of explanatory variables, we examined the correlation coefficients among the managerial variables and found that many of them had multicollinearity, so we finally selected the following two explanatory variables.

Explanatory variables
(1) Salary costs: Personnel costs for the year.
(2) Academy operating costs: Total costs for the seven years prior to the target year (2021).

The impact of (1) salary costs' increase is instant but short. In contrast, (2) academy operation costs take long time until the effects appear. To see the effect, we use 7-year total cost of Academy operations. For example, when analyzing the 2021 season, the target value is the ranking scores in 2021and the salary cost data in 2021 are used. Concerning Academy operating costs, data from 2014 to 2020 are used. Data were taken from the site of J-League official [2]. This is about the financial data written in English, which includes all clubs belonging to J-League.

The regression analysis method was the XGBoost algorithm by scikit-learn package [3] [4]. Its GitHub site [5] has more information on the algorithm. Explanatory variables are, in advance, standardized for each variable shown in Figure 2.

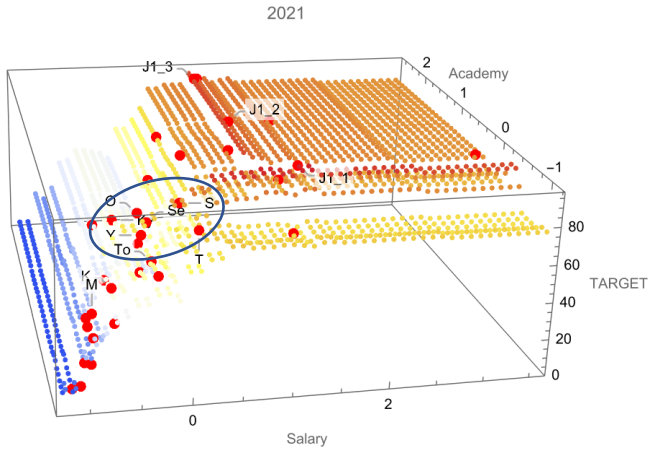

Figure 2. Resultant regression model with standardized salary and academy values and the ranking score as the target.

In Figure 2, the observation data represented by red points are plotted in a three-dimensional regression model *f(X)*. The warm colours such as brown correspond to higher target values.

In general, the more the salary and Academy operating costs, the higher the rank score. However, there are some exceptional medium-sized teams which have high rank scores even though their financial resources are limited. The teams will be later described in Section 4.

III. SHAPLEY VALUE

In this section, Shapley and SHAP values that we used are explained.

Shapley values are solutions in a game theory by multiple players [6-8] [9]. If n players work together, the profit (for example, 900 EURO) is divided to them. Then how they should divide the profit? How much is the individual player's contribution? Shapley found the unique solution of this question.

The main concept of Shapley values is **characteristic function** $\nu(X)$:

$$\nu : 2^n \rightarrow R$$

where n is the number of players and R means the profit by the subset of players. For example, there are three players A, B, and C. If A and B work together, the profit is 600 EURO. If C works together with A and B, then the profit becomes 900 EURO. The characteristic function defines the profit corresponding to any subset of players.

If n is 3, then the number of subsets is 2 x 2 x 2 = 8. In a real world (not in a theoretical world), it is too difficult to define the characteristic function. However, if such a characteristic function is given or can be found, each player's profit can be calculated using the following formula:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[\nu(x_{S \cup \{i\}}) - \nu(x_S)\right]$$

where ϕ_i is the Shapley value for player i, F is a set of players, and S is a subset of F which does not include i-th player, $S \subseteq F \setminus \{i\}$.

$|F|!$ is the permutations of the number of F. The term $[\nu(x_{S \cup \{i\}}) - \nu(x_S)]$ is player i's marginal contribution to the profit by S; the function ν is evaluated with the input of the player set $(x_{S \cup \{i\}})$ and then the function ν is evaluated with the input of the player set (x_S). The difference $[\nu(x_{S \cup \{i\}}) - \nu(x_S)]$ is the key part of the Shapley formula.

The term $|S|!\ (|F| - |S| - 1)!\ /|F|!$ expresses the appearance possibility of $x_{S \cup \{i\}}$. First S exists and then player i comes, and finally the left members of which number is $(|F| - |S| - 1)$ attend. Equality of appearance possibility is supposed here.

Finally, the sum of weighted differences is calculated that becomes the player i's profit. The easy explanation of the formula was described by Roth in [8].

Lundberg et al. modified the original Shapley value, so that we can use Shapley values in the machine learning regression analysis [10] [11] [12]. The customized Shapley value is called SHAP values. The differences are as follows:

(1) SHAP is defined for each explanatory variable i, instead of player i.
(2) Each data has its own **characteristic function**.
(3) The characteristic function is calculated using the regression prediction model $f(X)$

 ISBN: 978-1-68558-056-8

In this case of this paper, each football team has its characteristic function which is calculated using the regression prediction model $f(X)$. The regression model is visually shown in Figure 2. The vertical axis means the predicted target values.

Using the resultant regression model $f(X)$, an individual team's characteristic function $v_{team}(X)$ is calculated. **Function $v_{team}(X)$ predicts the team's target value for any subset of explanatory variables.** The characteristic function for each team reflects the team's behavioral characteristics.

If there is a missing value among the predictors, the value of *f(X)* cannot be calculated. Lundberg's idea is to input the average value of the predictor variable for the missing parameter [10]. By this idea, they could solve the problem that the characteristic function could not be defined in a real world. In Operational Management field, the concept of **an industry average value** is very important. In industry analysis, we firstly investigate whether a particular company's value is above or below the industry average. We think that this solution is reasonable from the industry analysis viewpoint as well.

In this paper, we express each predictor's SHAP value as "predictor_SHAP" such as **Academy_SHAP**.

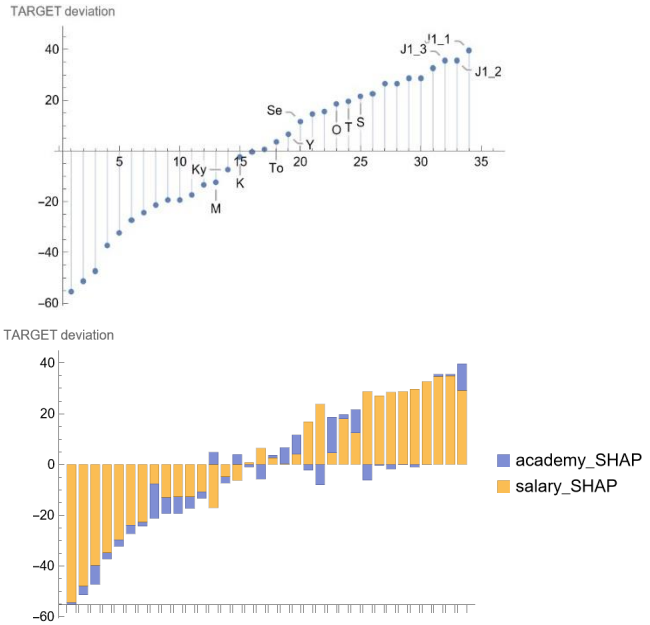

Figure 3. Target deviation values are divided to 2 SHAP values.

The resultant two SHAP values per team are visually illustrated in Figure 3. The x axis shows the 34 football teams. The y axis shows the deviation of target value which is its target value minus the average. The total SHAP values per team approximately becomes the deviation (see Figure 3).

Theoretically, the sum of SHAP values per data becomes equal to the target deviation of the data. However, if the fitting level of *f(X)* is low, the SHAP total is not equal to the deviation.

In the lower bar chart in Figure 3, it is found that Salary_SHAP is much greater than Academy_SHAP. This means that in many teams the dominant factor of the target (ranking score) value is the salary expenses. However, some teams have significantly large Academy_SHAP. The ranking top team (see the right-end bar) has large Academy_SHAP, compared to others. In the next section, we shall evaluate the effect.

In a machine-learning based regression analysis, SHAP values are widely used for in various fields [13, 14]. Concerning football players, there are many researches using Shapley/SHAP values.

Sizov et al. use Shapley values to determine the salary prices or values of football players [15]. Hiller uses Shapley values to determine the performance/importance of players in each team of the Bundesliga in the season 2012/2013 [16]. Buzzacchi1 et al. use Shapley values to evaluate ranking of football managers in the Italian Serie A [17]. Marc Garnica-Caparrós uses SHAP values to understand gender differences in professional European football [18]. For example, it says that a high number of ground duels increases the probability of the model classifying a female player.

Although there are many researches by SHAP-based approach in the football field, as above mentioned, the target is the players' performance or managers' skills. As far as we know, there are **no** football team's management strategy evaluation by the SHAP approach. Our research is the first football teams' managerial structure evaluation by SHAP values.

In other fields except football, managerial researches by using SHAP values exist, as industry analysis [19][20] [21]-[23] [24].

IV. ACADEMY DEVELOPMENT LEBEL MEASUREMENT

In this section, we propose a measurement of academy training achievement level using the SHAP values described in the previous section.

First, we consider meanings of the Academy_SHAP value. Even if the Academy's operating expenses are large, if the Academy does not generate results, the ranking score does not increase and the Academy_SHAP value does not increase. In addition, even if players from the academy play more games, they do not always play important roles to obtain points of victory. Just Academy_SHAP is not sufficient to express the Academy development level, because there may be other hidden factors that contribute to the improvement in the ranking. In general, it is difficult to find causal relationships in policy and strategy evaluation [25].

To solve the problem, Matsuhashi defined the measurement for Academy development levels, based on the SHAP values. A new concept titled **"percentage of Academy graduates' participant ratio"** is introduced. This is the ratio of the number of appearances by players from the Academy to the number of league games available in a season (For example, for J1 in 2019, 34 games x 14 players). The calculation of the percentage of Academy graduates' participant ratio is defined as A/B where

- A: Total number of games from 2019 to 2021 played by Academy graduate players since 2011.
- B: The number of available slots for the three-year period from 2019 to 2021.

Matsuhashi's Measure = (Percentage of Academy graduates' participant ratio) x **(Academy_SHAP value).**

Next, we evaluate the performance of the measure.

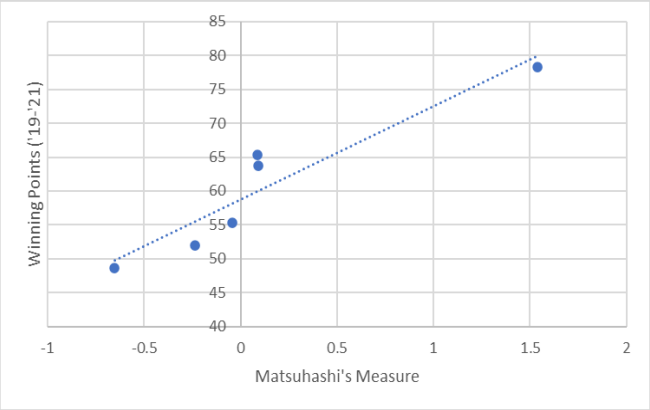

Figure 4. Relation between Matsuhashi's Measure and the average winning points with the correlation coefficient 0.94.

In Figure 4, the operating revenues top 6 teams in J-League are selected and the data are plotted. The x-axis shows Matsuhashi's Measure and the y-axis shows the average winning points from 2019 to 2021. These highest operating revenue teams are likely to be strongest ones as shown in Figure 1. The correlation coefficient of the relationship is 0.94 (see Figure 4). **This high correlation value indicates that the Matsuhashi's measure is highly correlated with the ranking score in the top group**.

The reason why the 3-year average winning points are used is that we would like to investigate the performance in the sustained period. From the relationship, we can say that **one of elements for maintaining strong teams is an excellent academy development base, and that Matsuhashi's Measure could measure the Academy development level with high accuracy**.

Reversely, what is the feature of the highest Matsuhashi's Measure team? In Table 1, the top 11 teams with the highest Matsuhashi's Measure values for three years through 2021 are listed. The names in yellow indicates a small or medium sized teams. In the scale-based clustering in Figure 1, these 11 teams are belonging to green or red clusters.

In the previous work, Matsuhashi described the followings [1]: *Although the ordinary strategy is to improve the ranking by increasing the financial investment, there were some medium-sized clubs that achieve high performance by investing in the Academy operation expenses. The Matsuhashi's Measure was effective in identifying these growing medium-sized clubs.*

TABLE 1. TEAMS WITH HIGHEST MATSUHASHI'S MEASURE VALUES.

Name	Matsuhashi's Measure ('19-'21)
S	2.074
J1_1	1.540
Y	0.842
O	0.677
T	0.323
Se	0.283
J1_2	0.091
J1_3	0.088
K	0.088
To	0.018
M	0.005

Using the resultant regression model, the highest Matsuhashi's Measure 11 teams are marked in the regression model (see Figure 5 and 6). The blue arrow marks depict the J1_1 to J1_3 which are the large-scaled teams in J1. The number of medium-sized teams in Table 1 is 8. These clubs have potential to be in a transition state to the large-scaled teams. Among them, the higher 5 teams are marked in the large circle, and the other lower 3 teams are marked in the small circle in Figure 5. The teams in the large circle can be identified ***growing medium-sized teams which produce high ranking scores.***

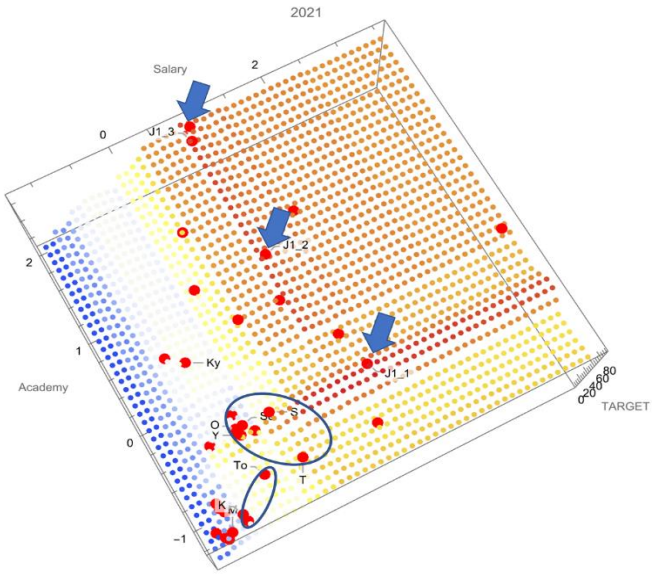

Figure 5. Highest Matsuhashi's Measure 11 teams on the regression model.

The closer look of the transition-state teams is shown in Figure 6. The 5 teams in the large circle have higher ranking score (target) values, compared with the teams in the small circle. Especially the team “S” with the highest Matsuhashi’s Measure has the highest target value. The medium-sized teams could be divided to the upper 5 teams and the lower 3 teams by target values as well as by the Matsihashi's Measure.

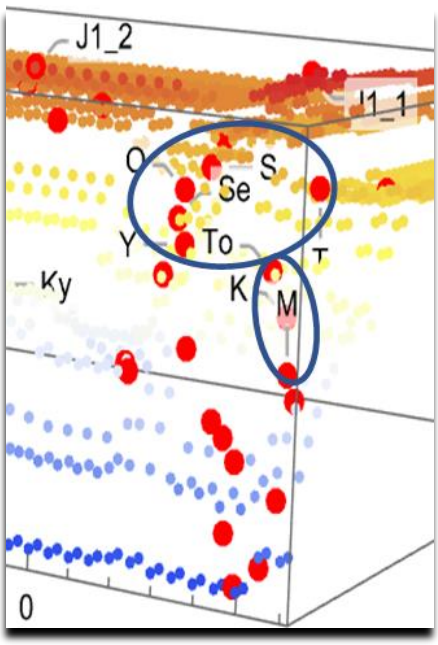

Figure 6. A closer look at highest Matsuhashi’s Measure teams on the regression model.

V. DISCUSSION

In this section, the reliability of Matsuhashi’s Measure (hereafter MM) will be discussed.

From the previous section result, it was found that the highest MM teams include

(1) the strongest three teams in 2021 which are large-scaled (J1_1 to J1_3), and
(2) effectively growing teams in 2021 which are medium-scaled.

When we compare MM values of J1_1 through J1_3, the strongest J1_1 team’s MM is larger by approximately 17 times (divide 1.54 by 0.091). One of driving forces of J1_1’s strength may be the Academy development level. Among the large-scaled teams, the high correlation 0.94 is found between MM values and the average winning points (see Figure 4). Among the large-scaled teams, Academy development level is important for the sustainability of the high performance and the level may be measured by MM.

Among medium-sized teams, there is high correlation between MM values and the target ranking scores (see Table 1 and Figure 6). For a medium-sized team, driving forces for growing is the Academy development level and the level may be measured by Matsuhashi’s Measure,

This study focuses on the relationship between academy development and annual ranking. The relationship represented by Matsuhashi's Measure may just a correlation and may not be a causal relationship. More analyses are necessary to prove the causal relationship [25]. This is our future research theme. At this stage, we can say that academy development is one of dominant factors to maintain its high performance for large-scaled teams and one of the effective approaches as a growth pattern for small and medium-sized teams.

VI. CONCLUSION

In this paper, we conducted a regression analysis of J-League results and analyzed the results using Shapley values, or more precisely, SHAP values. The novelty of the method is that the SHAP value is used to evaluate the contribution of the individual explanatory variables to the target value, taking into account the structural characteristics of individual football teams.

Based on the SHAP of Academy costs, the Academy development achievement measure named Matsuhashi’s Measure is defined as **(Percentage of Academy graduates’ participant ratio) x (academy_SHAP value).**

Among the large-scaled 6 teams, the correlation between Matsuhashi’s Measure values and average winning points was 0.94 which is very high. For the large-scaled teams, Academic development may be one of their sustainability factors. Then Matsuhashi’s Measure may measure the stability levels.

On the other hand, for medium-sized teams, Academy development gives another approach for growing. In such transition state teams with highest Matsuhashi’s Measure values, common feature can be found that they generate high performance effectively under the limited budgets. We can identify these teams’ positions visually on the regression model. These medium-sized teams have improved their scores by increasing the level of academy development achievement. This fact gives hope to smaller teams.

In conclusion, Academy development gives stability of high ranking to large and strongest teams and for the medium-sized teams, sustainable growth. In the paper, to measure the Academy development level, Matsuhashi’s Measure can be used with high accuracy.

Lastly, we shall describe the novelty of this analysis method. As company performance analysts, our interests exist on finding another approach to growth or success other than large investment. Observing the regression model, we may identify medium-sized but high-performance companies. Investigating the managerial states of these companies in the transition state, we would be able to find another approach strategy specific to the industry field. In the paper, the target industry field was the football team management. Then the recommended strategy was Academy training which was effective for sustained growth.

We shall continue to clarify the cause and result relation for high performance in football team strategies.

References

[1] S. Matsuhashi and Y. Shirota, "Finding of Corporate Growth Patterns by Shapley Values -- Case Study of Academy Development in the J-League -- " in *IEICE Technical Report* Kyoto, 2022/12/22-23 2022: IEICE, Technical Committee on

Infomation-Based Induction Sciences and Machine Learning (IBISML), p. (in printing).

[2] J. League. "Japan League Offcial Site. Individual Club Management Information | 公益社団法人 日本プロサッカーリーグ（Ｊリーグ） (jleague.jp) (accessed 2023/02/08).

[3] O. Kramer, "Scikit-learn," in *Machine learning for evolution strategies*: Springer, 2016, pp. 45-53.

[4] G. Hackeling, *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, 2017.

[5] XGBoostDevelopers. "XGBoost Decumentation (Revision 534c940a.)." https://xgboost.readthedocs.io/en/stable/ (accessed 2022/11/13.

[6] L. S. Shapley, "A value for n-person games, Contributions to the Theory of Games, 2, 307–317," ed: Princeton University Press, Princeton, NJ, USA, 1953.

[7] A. E. Roth, "Introduction to the Shapley value," *The Shapley value*, pp. 1-27, 1988.

[8] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.

[9] E. Winter, "The shapley value," *Handbook of game theory with economic applications*, vol. 3, pp. 2025-2054, 2002.

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[11] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[12] S. M. Lundberg and S.-I. Lee, "Consistent feature attribution for tree ensembles," *arXiv preprint arXiv:1706.06060*, 2017.

[13] D. Lubo-Robles, D. Devegowda, V. Jayaram, H. Bedle, K. J. Marfurt, and M. J. Pranter, "Machine learning model interpretability using SHAP values: Application to a seismic facies classification task," in *SEG International Exposition and Annual Meeting*, 2020: OnePetro.

[14] A. Joseph, "Shapley regressions: A framework for statistical inference on machine learning models," presented at the King's Business School Working Paper, 2019.

[15] G. Sizov, P. Oztürk, and K. Valle, "The Use of Game Theory in Feature Selection."

[16] T. Hiller, "The importance of players in teams of the German Bundesliga in the season 2012/2013–a cooperative game theory approach," *Applied Economics Letters*, vol. 22, no. 4, pp. 324-329, 2015.

[17] L. Buzzacchi, F. Caviggioli, F. L. Milone, and D. Scotti, "Impact and Efficiency Ranking of Football Managers in the Italian Serie A: Sport and Financial Performance," *Journal of Sports Economics*, vol. 22, no. 7, pp. 744-776, 2021.

[18] M. Garnica-Caparrós and D. Memmert, "Understanding gender differences in professional European football through machine learning interpretability and match actions data," *Scientific Reports*, vol. 11, no. 1, pp. 1-14, 2021.

[19] K. Yamaguchi, "Feature Importance Analysis in Global Manufacturing Industry," *International Journal of Trade, Economics Finance*, vol. 13, no. 2, pp. 28-35, 2022. [Online]. Available: http://www.ijtef.com/vol13/719-UT0036.pdf.

[20] Y. Shirota, K. Kuno, and H. Yoshiura, "Time Series Analysis of SHAP Values by Automobile Manufacturers Recovery Rates," in *2022 6th International Conference on Deep Learning Technologies (ICDLT)*, 2022: ACM, pp. 135-141.

[21] K. Kuno and Y. Shirota, "Time Series Analysis of Shapley Values in Machine-Learning Regression," *IEICE Technical Report; IEICE Tech. Rep.*, 2022.

[22] T. Hashimoto, Y. Shirota, and B. Chakraborty, "SDGs India Index Analysis using SHAP," presented at the International Electronics Symposium (IES) 2022, Surabaya, Indonesia and online, 2022/8/9-11, 2022.

[23] M. Fujimaki, E. Tsujiura, and Y. Shirota, "Automobile Manufacturers Stock Price Recovery Analysisat COVID-19 Outbreak," in *POM Nara 2022*, Online, 2022.

[24] Y. Shirota, M. Fujimaki, E. Tsujiura, M. Morita, and J. A. D. Machuca, "A SHAP Value-Based Approach to Stock Price Evaluation of Manufacturing Companies," in *2021 4th International Conference on Artificial Intelligence for Industries (AI4I)*, 2021: IEEE, pp. 75-78.

[25] E. Duflo, R. Glennerster, and M. Kremer, "Using randomization in development economics research: A toolkit," *Handbook of development economics*, vol. 4, pp. 3895-3962, 2007.

 ISBN: 978-1-68558-056-8

The Internet of Things system combined with the cloud platform applied to the data collection and analysis of the elderly's home life style

Bing-Hong Jiang
Institute of Mechatronic Engineering of NTUT
Taipei, Taiwan
Email: t110408015@ntut.edu.tw

Jung-Tang Huang
Institute of Mechatronic Engineering of NTUT
Taipei, Taiwan
Email: jthuang@ntut.edu.tw

***Abstract*—This study combines Google Cloud Platform (GCP), Google Assistant, Firebase and MongoDB for data streaming and storage through smart bracelets, smart amulets (9-axis IMU), and smart speakers (Google Home Next Mini), and the collected data changes are displayed on the webpage immediately to form a home care internet of things system for the elderly. The experiment visited five groups of families for actual testing, with ten people experimenting for one week, all wearing the smart bracelet and the smart amulet at the same time. The bracelets and amulets were transmitted by Bluetooth broadcasting, and received various physiological information and posture changes through the peripheral devices, which were stored in Firestore (Firebase's no-SQL) and MongoDB and analyzed by Cloud Function to provide relevant recommendations based on the elderly's body information. Finally, the results of the analysis and recommendations are broadcasted by smart speakers to improve the bad habits of daily life. Lightweight and inexpensive wearable devices will reduce the discomfort caused by wearing many sensors in the past, and at the same time will reduce the cost of setting up a home care Internet of Things system, which will assist the elderly in the field of home care by creating a long-term automated care system, giving a major boost to the issue of elderly care in an aging society.**

Keywords-IoT; Smart Healthcare; Data flow; Big Data Analysis; Cloud Services;

I. INTRODUCTION

Many care systems collect various physiological information and activity status of the user through many wireless technologies and sensors [1] and receive this information through peripheral devices or mobile applications so that the user can confirm the current physiological status [2] [3], thus forming a simple care system. In addition, there are also many studies that use a large number of sensors to collect large amounts of data for machine learning analysis of sleep and behavior of the elderly, aiming to confirm detailed sleep status and rhythm of the elderly [4], etc. Sampling of frailty gait and fall patterns, or deep learning or visual analysis training through open-source datasets provided by medical institutions [5], predicts or confirms the occurrence of diseases.

Although the aforementioned studies have employed a wealth of research methods and techniques to perform different aspects of analysis, they have unfortunately not been successfully applied to real-world situations. The extensive use of sensors causes user reactions and inconvenience, as well as tension; it does not maximize the value of the quantified data and causes a decrease in willingness to use. These problems make the system less intelligent and personalized.

Therefore, this study only requires the wearing of the smart bracelet and the smart amulet to improve the discomfort and rebound caused by excessive wearing of sensors in the past. The physiological data from the smart bracelet analyzed the daily physiological status of the subject, and the data from the smart amulet analyzed the walking stability. Several studies in the past found that the duration of TUG was closely related to moderate and severe Parkinson's disease [6]. The duration of TUG can be used to determine whether there is a risk of falling. Finally, the analyzed results were communicated to the subjects through the smart speaker (Google Home Nest mini) as an information disseminator and appropriate suggestions were given.

In Section 2, the experimental system architecture is introduced, using Raspberry Pi combined with smart bracelets and smart amulets with cloud platform services to form an Internet of Things, collecting subjects' physiological data, activity posture and movement status in a lightweight way to establish a daily physiological model. In Section 3, the actual wearing of the smart bracelet and the smart amulet will be carried out for daily physiological status and posture monitoring experiment, and the data will be collected and the subjects will be judged by the "Timed Up and Go Test (TUG) evaluation standard" to determine whether they have weakness symptoms. In Section 4, the data collected from the smart bracelet and the smart amulet are analyzed and the conclusions of this paper and the future directions of optimization are summarized.

II. SYSTEM ARCHITECTURE

The Bluetooth devices used in this study are shown in Table 1. A complete care system is constructed by lightweight and inexpensive devices. The smart bracelet and smart amulet are the only devices that need to be worn. The rest of the devices are fixed and placed to solve the problem of wearing too many sensors while ensuring high accuracy of data.

The smart bracelet can detect the subject's heart rate, blood pressure, body temperature, step count, walking mileage, calories and other daily physiological data; the smart amulet

 ISBN: 978-1-68558-056-8

can collect the subject's posture, movement changes, indoor positioning and other data. Since the development of smart bracelets in the market has been quite complete and the types of physiological data measured are quite comprehensive, considering the generality of the future experimental process, we designed the smart amulet to receive data from the broadcast package of commercially available smart bracelets. At the same time, because the smart bracelet function is well developed, we choose not to include the smart bracelet function when developing the smart amulet.

TABLE I. DEVICES USED IN IOT CARE SYSTEMS

Device	Function	Advantage
Smart Bracelet	Blood pressure, Step, Mileage, Temperature, Heart rate and Calorie	with bracelet protocol, Cheap
Smart Amulet (9-axis IMU)	Attitude, Motion, Height monitoring, Fall monitoring, Emergency alert and indoor positioning	With 9-axis sensor, accurate identification
Smart Speaker (Google Home Nest Mini)	Notifications, Conversations, Sentence Collection and Care	Google has a series of services and functions
Beacon	Indoor positioning	Accurate positioning function
Edge Device (Raspberry Pi 4B)	Collect Bluetooth device data and store in database	Speed up data processing and response time

This study is a combination of cloud platform and Internet of Things application, as shown in Figure 1. The data is received from the smart amulet, merged and sent to the edge device, and the edge device sends the data to the cloud database (firestore) of Google cloud platform (GCP) for storage via MQTT, and stored in the local database (MongoDB). The data stored in firestore is classified, organized and analyzed by the cloud function provided by GCP, and the data analysis results are displayed on the web page in real time so that the subjects can view the current physiological data; the data in the local database is used as a data set for future training of the machine learning classifier to maximize the value of the data. The analyzed results are actively pushed through the smart speaker using the data stream integration function of the cloud platform.

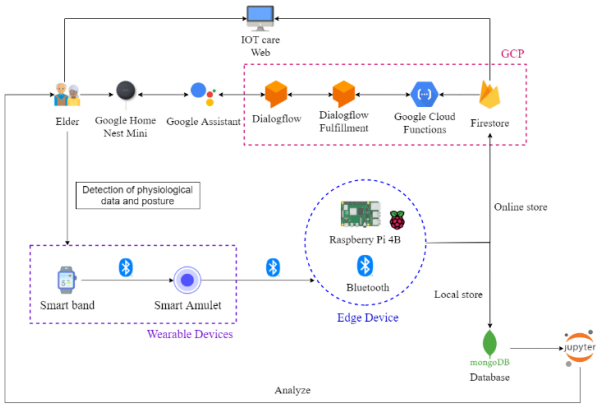

Figure 1. IoT System Architecture Diagram.

In order to realize the function of active push message by speaker, this study uses the triggering function of Cloud Functions and Cloud Pub/Sub to realize it, the data change triggers Cloud Functions through HTTP, Cloud Pub/Sub sends the object to the corresponding topic, the program in Raspberry Pi will subscribe to the same topic, through Google Text to Speech API receives and processes the text content in the object, converts the text content into voice messages, and finally pushes the voice messages actively through Google Speaker, as shown in Figure 2.

When the smart amulet detects that the subject is in an emergency situation, it allows the speaker to make an emergency broadcast through the data stream. Since the smart speaker is placed in the subject's home, the microphone radio system will only be turned on when the smart speaker hears the wake-up call, taking into account the subject's privacy concerns. In the absence of a wake-up call, the subject's privacy is protected.

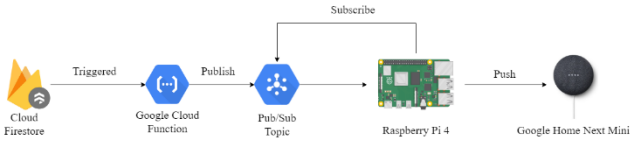

Figure 2. Flowchart of active broadcasting by speakers

III. METHOD

In this section, we describe in detail the types of subjects and the methods used to collect three types of data: physiological data, daily posture and exercise behavior.

A. Subjects

The subjects were 10 people, 5 elderly people aged 70 to 80 years and 5 young people aged 24 to 26 years, and the system was set up in 5 households. The height and weight of the elderly were as shown in Table 2, and the height values shown in Table 2 were measured without hunchback. The purpose of including young subjects in this experiment is to collect comparative data for future training of the frailty classifier. All data were transferred to the cloud and local databases through the edge device. The experiments were conducted with the consent of the subjects who received daily data collection for this experiment.

TABLE II. SUBJECT'S PHYSICAL INFORMATION

Subjects (Elder)	Height (cm)	Weight (kg)	Humpbacked	Illness & Injury
Elder 1	158	62	No	Had ankle surgery
Elder 2	155	58	Yes	Bipolar disorder & Effusion of knee joint
Elder 3	155	45	Yes	Had knee surgery

 ISBN: 978-1-68558-056-8

Subjects (Elder)	Height (cm)	Weight (kg)	Humpbacked	Illness & Injury
Elder 4	157	58	Yes	Back sprain
Elder 5	168	65	No	None

B. Physiological State Data and Daily Posture Collection

During the experiment, the subjects will be asked to wear the smart bracelet and the smart amulet for 7 days. Except for washing, they would wear them during the rest of the time. The smart bracelet is shown in Figure 3. The smart amulet was attached to the chest by means of biocompatible adhesive, as shown in Figure 4. The purpose of using the adhesive is to control the correctness of data collection, to ensure that the movement of the amulet is consistent with the body movement and does not affect the data interpretation of the 9-axis sensor. If the amulet is worn by hanging, it will cause the smart amulet to shake, which will generate noise and affect the sensor's interpretation.

The 9-axis sensor in the smart amulet includes 3-axis accelerometer, 3-axis gyroscope and 3-axis magnetometer. The coordinate system is shown in Figure 5, which can measure the acceleration, rotation angle and geomagnetic direction respectively, and calculate the subject's posture and movement distance. The G-value (1) can be calculated by taking the root of the sum of the squared acceleration of each axis. The change of G-value is used to determine whether the subject is active or not. If the subject is at rest, the G value is equal to the gravitational acceleration (about 1G).

$$\text{G-value} = \sqrt{(ACC_x)^2 + (ACC_y)^2 + (ACC_z)^2} \quad (1)$$

Through the orientation algorithm (gradient descent algorithm) proposed by Madgwick [7], the acceleration value, magnetometer value, and gyroscope value are calculated to derive the 3-axis rotation angle, which are Roll, Yaw, and Pitch. Observation of Roll, Yaw, and Pitch can analyze the walking deflection condition of the subject.

Figure 3. Smart bracelet to wear on the wrist.

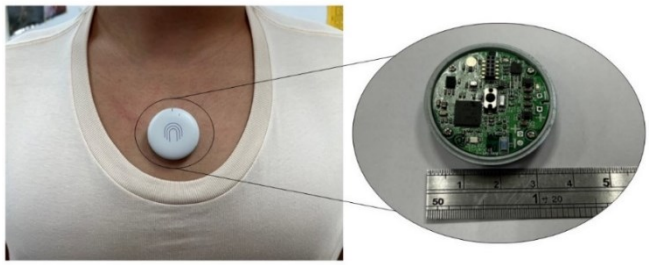

Figure 4. Smart amulet (9-axis IMU) developed and designed by our laboratory

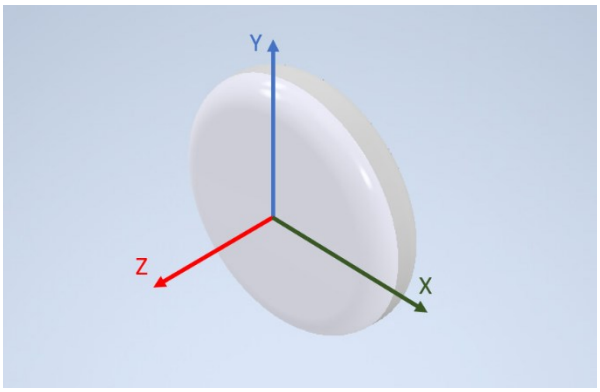

Figure 5. Smart amulet (9-axis IMU) Coordinate System

C. Behavioral Data Collection

The Timed Up and Go Test (TUG) frailty assessment standard was conducted to check the movement changes of the subject and to determine whether the subject had frailty symptoms. Time Up and Go Test uses the standard TUG protocol and starts from the center of the foot and goes forward 3m, using tape at the 3m mark and turning around the cross mark. The TUG experiment was conducted using a chair with no back rest. During the experiment, the subjects will wear smart amulets and smart bracelets to collect physiological state data and experimental posture data. Finally, the test results and exercise performance data of young and old people are compared to analyze the differences and train the classifier model.

IV. Results and Discussion

Based on the system architecture and the experimental method, the experiments are conducted to validate the care IoT cloud system designed in this study, and finally the collected data are analyzed completely.

A. Physiological Information and Daily Posture Collection Results

The data collected during the experiment will be uploaded and stored to the cloud database through the edge device, and the current posture changes will be displayed through the webpage for the subjects to view in real time, as shown in Figure. 6. In the future, the system will be developed in such a way that caregivers can check the activity of the elderly in

 ISBN: 978-1-68558-056-8

real time through the webpage. For the elderly with frail symptoms, the caregiver can check whether there is a fall in real time. If unfortunately a fall occurs, the fall posture of the smart amulet will be sent to the cloud platform to trigger the speaker to broadcast a fall warning message, and through the screen warning and broadcast warning to achieve double reminders to avoid the regret caused by negligence.

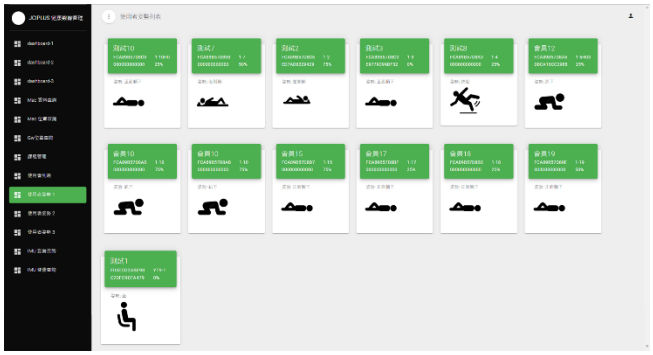

Figure 6. The status of smart amulet is displayed instantly on the web

The physiological data and postural changes collected daily were plotted for analysis. The graphs were used to clearly analyze the physiological data and postural distribution of the subjects during one hour at a point in time, as shown in Figure. 7 and 8. Figure. 7C shows that the blood pressure changes during the first 30 minutes implied a trend of pre-hypertension. According to the criteria for hypertension published by the American Heart Association [8], a diastolic blood pressure between 120 mmHg and 129 mmHg and a systolic blood pressure below 80 mmHg are the criteria for prehypertension. However, the blood pressure status returned to normal in the last 30 minutes, and the heart rate was higher in the first 30 minutes when compared with the heart rate variation graph in Figure. 7A. At the same time, the posture distribution in Figure. 8 showed a prolonged sedentary state, which was verified with the experimental activity records, and the subject was watching a movie at that time, which was presumed to be caused by the tension of the drama.

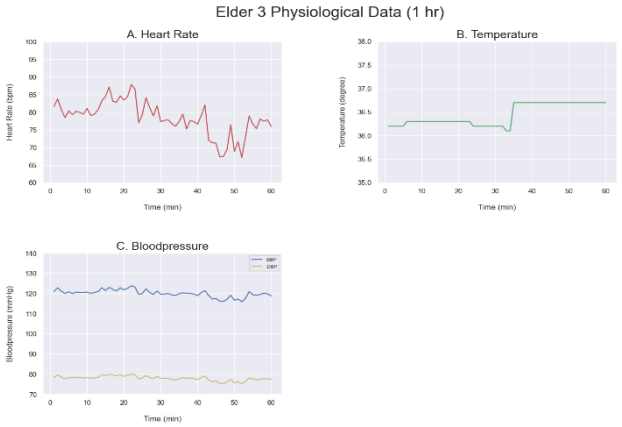

Figure 7. Line graph of the physiological data changes during a certain hour of the experiment (the same sampling time as in Figure. 8).

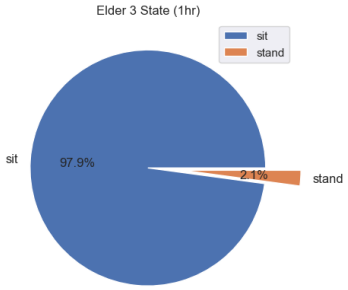

Figure 8. The pie chart of the posture distribution during an certain hour of the experiment (the same sampling time as Figure. 7).

B. *Daily Behavior Analysis*

The results of the TUG experiment with the same conditions for young person 1 and elder 1 are shown in Figure. 9. The 3-axis acceleration and G-value are magnified 512 times for easy observation and plotting on the graph. From Figure. 9A and 9C for comparison of the difference in 3-axis acceleration changes, we can observe that the amplitude of G-value of young person 1 is larger than that of elder 1. By comparing the amplitudes and observing the experimental procedure, it was inferred that the young people walked at a larger pace and faster speed.

On the other hand, comparing the difference of Z-axis acceleration, the maximum amplitude of Z-axis acceleration reached -400 as shown in the red circles in Figure. 9A and 9C, indicating that elder 1 was leaning forward than young person 1 in getting up, which could be inferred from observing the experimental procedure that elder 1's leg muscles were relatively weak and needed to be guided to stand by body strength. At the same time, the TUG test time of elder 1 was greater than 12 seconds, and it is presumed that there may be a risk of falling [9].

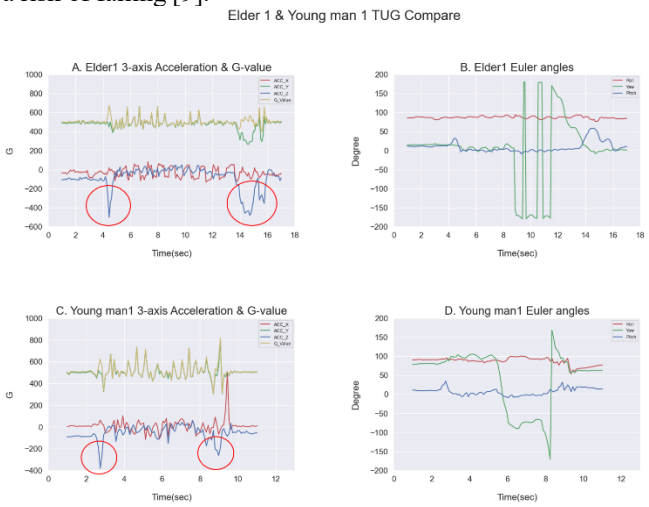

Figure 9. Comparison chart of TUG experimental data between the elderly and young people.

V. Conclusion and Future Work

Through the combination of physiological data and behavioral posture analysis, the care system can monitor the living condition of the elderly at home more effectively. The smart bracelet and the smart amulet are combined with the edge device to form an Internet of Things for data collection and pre-processing. The edge device uploads data to the cloud platform database and the local database for storage, and the cloud platform further analyzes the data and broadcasts the results through the smart speaker; the local database stores the data for long-term use as a training data set for the classifier model in the future. The system can effectively observe the fall risk of the elderly. The data stored in the cloud database can be displayed on the web page in real time, and the fall can be broadcasted through the smart speaker in real time to avoid regrets. The experimental results confirm that the system designed in this experiment is a cost-effective tool for daily care of the elderly by combining the Internet of Things with the cloud platform. Although this study has completed the design of the elderly care IoT system and conducted some experiments and studies, the hidden risks and varying degrees of cooperation in inviting elderly people to participate in the experiments have led to relatively limited experiments and data collection, and limited sampling target groups. In the future, we plan to enter more homes to conduct experiments and optimize the system function to conduct more realistic and long-term experiments as the primary improvement point. It is believed that a more comprehensive and perfect personalized care system will be established in the near future.

References

[1] R. K. Kodali, G. Swamy, and B. Lakshmi, "An implementation of IoT for healthcare," 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2015, pp. 411-416, doi: 10.1109/RAICS.2015.7488451.

[2] M. M. Khan, T. M. Alanazi, A. A. Albraikan, and F. A. Almalki, “IoT-Based Health Monitoring System Development and Analysis,” Security and Communication Networks , pp. 1-11, April 2022.

[3] P. Visutsak, and M. Daoudi, "The smart home for the elderly: Perceptions, technologies and psychological accessibilities: The requirements analysis for the elderly in Thailand," 2017 XXVI International Conference on Information, Communication and Automation Technologies (ICAT), Sarajevo, Bosnia and Herzegovina, 2017, pp. 1-6, doi: 10.1109/ICAT.2017.8171625.

[4] F. Sun, W. Zang, R. Gravina, G. Fortino, and Y. Li, “Gait-based identification for elderly users in wearable healthcare systems,” Information Fusion, pp. 134-144, 2020.

[5] M. C. H. Yeh, et al, “Artificial Intelligence–Based Prediction of Lung Cancer Risk Using Nonimaging Electronic Medical Records: Deep Learning Approach,” Journal of Medical Internet Research, vol. 23, pp. 1-13, August 2021.

[6] G. Sprint, D. J. Cook and D. L. Weeks, "Toward Automating Clinical Assessments: A Survey of the Timed Up and Go," in IEEE Reviews in Biomedical Engineering, vol. 8, pp. 64-77, 2015, doi: 10.1109/RBME.2015.2390646.

[7] S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan, "Estimation of IMU and MARG orientation using a gradient descent algorithm," 2011 IEEE International Conference on Rehabilitation Robotics, Zurich, Switzerland, 2011, pp. 1-7, doi: 10.1109/ICORR.2011.5975346.

[8] "Understanding Blood Pressure Readings | American Heart Association," Heart, Feb 2023. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

[9] "Timed Up and Go Test (TUG) - Physiopedia," Physio-pedia, Feb 2023. https://www.physio-pedia.com/Timed_Up_and_Go_Test_(TUG)

 ISBN: 978-1-68558-056-8

A Hypergraph Approach for Logic-based Abduction

Qiancheng Ouyang
LISN, CNRS
Université Paris-Saclay
email: oyqc@lisn.fr

Tianjiao Dai
LISN, CNRS
Université Paris-Saclay
email: dai@lisn.fr

Yue Ma
LISN, CNRS
Université Paris-Saclay
email: yue.ma@lisn.fr

***Abstract*—Abduction reasoning, which finds possible hypotheses from existing observations, has been studied in many different areas. We consider an abduction problem that takes into account a user's interest. We propose a new approach to solving such an abduction problem based on a hypergraph representation of an ontology and obtain a linear algorithm for a description logic.**

***Index Terms*—Abduction; Hypergraph; Description Logic**

I. Introduction

Abduction reasoning aims to generate a possible hypothesis for a given observation. Abduction has been applied in many artificial intelligence (AI) areas, such as machine learning, logical programming, and statistical relational AI [7].

We focus on *logical-based abduction* [4] over *description logic ontologies*. Here, ontologies consist of *axioms* that state the relationship of different *concepts* and *relationships* over a specific domain. Then, our abduction problem consists of three parts: (i) a given background knowledge (i.e., an existing ontology $\mathcal{O}$); (ii) a set of hypotheses (i.e., a set of axioms $\mathcal{H}$) and (iii) a given conclusion (i.e., a single axiom). There have been many studies of abduction over different ontologies, such as the complexity of abduction over $\mathcal{EL}$ [2] and their application to repairing ontologies [8], abduction over $\mathcal{EL}$ by translation to first-order logic [5], forgetting-based abductive reasoning over expressive ontology $\mathcal{ALC}$ [3], and signature-based abduction over more expressive $\mathcal{ALCOI}\mu$ [6].

We propose a new solution (section IV) of abduction over a special $\mathcal{EL}$-ontology (free of role restrictions) in Section III, based on a hypergraph representation of ontologies.

II. Preliminaries

An *ontology* $\mathcal{O}$ is a set of *axioms* of the form $A_1 \sqcap \cdots \sqcap A_n \sqsubseteq B$, where A_i, B are called *concepts*. An interpretation $\mathcal{I} = \langle \Delta^I, \cdot^I \rangle$ consists of a non-empty domain Δ^I and a mapping $\cdot^I$ that maps each concept to a subset $A^I \subseteq \Delta^I$. A model of $\mathcal{O}$ is an interpretation that for each $A_1 \sqcap \cdots \sqcap A_n \sqsubseteq A \in \mathcal{O}$, we have $A_1^{\mathcal{I}} \cap \cdots \cap A_n^{\mathcal{I}} \subseteq A^{\mathcal{I}}$. We say $\mathcal{O} \models A'_1 \sqcap \cdots \sqcap A'_n \sqsubseteq B'$ iff for any models $\mathcal{I}$ of $\mathcal{O}$, we have $(A'_1)^{\mathcal{I}} \cap \cdots \cap (A'_n)^{\mathcal{I}} \subseteq (B')^{\mathcal{I}}$.

A *(directed) hypergraph* $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$ consists of a node set $\mathcal{V} = \{v_1, v_2, \cdots, v_n\}$ and a hyperedge set $\mathcal{E} = \{e_1, e_2 \cdots, e_m\}$, where $e_i = \langle T(e_i), f(e_i) \rangle$ with $T(e_i) \subseteq \mathcal{V}$ being a subset and $f(e_i) \in \mathcal{V}$ being a node. Note that a classical hyperedge can have multiple nodes in its head, which we require to be a singleton for computing abduction.

Definition 1 ([1])**.** Given a hypergraph $\mathcal{H} = \{\mathcal{V}, \mathcal{E}\}$, assume $S \subseteq \mathcal{V}$ and $v \in \mathcal{V}$. A *hyperpath* from S to v is a sequence $h = [e_1, e_2, \cdots, e_n]$ of hyperedges such that (i) $f(e_n) = \{v\}$; (ii) for $i = 1, \cdots, n$, $T(e_i) \subseteq S \cup \{f(e_1); \cdots, f(e_{i-1})\}$; (iii) for $i = 1, \cdots, n$, $f(e_i) \in \bigcup_{i<j\leq n} T(e_j)$.

III. Abduction Problem

We consider an abduction problem that takes into account a user's interests represented by a set of concepts Σ.

Definition 2. An abduction problem is a tuple

$$\langle \mathcal{O},\ \Sigma,\ A_1 \sqcap \cdots \sqcap A_n \sqsubseteq B \rangle,$$

where $\Sigma = \{A', B', \cdots\}$ is a set of concept names. A *solution* of this problem is a (minimal) ontology

$$\mathcal{H} = \{A'_1 \sqcap \cdots \sqcap A'_n \sqsubseteq B' \mid A'_i, B' \in \Sigma, n \geq 0\}$$

such that $\mathcal{O} \cup \mathcal{H} \models A_1 \sqcap \cdots \sqcap A_n \sqsubseteq B$. A solution $\mathcal{H}$ is called a *hypothesis* with respect to Σ.

Example 1. Let an ontology $\mathcal{O}_0$ be:

$$peopleWithDiploma \sqsubseteq doctor$$
$$peopleHasPaper \sqsubseteq researcher$$
$$doctor \sqcap employeeWithUniversityChair \sqsubseteq professor$$

$\mathcal{O}_0$ can not derive the following axiom α_0:

$$\alpha_0 : doctor \sqcap employeeWithUniversityChair \sqsubseteq researcher$$

although it should be true. Consider $\Sigma_0 = \{professor,\ peopleHasPaper\}$. If we add a hypothesis $\mathcal{H}_0 = \{professor \sqsubseteq peopleHasPaper\}$, we have $\mathcal{O}_0 \cup \mathcal{H}_0 \models \alpha_0$. Therefore, $\mathcal{H}_0$ is a solution of the abduction problem $\mathcal{A}_0 = \langle \mathcal{O}_0,\ \Sigma_0,\ \alpha_0 \rangle$. It is clear that $\mathcal{H}_0$ is also a minimal solution to the abduction problem. But there is no solution to $\mathcal{A}_0$ if $\Sigma_0 = \{professor,\ peopleWithDiploma\}$.

IV. A Hypergraph-based Algorithm

We now present a method of finding a (minimal) solution to the abduction problem using hypergraphs.

Definition 3. For each set $\mathcal{O}$ of axioms, we define a hypergraph $H_{\mathcal{O}} = (\mathcal{N}_h, \mathcal{E}_h)$, where $\mathcal{N}_h := \{N_{A'} \mid A' \in \mathsf{N_C}\}$ and

$$\mathcal{E}_h := \{\{N_{A'_1}, \cdots, N_{A'_n}\} \to N_{A'} \mid A'_1 \sqcap \cdots \sqcap A'_n \sqsubseteq A' \in \mathcal{O}\}$$

Example 2 (Example 1 cont'd)**.** By definition, the hypergraph $H_{\mathcal{O}_0}$ of $\mathcal{O}_0$ is shown in Figure. 1. Now, we add an edge

 ISBN: 978-1-68558-056-8

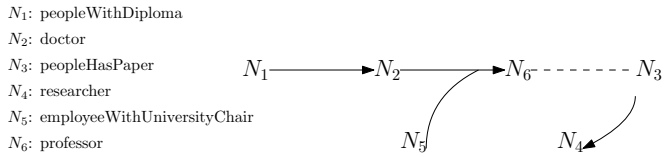

Fig. 1: The hypergraph representation $H_{\mathcal{O}_0}$ of $\mathcal{O}_0$ in Example 1

$\{N_6\}\rightarrow N_3$ to the hypergraph $H_{\mathcal{O}_0}$. Then, we can find a hyperpath h from $\{N_2,\ N_5\}$ to N_4:

$$h = [\{N_2, N_5\}\rightarrow N_6,\ \{N_6\}\rightarrow N_3,\ \{N_3\}\rightarrow N_4]$$

Theorem 1. Given an ontology $\mathcal{O}$ and its associated hypergraph $H_{\mathcal{O}}$, an ontology $\mathcal{H}$ is a (minimal) solution to the abduction problem $\langle \mathcal{O}, \Sigma, A_1 \sqcap \cdots \sqcap A_n \sqsubseteq B\rangle$ iff $H_{\mathcal{H}}$ is a (minimal) hypergraph such that (i) All nodes in $H_{\mathcal{H}}$ are of the form $N_A,\ A \in \Sigma$, and (ii) There exists a hyperpath from $N_{A_1}, \cdots, N_{A_n}$ to N_B in $H_{\mathcal{O}} \cup H_{\mathcal{H}}$.

Example 3 (Example 1 cont'd)**.** By Theorem 1, to solve the abduction problem $\mathcal{A}_0$, it is enough to find an $H_{\mathcal{H}}$ such that there exists a hyperpath from $\{N_2, N_5\}$ to N_4 in $H_{\mathcal{O}_0} \cup H_{\mathcal{H}}$. The hypergraph $H_{\mathcal{H}}$ consists of a single edge $\{N_6\}\rightarrow N_3$ satisfying the requirement, leading to the hyperpath given in Example 2 as the minimal solution of the problem.

Before stating our main Algorithm 2, we define a property of *saturation* for a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and $V \subset \mathcal{V}$. We define $U \subset V$ to be *saturated* (under V) if there exists $e \in \mathcal{E}$ such that $T(e) = U$ and $f(e) \notin V$. For example, in Fig. 1, if we have $V = \{N_1, N_2, N_5\}$, then $\{N_1\}$ and $\{N_2, N_5\}$ are saturated under V, while other subsets of V are not. Algorithm 1 finds all vertices approachable from V in run-time $O(|\mathcal{E}|)$.

Proposition 1. For a hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ and $V \subset \mathcal{V}$, $v \in Span(\mathcal{V}, \mathcal{E}, V)$ iff. there is a hyperpath from V to v.

Algorithm 1: $Span(\mathcal{V}, \mathcal{E}, V)$

input : hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$, set $V \subset \mathcal{V}$
output: $W \subset \mathcal{V}$ of all vertices spanned from V

```
1  W = V.
2  U = {U, U is saturated under V}.
3  while U ≠ ∅ do
4  |  choose U ∈ U,
5  |  while there exists v ∈ V \ W such that (W, v) ∈ E
   |   do
6  |  |  put v into W;
7  |  |  put all saturated sets containing v into U.
8  |  end
9  |  remove U.
10 end
11 return W
```

In Algorithm 2, we first check if v can be directly reached by V (Line 1-3). Then, we check if the aiming hypergraph exists (Line 4-11). These 2 steps have run-time $O(|\mathcal{E}|)$. In the minimizing step, for each $e \in \mathcal{E}'$, we check only once if e can be deleted. Hence, the total run-time is $O(|\Sigma||\mathcal{E}|)$.

Algorithm 2:

input : hypergraph $H = (\mathcal{V}, \mathcal{E})$, $\Sigma \subset \mathcal{V}$, $S \subset \mathcal{V}$, $v \in \mathcal{V}$
output: hypergraph $\mathcal{H}$ on Σ

```
1  V = Span(V, E, S).
2  if v ∈ V then
3  |  return empty graph.
4  else
5  |  if Σ ⊂ V or Σ ∪ V = ∅ return non-existence.
6  |  Σ \ V = {v_1, ..., v_m},
7  |  choose m hyper-edges E' = {e_1, ..., e_m} where
   |   T(e_i) ⊂ Σ ∩ V and f(e_i) = v_i for 1 ≤ i ≤ m.
8  |  V = V ∪ Σ.
9  end
10 if v ∉ Span(V, E ∪ E', V) then
11 |  return non-existence.
12 else
13 |  minimize E' (check if there exists e ∈ E' such that
   |   E' − e satisfies until we get a minimal size).
14 |  return H = (Σ, E').
15 end
```

We explain Algorithm 2 via the following example.

Solution of Example 3. (via Algorithm 2)

1) $H = H_{\mathcal{O}}, \Sigma = \{N_3, N_6\}, S = \{N_2, N_5\}, v = N_4$.
2) Line 1: $V = Span(\mathcal{V}, \mathcal{E}, S) = \{N_2, N_5, N_6\}$.
3) Line 7: $\mathcal{E}' = \{\{N_6\} \rightarrow N_3\}$.
4) Line 10: $v \in Span(\mathcal{V}, \mathcal{E} \cup \mathcal{E}', V = \mathcal{V})$.
5) Line 13: we see $\{\{N_6\} \rightarrow N_3\}$ cannot be deleted. It returns $\mathcal{H} = (\Sigma, \mathcal{E}')$.

Theorem 2. For Algorithm 2, the output $\mathcal{H}$ is a minimal hypergraph satisfying the conditions (i) and (ii) in Theorem 1.

V. Conclusion

In this work, we introduce a hypergraph-based algorithm for solving abduction problems over $\mathcal{EL}$-ontologies that do not have role restrictions, which have a linear time complexity w.r.t. the size of the input ontology. As for future work, we plan to implement our algorithm and extend it to handle general $\mathcal{EL}$-ontologies with role restrictions, as well as more expressive ontologies such as $\mathcal{ALC}$.

Acknowledgment. We thank Hui Yang for bringing our attention to the topic and the discussion with us.

References

[1] A. Giorgio and L. Luigi, "Directed hypergraphs: Introduction and fundamental algorithms-a survey", Theoretical Computer Science, vol. 658, pp. 293–306, 2017.
[2] M. Bienvenu, "Complexity of abduction in the EL family of lightweight description logics", Proc. of KR'08, 2008, pp. 220–230.
[3] W. Del-Pinto and R. A. Schmidt, "Abox abduction via forgetting in ALC", Proc. of AAAI'19, 2019, pp. 2768–2775.
[4] T. Eiter and G. Gottlob, "The complexity of logic-based abduction", J. ACM, vol. 42 (1), 1995, pp. 3–42.
[5] F. Haifani, P. Koopmann, S. Tourret, and C. Weidenbach, "Connection-minimal abduction in *EL* via translation to FOL", Proc. of IJCAR'22, 2022, pp. 188–207.

 ISBN: 978-1-68558-056-8

[6] P. Koopmann, W. Del-Pinto, S. Tourret, and R. A. Schmidt “Signature-based abduction for expressive description logics”, Proc. of KR’20, 2020, pp. 592–602.
[7] Sindhu V. Raghavan, “Bayesian Abductive Logic Programs: A Probabilistic Logic for Abductive Reasoning”, Statistical Relational Artificial Intelligence, Proc. of IJCAI’11, 2011, pp. 2840–2841.
[8] F. Wei-Kleiner, Z. Dragisic, and P. Lambrix, “Abduction framework for repairing incomplete EL ontologies: Complexity results and algorithms”, Proc. of AAAI’14, 2014, pp. 1120–1127.

 ISBN: 978-1-68558-056-8

Robust Representations in Deep Learning

Shu Liu
The Computational and Data Science PhD Program)
Middle Tennessee State University
Murfreesboro, TN 37132, USA
email: sl6b@mtmail.mtsu.edu

Qiang Wu
Department of Mathematical Sciences
Middle Tennessee State University
Murfreesboro, TN 37132, USA
email: qwu@mtsu.edu

***Abstract*—Deep neural networks are playing increasing roles in machine learning and artificial intelligence to handle complicated data. The performance of deep neural networks depends highly on the network architecture and the loss function. While the most common choice for loss function is the squared loss for regression analysis it is known to be sensitive to outliers and adversarial samples. To improve the robustness, we introduce the use of the correntropy loss to the implementation of deep neural networks. We further split the neural network architecture into a feature extraction component and function evaluation component and design four two-stage algorithms to study which component is more impacted by the use of the robust loss. The applications in several real data sets indicates that the robust deep neural networks can efficiently generate robust representations of complicated data and the two-stage algorithms are consistently more powerful than their one-stage counterparts.**

***Index Terms*—deep neural network, LSTM, correntropy loss, robustness**

I. Introduction

The history of artificial neural network dates back at least to the perceptron invented by Rosenblatt [1]. After more than half a century's development, artificial neural networks are playing increasing roles in modern machine learning and artificial intelligence applications. Although it has been proved that the feed-forward neural network with a single hidden layer and sigmoid activation function can approximate any arbitrarily complex continuous mapping with arbitrary precision [2]–[4], more and more evidence shows that deep neural networks could more powerful [5]–[7]. In the past decade, along with the fast development of hardwares and computational power, deep neural networks have been successfully applied to computer vision, speech and audio recognition, language processing,customer relationship management, and many other fields.

The performance of deep neural networks highly depends on the network architecture and the loss function. In the context of supervised learning, three main neural network architectures are popularly used. The fully connected neural networks, the convolutional neural networks, and the recurrent neural networks. While the fully connected neural networks could be more widely applied to any structured data set, convolutional neural networks have been shown powerful for image analysis and computer vision, and recurrent neural networks have been successfully used in time series data, such as speech recognition and natural language processing. Regarding the loss functions, the least square loss and cross-entropy loss are commonly used for regression analysis and classification tasks, respectively.

Robustness concerns may arise in practice when the data is contaminated by outliers. For instance, Rare body poses in human pose estimation, unlikely facial point position in facial landmark detection, imprecise ground-truth annotations,and label misspecification all may result abnormal samples in image processing, outliers may present in financial data due to heavy tailed distributions, and system shock could produce extreme and erratic values in signal processing. In these situations, there are needs to develop deep learning robust approaches because least square loss and cross entropy are well known to be unrobust and sensitive to outliers. Some efforts have been done in the literature, e.g., [8]–[10]. While there are multiple ways to promote algorithm robustness, the most common approach is to adopt a robust loss to train the neural networks and typical examples include the Huber's loss, the Tukey's biweight loss, the truncated least square loss, the Cauchy loss, and the correntropy loss.

In this paper, we propose to build robust deep neural networks by the correntropy loss. We will not only verify its effectiveness, but also thoroughly explore where robustness comes from. To be precise, we recall that a deep neural network is usually regarded as the combination of two components, the feature extraction component and the function evaluation component. We are particularly interested in the impact of the robust loss on these two components and will evaluate if both components are impacted or one is more impacted than the other. In order to make fair comparisons, we design two-stage algorithms and conduct a comparative study on several real world applications. The results indicate two surprising findings: (1) While the robust loss may impact both components, it seems the feature extraction part is more impacted. In other words, robust deep neural networks incline to produce more robust feature representations. (2) The two stage implementation of deep neural networks are always more powerful than the one stage approaches, regardless of the loss functions used.

The rest of the paper is organized as follows. In Section II, we introduce the deep neural networks. In Section III, we introduce the two stage algorithms to build deep neural networks. In Section IV, we apply the proposed algorithms to real world applications and present the results. We close with

 ISBN: 978-1-68558-056-8

conclusions and discussions in Section V.

II. Robust deep neural networks

In this section, we introduce two robust deep neural networks, the robust deep feed-forward neural network and the robust long short-term memory neural network.

A. The correntropy loss

Outliers are abnormal or extreme values in the data that significantly deviate from the rest of the observations. The appearance of a small amount of outliers may reduce the ability of statistical inference and hurt the predictive performance of machine learning models. While outlier detection and removal are used sometimes [11] [12], a more common approach for supervised learning is to adopt a robust loss function.

Given a data set $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$ with $\mathbf{x}_i \in \mathbb{R}^d$ representing the vector of d explanatory variables and y_i the response, the well-known least square method aims to minimize the the mean square error

$$\min_f \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i = f(\boldsymbol{x}_i)$ is the prediction of the response variable y_i by a hypothetical function f. The minimization process is conducted over a set of hypothesis functions which could be the set of linear functions in the traditional multiple linear regression or the set of nonlinear functions represented by a neural network architecture. A main advantage of the least square method is its optimality when the noise follows a Gaussian distribution while the main criticism is the lack of robustness when the Gaussianity is violated by outliers of heavy tailed distributions.

The use of correntropy loss for robustness has a long history. Its variant forms have been proposed as goodness-of-fit measures in the literature under different terminologies, such as the Welsch's loss [13], the inverted Gaussian loss [14], the exponential squared loss [15], the reflected normal loss [16], and the maximum correntropy criterion [17] [18]. In this paper we adot the form proposed in [19]:

$$\mathcal{L}(y_i, \hat{y}_i) = \sigma^2 \left(1 - \exp\left(\frac{(y_i - \hat{y}_i)^2}{\sigma^2}\right)\right),$$

where $\sigma > 0$ is a tunable parameter that trades off the robustness and fitting errors. A robust regression approach minimizes the mean correntropy loss

$$\min_f \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i).$$

There exist not only numerous empirical evidences in the literature to show the ability of correntropy loss to promote robustness, the theoretical guarantees were also investigated in recent studies [20]–[23].

B. Robust Deep Feed-forward Neural Network

A fully connected feed-forward neural network (FNN) consists of three parts: the input layer, the hidden layers, and the out put layer. The input layer has d neurons, representing the d features of the input data. Mathematically, for an input vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, the jth node of the input layer is given by $h_j^0 = x_j$. A FNN can have one or multiple hidden layers. It is called a shallow neural network if there is only one hidden layer and a deep neural network if there are two or more hidden layers. As we have already mentioned above, although a shallow neural network has the ability to well approximate arbitrarily complicated functions, there are both empirical and theoretical evidence that deep neural networks are more powerful in real applications. For hidden layers, the value of each neuron is computed from all neurons of the precedent layer by an affine linear mapping and an activation function: let $h_{l,j}$ denote the j-th node of the l-th layer and d_l be the number of neurons in the l-the layer. Then

$$h_{k,j} = a\left(\sum_{j=1}^{d_{l-1}} w_{l,j,k} h_{l-1,j} + b_{l,j}\right),$$

where $w_{l,j,k} \in \mathbb{R}$, $b_{l,j} \in \mathbb{R}$, and a is an activation function. The most popular choices for the activation function include the sigmoid function, the hyperbolic tangent function, and the rectified linear activation function (ReLU). The output layer of an FNN will produce predicted values for the response variable. For a regression analysis with a scalar response variable, the output layer contains one neuron by linear function of the last hidden layer:

$$\hat{y} = \sum_{j=1}^{d_L} w_{L,j} h_{L,j} + b_L,$$

where L denotes the number of hidden layers. Figure 1 shows an example of FNN with two hidden layers and a single output. The number of output neurons can be more than one for vector valued regression analysis or classification problems.

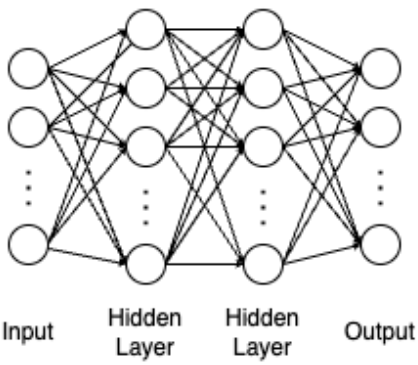

Fig. 1. A deep Feed-forward neural network with two hidden layers

The training of the weight and bias parameters of FNN requires a loss function to measure the error when $\hat{y}_i$ is used to predict the true response value y_i for each observation $\mathbf{x}_i$. In

this paper, we implement the robust deep feed-forward neural network (RFNN) by minimizing the mean correntropy loss.

C. Robust Long Short-Term Memory Neural Network

Long Short-Term Memory Neural Network (LSTM) was first introduced by Hochreiter and Schmidhuber [24]. In 1999, Felix et al. [25] introduced a forget gate mechanism based on Hochreiter and Schmidhuber's work, which enables LSTM to reset its own state to avoid network crashes. Several variant models were proposed since then. LSTM is a special kind of recurrent neural network and has been shown effective for time series analysis, speech recognition, language translation, and natural language processing due to its ability to memorize long and short term information.

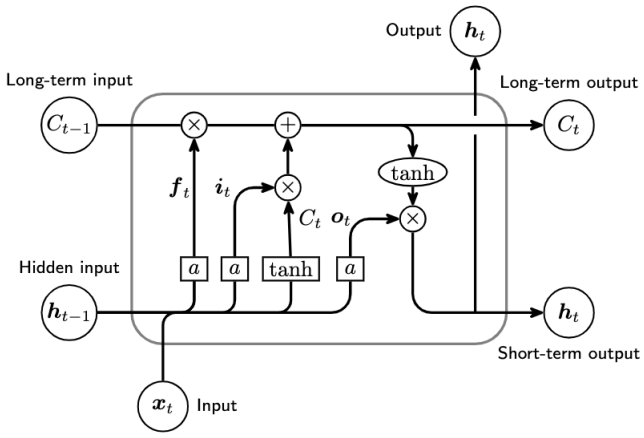

Fig. 2. LSTM network structure

Figure 2 shows the core structure of LSTM. The long term information is stored in the cell states and pass through the entire chain system of LSTM from beginning to end. Three layers will used to decide what information will be removed and what information will be added. The forget gate layer generate $\boldsymbol{f}_t$ of a vector of values between 0 and 1 based on the current input $\boldsymbol{x}_t$ and previous moment output h_{t-1} via affine linear transforms and sigmoid activation function:

$$\boldsymbol{f}_t = a\left(W_f \boldsymbol{x}_t + U_f \boldsymbol{h}_{t-1} + b_f\right)$$

where $W_{\boldsymbol{f}}$ is a matrix of weight coefficients and b_f a sequence of biases. The values of $\boldsymbol{f}_t$ determine the percentage of information in C_{t-1} that are allowed to pass through, in other words, the information $(1-\boldsymbol{f}_t) * C_{t-1}$ will be removed or "forgotten", where $*$ denotes element wise multiplication operator . A tanh layer will produce values representing candidate information :

$$\tilde{C}_t = \tanh\left(W_C \boldsymbol{x}_t + U_C \boldsymbol{h}_{t-1} + b_i\right),$$

and the input gate layer produced $\boldsymbol{i}_t$, again a vector of values between 0 and 1, by

$$\boldsymbol{i}_t = a\left(W_i \boldsymbol{x}_t + U_i \boldsymbol{h}_{t-1} + b_i\right)$$

to decide the percentage of candidate information $\tilde{C}_t$ to be added to the cell state. The cell state is then updated by

$$C_t = \boldsymbol{f}_t * C_{t-1} + \boldsymbol{i}_t * \tilde{C}_t.$$

After the cell state is updated, LSTM will use the output gate will first compute

$$\boldsymbol{o}_t = a(W_o \boldsymbol{x}_t + U_o \boldsymbol{h}_{t-1} + b_o)$$

and then use $\tanh(C_t)$ as weight coefficients to generate the output

$$\boldsymbol{h}_t = \boldsymbol{o}_t * \tanh(C_t),$$

which will be further used to produce the prediction of the response variable. In this paper a linear function

$$\hat{y}_t = \boldsymbol{w}^\top \boldsymbol{h}_t + b$$

is used. Given a sequence of time series $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$ and corresponding response series $y_1, y_2, \ldots, y_T$, the robust LSTM will minimize the mean correntropy loss

$$\frac{1}{T}\sum_{i=1}^{T} \mathcal{L}(y_t, \hat{y}_t)$$

over the historical period to estimate the network parameters.

III. Two Stage Algorithms

When the data are contaminated by outliers or are skewed and have heavy tails, robust algorithms are supposed to perform better. As we will see in Section IV below, our robust deep learning algorithms are indeed superior as expected when they are applied to real applications with robustness concerns.

The success of deep neural networks have been largely attributed to its ability to extract information from the complicated data. Therefore, it is commonly recognized that a deep neural network can be split into two parts: a feature extraction part and a function evaluation part, where the first part extract relevant features from the input data and second part use the features to build a decision function. As we are able to show the superiority of robust deep learning algorithms, we want to explore further and answer the questions that (1)"whether the robust deep learning algorithms promote robustness of feature extraction and lead to robust representation?" and (2) "which part of the network is more impacted by the use of robust loss?" To answer these questions, we propose a series of two stage algorithms.

For FNN and RFNN, we regard the part from the input to the last hidden layer as the feature extraction process and from the last hidden layer to output as the function evaluation part. We first run FNN and robust FNN to build two neural networks. Then we extract the features and run the linear regression with either least square (LS) approach or the robust regression (RR) with correntropy loss. This leads to four two-stage algorithms: FNN+LS, FNN+RR, RFNN+LS, and RFNN+RR, where the FNN+LS uses the features extracted from FNN and least square regression to predict the response variable, FNN+RR uses the features extracted from FNN and robust regression, RFNN+LS uses the features extracted from RFNN and least

square regression, and RFNN+LS uses the features extracted from RFNN and robust regression. If RFNN+LS outperforms FNN+LS, we are able to conclude that the feature extraction process by RFNN is robust. Otherwise the correntropy loss does not robustify the feature extraction process. On the other hand, if RFNN+RR outperforms RFNN+RR, the correntropy loss plays a role in the function evaluation process.

In LSTM we regards the values in output h_t are the features extracted from the input x_t and previous information. We can similarly design four two-stage algorithms to study the robust representation ability of RLSTM.

IV. Applications

In this section, we apply our algorithms to real-world applications and illustrate their effectiveness.

A. Airfoil Data Set

Airfoil Self-Noise Data [26] is a NASA data set which obtained from a series of aerodynamic and acoustic tests of two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel. It is a multivariate data set with 5 attributes (Frequency, Angle of attack, Chord length, Free-stream velocity and Suction side displacement thickness) measuring scaled sound pressure level. The data set contains 1503 instances. Figure 3 shows the histogram of the response variable. It is clearly left skewed.

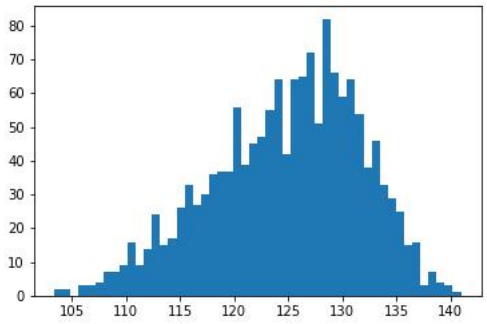

Fig. 3. Histogram of response variable for airfoil data

We randomly sampled 50% (Different split ratios do not significantly affect the final results) of the data as training set and remaining data as test set. We apply FNN, RFNN, and all four two-stage algorithms to build models and predict on the test set. The neural network contains two hidden layers with each hidden layer containing 64 hidden neurons. Tensorflow in Python is used to train the neural network with both the epoch and batch sizes selected as 50. The parameter σ is not sensitive and a value of $\sigma = 10$ is used. The experiments are repeated 50 times. The average mean absolute error (MAE) and the standard error (SE) of all six approaches are reported in Table I.

Firstly, we see that RFNN outperforms FNN, indicating the use of correntropy loss improves the robustness of the neural network estimation. Next, RFNN+LS outperforms FNN+LS and RFNN+RR outperforms FNN+RR, that is, when the same regression approach is used, using features from RFNN is always better. This means that the features extracted by RFNN is more informative and therefore we can claim that the RFNN helps to extract features more robustly. Thirdly, FNN+RR outperforms FNN+LS and RFNN+RR also outperforms RFNN+LS, but the improvement is not significant. This means that once the features have been extracted, further use of robust loss in the regression step does not help much. Lastly, it is surprising to see that FNN+LS outperforms FNN and RFNN+RR outperforms RFNN, indicating that when the same loss function is used, the two-stage algorithms are consistently better than traing the neural network directly. Further more, if we change the loss function in the second regression stage, the two-stage algorithms are still better.

B. Boston Housing Data Set

Boston Housing Data Set contains information collected by the U.S Census Service concerning housing in the area of Boston Massachusetts. It is a multivariate data set with 13 attributes measuring the median value of owner-occupied homes. It contains 506 instances. Figure 4 show the histogram of the home values. We can see outliers on the right end.

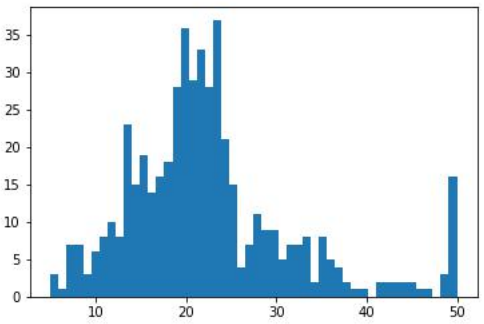

Fig. 4. Histogram of home values in Boston housing data

This data has been build in the sklearn module in Python where the training set and test set have been automatically separated. We merged them together and then randomly sampled 50% as training set and put the remaining data into the test set. The hyperparameters and analysis process are the same as the experiment for Airfoil data. The results are shown in Table I. The findings are very similar to the application in Airfoil data: RFNN is more robust than FNN. The use of correntropy seems play more roles in robust feature extraction while less roles in function evaluation. A follow-up regression stage helps further improve the performance.

C. Agroecosystem Data

This data is collected by Dr. Song Cui at the MTSU Department of Agriculture. It contains carbon, water and energy fluxes of a cool-season dominated pasture ecosystem

 ISBN: 978-1-68558-056-8

for 159 days from September 2, 2016 to May 3, 2017. For each day, there are 48 data points with each data point a summary of the relevant information in half an hour. If the data is complete, there should be 7632 data points in total. But due to power outage or system failure, a small number of data points are missing and the data actually contains 7265 data points. In this analysis, 12 driver variables that measure the radiation, humidity, temperature, wind, and some other features, were used to predict evapotranspiration flux. Outliers due to system shocks can be clearly seen from plot in Figure 5.

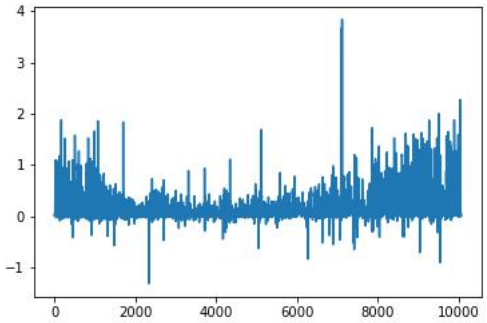

Fig. 5. Plot of evapotranspiration flux

The deep neural network with two hidden layers and 64 hidden neurons for each layer is again used for the analysis. A 50%-50% split is again used for training and test data split. The RFNN used $\sigma = 50$. The experiments are repeated 50 times and the average MAEs are reported for all six approaches in Table I. Similarly, we find the RFNN is able to lead to robust representation of the data and two-stage algorithms are more powerful.

D. CSI 300 Data

Per Wikipedia [27], the CSI 300 is a capitalization-weighted stock market index designed to replicate the performance of the top 300 stocks traded on the Shanghai Stock Exchange and the Shenzhen Stock Exchange. it is a gauge of Chinese stock market. In this experiment, the trading information (opening price, closing price, highest price, lowest price and volume) of CSI 300 from January 3, 2017 to December 29, 2018 (excluding weekends and holidays) used. Figure 6 show the closing prices. Stock prices are typical examples of time series involving sudden changes and abnormal values due to the reaction to government policies, economical indicators, and sentiments.

The analysis aims to predict the closing price based on the opening price, the previous day's opening price, closing price, highest price, lowest price and volume. As the price data can be viewed as time series, LSTM is appropriate. The results by six approaches are reported in Table II. Although a different network architecture is used in this experiment, the findings are still consistent with previous applications. The only difference is that the use of robust regression in

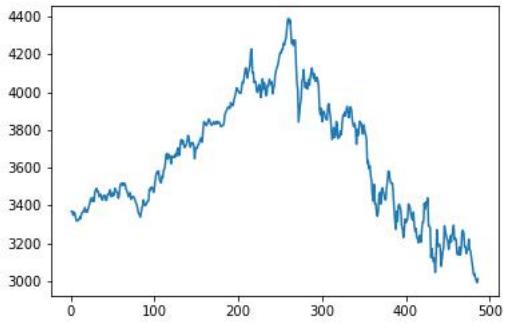

Fig. 6. Histogram of closing prices in CSI 300

TABLE I
MAE ON AIRFOIL AND BOSTON HOUSING DATA

Method	Airfoil	Boston Housing	Agroecosystem
FNN	0.2366 (0.0023)	0.2761 (0.0045)	0.1877 (0.0026)
FNN+LS	0.2235 (0.0016)	0.2719 (0.0040)	0.1720 (0.0007)
FNN+RR	0.2223 (0.0016)	0.2702 (0.0040)	0.1719 (0.0007)
RFNN	0.2279 (0.0018)	0.2706 (0.0035)	0.1779 (0.0014)
RFNN+LS	0.2173 (0.0014)	0.2681 (0.0035)	0.1714 (0.0009)
RFNN+RR	0.2161 (0.0014)	0.2669 (0.0035)	0.1713 (0.0008)

the second stage play more roles, as is evidenced the better performance of LSTM+RR and RLSTM+RR than that of LSTM+LS and RLSTM+LS, respectively.

TABLE II
MAE ON AIU AND CSI300 DATA

Method	CSI300
LSTM	0.2221 (0.0007)
LSTM+LS	0.2133 (0.0007)
LSTM+RR	0.2016 (0.0006)
RLSTM	0.2197 (0.0008)
RLSTM+LS	0.2116 (0.0007)
RLSTM+RR	0.2012 (0.0007)

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed to implement robust deep neural networks by using the correntropy loss and four two-stage algorithms. Simulation studies on four real data applications show that the robust deep neural networks are more efficient to handle data with outliers or skewed. Moreover, the robust deep neural networks are able to efficiently extract more informative features, indicating the entropy loss plays more roles in robust representation of the data.

The superiority of two-stage algorithms is a serendipity. The original motivation of these algorithms is to study how the robust loss plays roles in the network construction process, not for better performance. The simulations surprisingly show that all two-stage algorithms are consistently better than their one-stage counterparts, regardless the loss function used.

In this paper we have focused on the fully connected deep feed-forward neural networks and the LSTM for regression analysis. Convolutional Neural Network (CNN) is another

 ISBN: 978-1-68558-056-8

representative algorithms of deep learning is particularly confidential for its excellent performance in image processing and computer vision. We can similarly develop robust convolutional neural network. However, it seems CNN is more widely used in classification problems while the correntropy loss is more appropriate for regression analysis. So, we have omitted the study of robust CNN in this paper. But the idea of two-stage training is promising and it would be interesting to develop two-stage CNN algorithms with appropriate classification loss functions, such as the cross entropy loss.

ACKNOWLEDGMENT

This work is partially supported by NSF (DMS-2110826).

REFERENCES

[1] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[2] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[3] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural networks*, vol. 2, no. 3, pp. 183–192, 1989.

[4] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[5] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press Cambridge, 2016, http://www.deeplearningbook.org.

[8] B. Chen, L. Xing, H. Zhao, N. Zheng, and J. C. Principe, "Generalized correntropy for robust adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3376–3387, 2016.

[9] I. Goodfellow, P. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Communications of the ACM*, vol. 61, no. 7, pp. 56–66, 2018.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJzIBfZAb

[11] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[12] L. Yin, Q. Wu, and D. Hong, "Statistical methods and software package for medical trend analysis in health rate review process," *J Health Med Inform*, vol. 7, no. 219, 2016.

[13] J. E. Dennis Jr and R. E. Welsch, "Techniques for nonlinear least squares and robust regression," *Communications in Statistics-Simulation and Computation*, vol. 7, no. 4, pp. 345–359, 1978.

[14] K. P. Körding and D. M. Wolpert, "The loss function of sensorimotor learning," *Proceedings of the National Academy of Sciences*, vol. 101, no. 26, pp. 9839–9842, 2004.

[15] X. Wang, Y. Jiang, M. Huang, and H. Zhang, "Robust variable selection with exponential squared loss," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 632–643, 2013.

[16] F. A. Spiring, "The reflected normal loss function," *Canadian Journal of Statistics*, vol. 21, no. 3, pp. 321–330, 1993.

[17] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: properties and applications in non-Gaussian signal processing," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[18] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Science & Business Media, 2010.

[19] Y. Feng, X. Huang, L. Shi, Y. Yang, J. A. Suykens *et al.*, "Learning with the maximum correntropy criterion induced losses for regression." *J. Mach. Learn. Res.*, vol. 16, no. 30, pp. 993–1034, 2015.

[20] Y. Feng, J. Fan, and J. A. Suykens, "A statistical learning approach to modal regression," *Journal of Machine Learning Research*, vol. 21, no. 2, pp. 1–35, 2020.

[21] Y. Feng and Y. Ying, "Learning with correntropy-induced losses for regression with mixture of symmetric stable noise," *Applied and Computational Harmonic Analysis*, vol. 48, no. 2, pp. 795–810, 2020.

[22] Y. Feng and Q. Wu, "Learning under $(1 + \epsilon)$-moment conditions," *Applied and Computational Harmonic Analysis*, vol. 49, no. 2, pp. 495–520, 2020.

[23] ——, "A framework of learning through empirical gain maximization," *Neural Computation*, vol. 33, no. 6, pp. 1656–1697, 2021.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[26] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[27] Wikipedia contributors, "CSI 300 index," accessed December 11, 2022. [Online]. Available: https://en.wikipedia.org/wiki/CSI_300_Index

 ISBN: 978-1-68558-056-8

Graph Data Models and Relational Database Technology

Malcolm Crowe
Emeritus Professor, Computing Science
University of the West of Scotland
Paisley, United Kingdom
Email: Malcolm.Crowe@uws.ac.uk

Fritz Laux
Emeritus Professor, Business Computing
Universität Reutlingen
Reutlingen, Germany
Email: Fritz.Laux@reutlingen-university.de

***Abstract*—Recent work on database application development platforms has sought to include a declarative formulation of a conceptual data model in the application code, using annotations or attributes. Some recent work has used metadata to include the details of such formulations in the physical database, and this approach brings significant advantages in that the model can be enforced across a range of applications for a single database. In previous work, we have discussed the advantages for enterprise integration of typed graph data models (TGM), which can play a similar role in graphical databases, leveraging the existing support for the unified modelling language UML. Ideally, the integration of systems designed with different models, for example, graphical and relational database, should also be supported. In this work, we implement this approach, using metadata in a relational database management system (DBMS).**

Keywords—typed graph model; graph schema; relational database; implementation; information integration.

I. INTRODUCTION

For many years, the process of database implementation has included a conceptual data modeling phase, and this has often been supported by declarative structures using annotations or attributes [1]. Some recent DBMS have included metadata in the relational model to form a bridge with the physical database. This approach brings significant advantages in that the data model can be enforced across all applications for a single database. In previous work [2], we provided mapping rules for TGM so that data models can play a similar role in graphical databases, using the notations of UML [3]. During such early conceptual model building, incremental and interactive exploration can be helpful [4] as fully automated integration tools may combine things in an inappropriate way, and the use of data types [5] can help to ensure that semantic information is included not merely in the model, but also in the final database. In this short paper we report on such an implementation of TGM, using metadata in a relational DBMS [6], partly inspired by recent developments in the PostgreSQL community [7].

As with the original relational model, the Typed Graph Model (TGM) has a rigorous mathematical foundation as an instance of a Graph Schema.

The plan of this short paper is to review the TGM in Section II, and discuss the implementation details in Section III, including an illustrative example. Section IV provides some conclusions.

II. THE TYPED GRAPH MODEL AND INFORMATION INTEGRATION

We will construct a TGM for a database by declaring instances of nodes and edges as an alternative to specifying tables of nodes and edges.

A. Typed Graphs formalism

In this section we review the informal definition of the TGM from [2], using small letters for elements (nodes, edges, data types, etc.) and capital letters for sets of elements. Sets of sets are printed as bold capital letters. A typical example would be $n \in N \in \mathbf{N} \subseteq \wp(N)$, where N is any set and $\wp(N)$ is the power-set of N.

Let T denote a set of simple or structured (complex) data types. A data type $t:=(l,d) \in T$ has a name l and a definition d. Examples of simple (predefined) types are $(int,\mathbb{Z})$, $(char,ASCII)$, (%,[0..100]) etc. It is also possible to define complex data types like an order line $(OrderLine,(posNo,partNo,partDescription,quantity))$. The components need to be identified in T, e. g., $(posNo,int>0)$. Recursion is allowed as long as the defined structure has a finite number of components.

Definition 1 (Typed Graph Schema, TGS) *A typed graph schema is a tuple* $TGS=(N_S,E_S,\varrho,T,\tau,C)$ *where:*

- N_S *is the set of named (labeled) objects (nodes) n with properties of data type* $t:=(l,d) \in T$*, where l is the label and d the data type definition.*
- E_S *is the set of named (labeled) edges e with a structured property* $p:=(l,d) \in T$*, where l is the label and d the data type definition.*
- ϱ *is a function that associates each edge e to a pair of object sets (O,A), i. e.,* $\varrho(e):=(O_e,A_e)$ *with* $O_e,A_e \in \wp(N_S)$*.* O_e *is called the* tail *and* A_e *is called the* head *of an edge e.*
- τ *is a function that assigns for each node n of an edge e a pair of positive integers* (i_n,k_n)*, i. e.,* $\tau_e(n):=(i_n,k_n)$ *with* $i_n \in N_0$ *and* $k_n \in N$*. The function* τ *defines the min-max multiplicity of an edge connection. If the min-value* i_n *is 0 then the connection is optional.*
- *C is a set of integrity constraints, which the graph database must obey.*

 ISBN: 978-1-68558-056-8

The notation for defining data types T, which are used for node types N_S and edge types E_S, can be freely chosen: and in this implementation SQL will be used for identifiers and expressions, together with a strongly typed relational database engine. The integrity constraints *C* restrict the model beyond the structural limitations of the multiplicity τ of edge connections. Typical constraints in *C* are semantic restrictions of the content of an instance graph.

Definition 2 (Typed Graph Model) *A typed graph Model is a tuple TGM=(N,E,TGS,φ) where:*

- *N is the set of named (labeled) nodes n with data types from N_S of schema TGS.*
- *E is the set of named (labeled) edges e with properties of types from E_S of schema TGS.*
- *TGS is a typed graph schema as defined above..*
- *φ is a homomorphism that maps each node n and edge e of TGM to the corresponding type element of TGS, formally:*

$$\varphi: TGM \rightarrow TGS$$
$$n \mapsto \varphi(n) := n_S (\in N_S)$$
$$e \mapsto \varphi(e) := e_S (\in E_S)$$

The fact that φ maps each element (node or edge) to exactly one data type implies that each element of the graph model has a well-defined data type. The homomorphism is structure preserving. This means that the cardinality of the edge types is enforced, too. In this implementation, the declaration of nodes and edge of the TGM develops the associated TGS incrementally including the development of the implied type system *T*. Data type and constraint checking is applied for all nodes and edges before any insert, update, or delete action can be committed.

B. The Data Integration Process

The full benefit of information integration requires the integration of source data with their full semantics. We believe a key success factor is to model the sources and target information as accurately as possible. The expressive power and flexibility of the TGM allows to describe the meta-data of the sources and target precisely and in the same model, which simplifies the matching and mapping of the sources to the target. The tasks of the data integration process are:

1) model sources as TGS S_i (i = 1, 2, ..., n)
2) model target schema *T* as TGS *G*
3) match and map sources S_i with TGS *G*
4) check and improve quality
5) convert TGS G back to T again

Steps 3 and 4 can occur together in an interactive process once the basic model has been outlined. Such a process is crucial for EII and other data integration projects, which demand highly accurate information quality, which can be further improved with the use of different mappings.

To start the process, it may be necessary to collect structure and type information from a data expert or from additional information. Where sources are databases, the rigid structures provide a good starting point. Otherwise, the relevant data must first be identified together with its meta-data if available. This includes coding and names for the data items. The measure units and other meta-data provided by the data owner are used to adjust all measures to the same scale. The paper of Laux [5] gives some examples how to transform relational, object oriented, and XML-schemata into a TGS.

If the source is unstructured or semi-structured, e.g., documents or XML/HTML data, concepts and mechanisms from Information Retrieval (IR) and statistical analysis may help to identify some implicit structure or identify outliers and other susceptible data. If the data are self-describing (JSON, key-value pairs, or XML) linguistic matching can be applied with additional help from a thesaurus or ontology. Nevertheless, it is advisable to validate the matching with instance data or an information expert. We present two possible TGS for a single enterprise in UML notation in Figure 1. This little example demonstrates already the flexibility of the model in terms of detail and abstraction.

III. Implementation in the Relational Database Schema

The prerequisite for implementation of a typed graph modelling system is to have a strong type system in the RDBMS. If this is already available, then a graph modelling capability can be added relatively simply, with slight extensions to the normal SQL syntax for creating and altering structured types, and some metadata for distinguishing node and edge types from other kinds of structured types.

Then the main difference between a graph schema as described above and a schema in most DBMS is that columns and attributes of database tables have a predefined order. In addition, for a given node type or edge type, there is a single base table containing the instances of that type. One way to build a graph is to insert rows in these tables.

The aim of additional graph support in the DBMS is to simplify the tasks of graph definition and searching. We add CREATE and MATCH statements, which we describe next.

A. Graph-oriented syntax added to SQL

To the normal SQL CREATE syntax, we add an option for constructing a graph inline:

```
   Create: CREATE Node { Edge Node } {',' Node { Edge Node }}.
   Node: '(' NewG | id ')'.
   NewG: id {':' label } [doc] .
   –Edge: '-[' NewG ']->' | '<-[' NewG ']-' .
```

In this syntax, the strings enclosed in single quotes are tokens, including several new token types for the TGM. In corresponding source input, unquoted strings are used for case-insensitive identifiers and double quoted strings for case-sensitive identifiers, possibly containing other Unicode characters. As usual in SQL, string constants in input will be

single quoted, and doc is a JSON-like structure providing a set of properties and value expressions, possibly including metadata definitions for ranges and multiplicity.

Such declarative statements build a base table in the database for each label.

Nodes and edges and new node types and edge types can be introduced with this syntax. The database engine constructs a base table for each distinct label, with columns sufficient to represent the associated properties. These database base tables for node types (or edge types) contain a single row for each node (resp. edge) including node references. They can be equipped with indexes, constraints, and triggers in the normal ways.

To the normal SQL DML, we add the syntax for the MATCH query, which has a similar syntax, except that it may contain unbound identifiers for nodes and edges, their labels and/or their properties.

```
Match: MATCH Node {',' Node } [WhereClause]
Statement .
```

The first part of the MATCH clause defines a graph expression. We say that a graph expression is bound if it contains only constant values, and all its nodes and edges are in the database. The MATCH algorithm proceeds along the node expressions, matching more and more of its nodes and edges with those in the database by assigning values to the unbound identifiers. If we cannot progress to the next part of the MATCH clause, we backtrack by undoing the last binding and taking an alternative value. If the processing reaches the end of the MATCH statement, the set of bindings contributes a row in the default result, subject to the optional WHERE condition. These rows then act as a source of values for the following statement.

B. Outline of the usage of the TGM

Following the suggestion in [5] we will consider the use of the TGM in analysis, where an interactive process is envisaged. The nodes and edges contained in the database combine to form a set of disjoint graphs that is initially empty. Adding a node to the database adds a new entry to this set. When an edge is added, either the two endpoints are in the same graph, or else the edge will connect two previously disjoint graphs. If each graph in the set is identified by a representative node (such as the one with the lowest uid) and maintains a list of the nodes and edges it contains, it is easy to manage the set of graphs as data is added to the database.

If an edge is removed, the graph containing it might now be in at most two pieces: the simplest algorithm removes it from the set and adds its nodes and edges back in.

The database with its added graph information can be used directly in ordinary database application processing, with the advantage of being able to perform graph-oriented querying and graph-oriented stored procedures. The normal processing of the database engine naturally enforces the type requirements of the model, and also enforces any constraints specified in graph-oriented metadata. The nodes and edges are rows in ordinary tables that can be accessed and refined using normal SQL statements. In particular, using the usual dotted syntax, properties can be SET and updated, and can be removed by being set to NULL.

As the TGM is developed and merged with other graphical data, conflicts will be detected and diagnostics will help to identify any obstacles to integrating a new part of the model, so that the model as developed to that point can be refined.

C. An example

To get started with a customer-supplier ordering system we could have a number of problematic CREATE statements such as:

```
   CREATE
   (Joe:Customer {"Name":'Joe Edwards',
Address:'10 Station Rd.'}),
   (Joe)-[:Ordered {"Date":date'22/11/2002'} ]->
(Ord201:"Order")-[:Item {Qty: 5}]->("16/50x100" :
WoodScrew),
   (Ord201)-[:Item {Qty: 5}]->("Fiber 12cm":
WallPlug),
   (Ord201)-[:Item {Qty: 1}]->("500ml" :
RubberGlue)
```

Primary keys for edges are here being left to the engine to supply – they could be specified explicitly if preferred. Name, Order and Date are in double quotes because they are reserved words in SQL. By default, the entire CREATE statement shown is considered a single transaction by the database engine: if the syntax checker is happy with it, it will be automatically committed.

It is easy to criticize what the user offers here: and the graph would benefit from splitting up composite information such as Fibre 12cm and 16/8x100 to clarify the meaning of the components and facilitate processing. Such changes can be made by the designer later.

Assuming the database is empty before we start, the first line above, if committed, would create a new base table CUSTOMER (a NodeType)

```
   CREATE TYPE CUSTOMER as ("Name" char, ADDRESS
char) NodeType
```

The NodeType metadata flag adds as the first column a primary key column ID of type char so that the new CUSTOMER table has an initial row

```
        ('JOE','Joe Edwards','10 Station Rd.')
```

That would work. The next line defines four more base tables, two NodeTypes and two EdgeTypes:

```
   CREATE TYPE "Order" NodeType
   CREATE TYPE WOODSCREW NodeType
   CREATE TYPE ORDERED as ("Date" date)
        EdgeType(CUSTOMER,"Order")
   CREATE TYPE ITEM as (QTY int)
        EdgeType("Order",WOODSCREW)
```

 ISBN: 978-1-68558-056-8

This also will work, but is probably not what the analyst wanted, because the Item edge type connects to nodes of type WOODSCREW. If this is committed, we cannot later have an Item edge connecting to a WALLPLUG.

But nothing is committed yet, so when the database engine finds this difficulty, it simply replaces the specification `:WoodScrew` in the second line by `:WoodScrew:&1` , and similar changes to WallPlug and RubberGlue.

This adds a new anonymous base node type for these node types, with a system-generated name

```
CREATE TYPE &1 NodeType
```

and the node type proposal becomes

```
CREATE TYPE ITEM as (QTY int)
     EdgeType("Order",WOODSCREW)

CREATE TYPE WoodScrew UNDER &1
CREATE TYPE WallPlug UNDER &1
CREATE TYPE RubberGlue UNDER &1
```

The analyst can be advised that this has been done, and they can later choose a better name for the new NodeType &1 (maybe PRODUCT?). This process of generalization can be offered as a standard database transformation.

After the nodes and edges have been generated and the transaction commits, the node and edge data would be installed in the database as follows:

```
CUSTOMER ('Joe','Joe Edwards',
     '10 Station Rd.')
"Order" ('Ord201')
WOODSCREW ('16/50x100')
WALLPLUG ('Fiber 12cm')
RUBBERGLUE ('500ml')
ORDERED ('&2','Joe','Ord201',
     'date'22/11/2002')
ITEM ('&3','Ord201','16/50x100')
     ('&4','Ord201','Fiber 12cm')
     ('&5','Ord201','500ml')
```

This is satisfyingly neat. We see that while the metadata flag NodeType gave the node type a primary key as the first column ID that is a primary key, the metadata flag EdgeType has given the edge types three initial columns: ID, a primary key, LEAVING, a foreign key to the leaving node type, and ARRIVING, a foreign key to the arriving type. Note also that ITEM's arriving type is the new anonymous type &1.

It is noteworthy that this mechanism allows schemas to evolve bottom-up during the database design, as envisaged in [2]. The normal Schema-first strategy is still available, and the two approaches can be combined for convenience. Either way, the database will contain a rigorous and enforceable relational schema at all stages, since any declarations that would not be enforceable will be rejected before being committed to the database.

During refinement of the model, there are opportunities for adding constraints and other metadata. Such details, and the enhanced diagnostics mentioned above, are the subject of ongoing research. The conference presentation will provide an opportunity for a demonstration of the process and more details on MATCH.

IV. CONCLUSIONS

The purpose of this paper was to report some progress in our Typed Graph Modeling workstream. The work is available on Github [8] for free download and use and is not covered by any patent or other restrictions.

The current "alpha" state of the software implements all of the above ideas apart but lacks the suggested interaction with the model designer. The test suite includes a version of the running example together with others that demonstrate the integration of the relational and typed graph model concepts in Pyrrho DBMS.

REFERENCES

[1] Oracle, Oracle Product Documentation (Online), Available from: https://docs.oracle.com/javaee/7/tutorial/persistence-intro.htm#BNBPZ [retrieved: Feb, 2023]

[2] F. Laux and M. Crowe, "Information Integration using the Typed Graph Model", DBKDA 2021: The Thirteenth International Conference on Advances in Databases, Knowledge, and Data Applications, IARIA, May 2021, pp. 7-14, ISSN: 2308-4332, ISBN: 978-1-61208-857-0

[3] E. J. Naiburg, and R. A. Maksimschuk, UML for database design. Addison-Wesley Professional, 2001

[4] R. De Virgilio, A. Maccioni, A., R. Torloner, "Model-Driven Design of Graph Databases", in Yue, E. et al (eds) Conceptual Modeling, 33rd International Conference (ER 2014), Springer, Oct 2014, pp. 172-185, ISSN: 0302-9743 ISBN: 978-3-319-12205-2

[5] F. Laux, "The Typed Graph Model", DBKDA 2020 : The Twelfth International Conference on Advances in Databases, Knowledge, and Data Applications, IARIA, Sept 2020, pp. 13-19, ISSN: 2308-4332, ISBN: 978-1-61208-790-0

[6] M. Crowe, and F. Laux, "Database Technology Evolution", IARIA International Journal on Advanced is Software, vol 15 (3-4) 2022, pp. 224-234, ISSN: 1942-2628

[7] S. Shah, et al. The PostgreSQL Data Computing Platform (PgDCP) (Online), Available from: https://github.com/netspective-studios/PgDCP [retrieved: Feb 2023]

[8] M. Crowe, PyrrhoV7alpha, https://github.com/MalcolmCrowe/ShareableDataStructures [retrieved: Feb 2023]

 ISBN: 978-1-68558-056-8

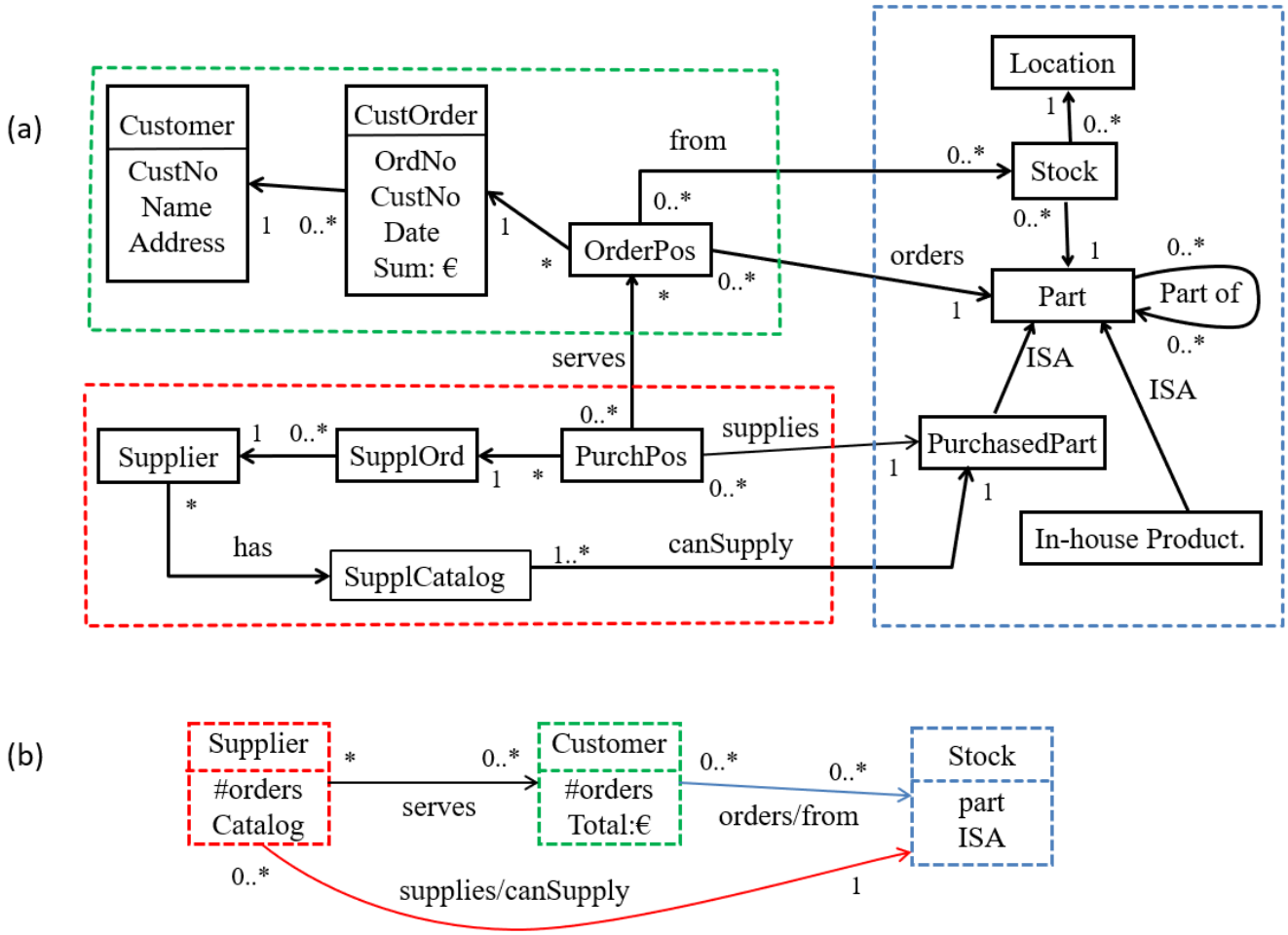

Figure 1. Example TGM of a commercial enterprise showing two levels of detail

Memory Efficient Data-Protection for Database Utilizing Secure/Unsecured Area of Intel SGX

Masashi Yoshimura
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan
email: yoshimura.masashi.yj6@is.naist.jp

Taisho Sasada
Division of Information Science
Nara Institute of Science and Technology
Research Fellow of the JSPS
Nara, Japan
email: sasada.taisho.su0@is.naist.jp

Yuzo Taenaka
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan
email: yuzo@is.naist.jp

Youki Kadobayashi
Division of Information Science
Nara Institute of Science and Technology
Nara, Japan
email: youki-k@is.naist.jp

***Abstract*—With the spread of cloud computing, database services have been provided on cloud platforms. As a Cloud Service Provider (CSP) has the highest privilege in the cloud platform, the CSP can get any data from the database even if a tenant admin secures all components, such as OS, database software, etc. as long as the database runs on the cloud. That is why CSP has become a new threat source in cloud-based databases. Trusted Execution Environment (TEE) is a key technology to protect memory, process, and storage against data theft by a CSP. It creates a secure area on the memory where the process outside the secure area cannot access, thereby preventing any access from CSP. However, since the secure area only has a limited amount of memory resources on a server, the rest memory resources keep vacant even when TEE exhaustively uses its allocated memory resources. In the case of the high-load database running on the secure area, almost all queries slow down due to being full of consumed memory despite most of the memory being free in the unsecured area. In this study, we design an efficient memory management mechanism for TEE-based secure database that effectively uses the resources of both the secure and unsecured areas; the proposed system handles only sensitive queries and data in the secure area while others in the unsecured area. Experimental results show that our system improves both resource utilization efficiency and execution speed compared to the system processing all data in the secure area.**

***Index Terms*—Data Protection; RDBMS; Intel SGX; Trusted Execution Environment; Cloud Computing.**

I. Introduction

Along with the widespread of using cloud platforms, most services come to running on a cloud platform. Although the cloud is very useful for flexible service management adapting to time-varied workloads, it creates new threats to cybersecurity. A cloud platform runs tenant processes on the top of the virtualization layer, such as a hypervisor, and thus all processes, memory, and storage are accessible from the virtualization layer. That is, even if tenant admins strictly secure their OS, service processes, and data, CSP is able to affect processes, obtain data from storage or memory, etc.

One of the major and important systems running on the cloud is a relational database management system (RDBMS), which basically handles important data on the business. As most business is driven by data these days, protecting data is extremely essential for a database. For this purpose, most RDBMS, such as MySQL or PostgreSQL, has an encryption function that encrypts data on storage and protects against data theft. Nevertheless, as mentioned before, CSP can compromise even such encryption by taking process/memory from the virtualization layer. Therefore, it is necessary to protect data on RDBMS even on cloud systems consistently.

Existing studies provide a Trusted Execution Environment (TEE)-based solution to protect the data of the database [1] [2]. TEE creates a secure area where user programs are decrypted and executed directly on a CPU. As any process outside the secure area cannot access process/memory in the secure area, a database working in the secure area can protect its data in any case. Although TEE protects data even on the cloud, TEE has a limitation on the secure area; the secure area can only use quite less memory than what hardware has to run the program. That is why the performance of a high-load database hits the ceiling even if its physical hardware has more memory and most of it is still vacant. Studies [1] [3] run a database in the secure area and face the problem, while a study [2] migrates most of the process onto the unsecured area but increasing communication between secure and unsecured area, and finally these interactions become a performance bottleneck.

To overcome this problem, we propose a system that protects data in RDBMSs that extends the upper limit of TEE-based database performance. The main idea is to divide the whole processes of a database into two: a series of processes for sensitive data and a series of processes for other non-sensitive data. The proposed system allows the user to define the confidentiality of each column when creating the table and executes the former process on the secure area while the latter on the unsecured area, separately. The proposed system also reduces

 ISBN: 978-1-68558-056-8

the number of communication between the secure/unsecured areas because each data is handled in the secure or unsecured area all the time. The basic design of the RDBMS is based on open-source Postgresql [3]. We implement our proposal using Intel Software Guard Extensions (SGX), a hardware-based TEE of Intel CPU. For performance evaluation, we compare our proposal with the system processing all data in the secure area. We confirmed that the proposed method reduces secure area usage by 44% compared to existing methods and runs over 3.3× faster than existing methods when dealing with a large amount of data.

The structure of this paper is as follows: In Section 2, we explain the related works of databases using TEE. In Section 3, we describe the preliminary of Intel SGX. In Section 4, we explain the sensitive information and the design of our proposed system. In Section 5, we show the result of the experimental evaluation. In Section 6, we discuss the limitation of our proposed system and security vulnerabilities and in Section 7, we conclude our contribution.

II. Related Work

There are several TEE-based solutions to protect the data of RDBMS as related work [1]–[3]. EnclaveDB [1] is a system that makes Hekaton, an in-memory database engine included in Microsoft SQL Server, available as an SGX application. EnclaveDB ensures data confidentiality but handles all types of queries in the secure area. Therefore, the unsecured area has much vacant memory. CryptSQLite [3] proposed a system that ensures the data confidentiality and integrity of SQLite by storing all data in a secure area. That is the available memory in the unsecured area remains free. StealthDB [2] executes most of the database processes in the unsecured area while few sensitive processes are in the secure area. Although it can use the memory of both secure and unsecured areas, processes on the secure and unsecured areas require many interactions to handle a series of query processing. Encryption/decryption of data is necessary to protect data from going back and forth between secure and unsecured areas. Although this design contributes to reducing the load on the secure area, the interaction and the encryption/decryption are a very huge burden for the database. This results in a large overhead for even a simple SELECT statement that traverses a large amount of data. The proposed method divides database processes for the secure or unsecured area, similar to EnclaveDB or StealthDB, but makes these processes independent so as to avoid communication between secure and unsecured areas as much as possible. From this design, we realize the efficient use of memory in the secure area as well as the avoidance of performance bottlenecks that happened in the communication between the secure and unsecured area.

III. Intel SGX

Intel SGX utilizes the cryptographic engines in Intel CPU to create an isolated environment (secure area) called Enclave. We store data in Enclave, protecting program execution with guaranteed confidentiality and integrity. As Enclave does not provide storage, Intel SGX provides a function called Sealing/Unsealing. This function encrypts data using a key stored in the CPU; nobody except the CPU decrypts it. Moreover, for integrity assurance, Intel SGX provides a verification mechanism called Remote Attestation (RA), which can verify the integrity of programs within Enclave and the remote SGX platform. Thus, a client communicating with a remote SGX platform can send and receive data securely to and from CSPs using TLS sessions generated by RA.

Although Intel SGX provides a useful mechanism for protecting processes and memory, data, the mechanism inevitably includes several performance overheads. First, there is an Enclave size limitation for each Intel CPU version. For example, the 6th to 10th-generation Intel CPUs have a size limit of 128 MB, and the 3rd-generation Xeon scalable processors have a maximum size limit of 512 GB. Second, Intel SGX supports Enclave paging, but page swapping incurs an overhead of about 40,000 CPU cycles due to page copy and context switches and so on [4]. Therefore, when we try to use many Enclave areas, much overhead is incurred. Third, SGX applications provide a transition between the Enclave process and the unsecured process. The transition function from an unsecured process to an Enclave process is called Ecall and the reverse transition function is called Ocall. However, during Ocall/Ecall, SGX performs context switches and flushes Translation Lookaside Buffer, resulting in an overhead of about 8,000 to 17,000 CPU cycles [5].

IV. Proposed Method

Before going into the detail of our proposal, we describe the threat model. Our threat model is information leakage at a database server. The adversary is a malicious CSP that has free access to memory and storage. Specifically, the adversary can steal data in memory by memory dump or cold boot attacks and data in storage by physically obtaining storage devices. Note that the retrieval of sensitive data or encryption keys directly from Intel SGX [6] [7] is beyond the scope.

As a general database that supports many kinds of queries, this paper limits the target queries for simplicity to basic CRUD operations (CREATE, INSERT, SELECT, UPDATE, and DELETE). Note that the support of queries such as subqueries and joins is future work.

For making processes independent for the secure and unsecured area, we design a table with secure columns. We focused on the fact that database tables generally consist of several columns having either sensitive data or non-sensitive data. For example, suppose that a table in a database that stores company employee information includes name, age, and hometown. In this case, only the hometown is non-sensitive information because of public data while name and age are sensitive. As these sensitive or non-sensitive data are different in a table, the proposed system allows users to define the confidentiality of data on a column-by-column basis when the table is created. In the proposed system, only the processing of sensitive data is performed in the secure area, and all other processing is performed in the unsecured area in order

 ISBN: 978-1-68558-056-8

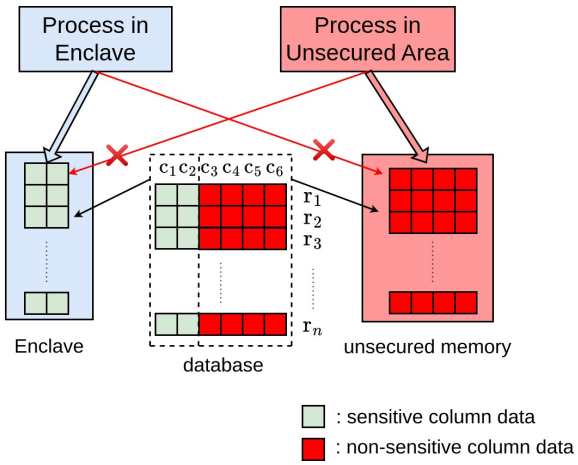

Figure 1. Proposed System Overview.

to improve resource utilization efficiency. Figure 1 shows this concept where columns of sensitive data, called sensitive columns, (c1 and c2) are treated in the secure area while columns of non-sensitive data, called non-sensitive columns, are processed in the unsecured area.

A. Sensitive Information

The proposed architecture is shown in Figure 2. We handle four types of information that may contain sensitive data in the proposed architecture: (1) queries, (2) query trees, (3) plan trees, and (4) sensitive buffer pools. We describe them one by one.

(1) Query: As queries may contain sensitive data (e.g., INSERT statements including sensitive data or queries containing WHERE clauses that specify values of sensitive data), all queries must be processed in the secure area. If the query does not contain sensitive data, data processing for the query is done in the unsecured area.

(2) Query Tree: The query tree is an abstract syntax tree of a query and therefore contains the same sensitive information as the query. Therefore, our system processes the query tree in the secure area.

(3) Plan Tree: A plan tree is a tree structure that shows the optimal query plan for a query tree. As a plan tree is used to issue instructions to process data in reality, it must be separated for the secure and unsecured areas in accordance with sensitive or non-sensitive columns. To avoid communication between two areas, we here have to make two different (non-related) plan trees for secure and unsecured areas.

(4) Buffer Pool: Since each record contains both non-sensitive and sensitive data, the non-sensitive data of all records is deployed in the non-sensitive buffer pool in the unsecured area and the sensitive data of all records is deployed in the sensitive buffer pool in the secure area. Each buffer pool is a fixed-length array of pages with a specified size (8KB, the same as the default setting of PostgreSQL), as in general RDBMS.

B. Design details of each module

The proposed system consists of 13 modules. We describe the function and key points of each module in the proposed method.

Communication Process, Decryption: The Communication Process (① in Figure 2) performs RA and query reception. A database client performs RA verification to determine whether or not the cloud server's platform (CPU) and the secure area can be trusted. If the client accepts the RA verification results and trusts the server platform and the secure area, the client encrypts the query using the symmetric key generated in the RA process and sends it to the cloud. After the communication process receives the encrypted query, it is sent to the secure area, and Decryption (② in Figure 2) decrypts the query using the symmetric key held within the secure area.

Parser: Parser (③ in Figure 2) generates a query tree from the query in the secure area.

Query Planner: The query planner (④ in Figure 2) generates a tree structure data called a plan tree representing the optimal query plan. The query planner optimizes a query tree so as to process the query efficiently, reducing the number of computations or data access to storage. The optimization is done for the entire query tree without taking care of non-sensitive or sensitive columns in this module.

Query Separator: The query separator (⑤ in Figure 2) can divide the optimal plan tree into a sensitive plan tree that handles only sensitive columns and a non-sensitive plan tree that handles only non-sensitive columns. If a query tree includes one or more sensitive columns, the query separator creates a new plan tree, called a sensitive plan tree, including sensitive columns only, which has the same structure as the original plan tree. Regarding non-sensitive columns, it makes the same tree remain non-sensitive columns only, which is called a non-sensitive plan tree. However, there are some queries that cannot be simply divided. Figure 3 shows the division of the plan tree of `SELECT (name, club) FROM USER WHERE country = 'Japan';`. The table USER consists of three columns: `name`, `club`, `country`. Only `name` is a sensitive column in this case. In this query, the `name` column data to be selected depends on `WHERE country = 'Japan'`, but the `country` column cannot be included in the sensitive plan tree because it is a non-sensitive column. In the proposed system, the sensitive plan tree generated by the division has an empty WHERE clause. After finishing the process of the non-sensitive plan tree, the WHERE clause of the sensitive plan tree refers to the identifiers ($id_1, id_2, ..., id_n$ in Figure 3) of records satisfying `WHERE country = 'Japan'`. Division of the plan tree is particularly important in the processing of complex queries with multiple subqueries, and it is the future work to design algorithms for processing like that queries efficiently utilizing the secure areas and reducing the number of communications between the secure and unsecured areas.

Query Executor: The Query Executor generates specific data processing instructions according to a plan tree. As the query separator makes two plan trees, non-sensitive and sensitive plan trees, the query executor is also required to be two in order to execute these two trees in the secure and unsecured area,

 ISBN: 978-1-68558-056-8

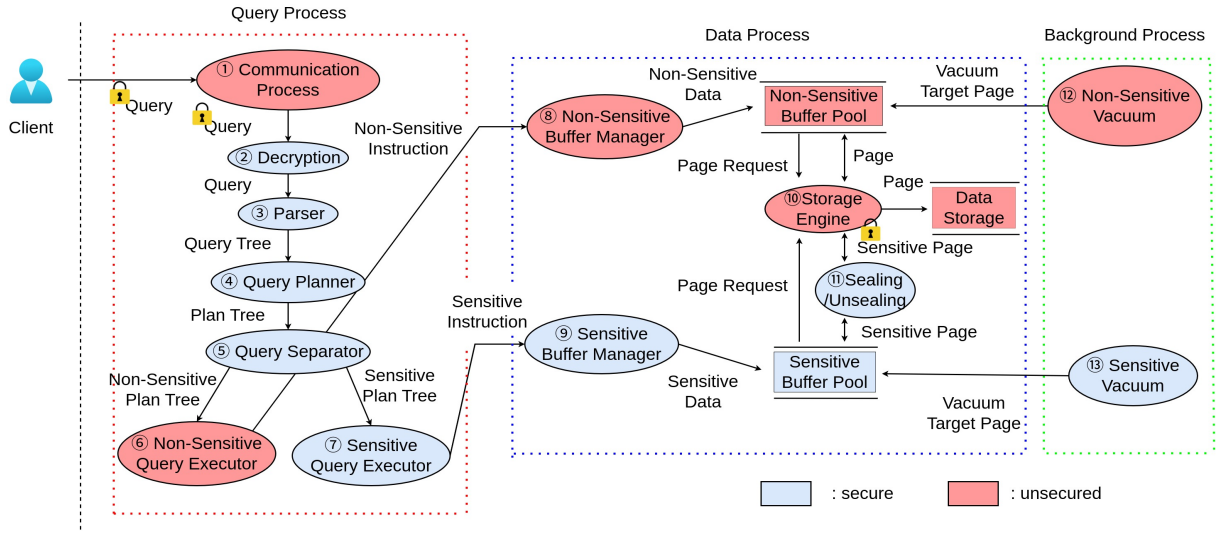

Figure 2. Architecture of the Proposed System.

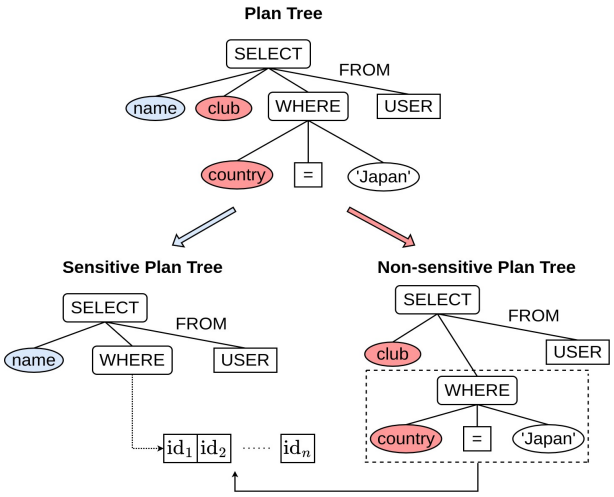

Figure 3. Division of the Plan Tree.

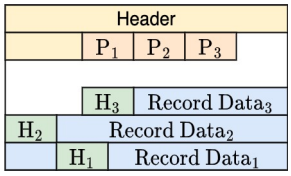

Figure 4. Data Page Structure in PostgreSQL.

respectively. We call it running in the secure area as a sensitive query executor while it is in the unsecured area as a non-sensitive query executor. Their function of them is identical except for the area they are running.

Buffer Manager: As a query is processed in parallel in both the secure and unsecured area, two buffer managers work in those areas, respectively, called sensitive and non-sensitive buffer managers. The functionality of the buffer manager is to handle the buffer pool based on data processing instructions issued by the Query Executor. The buffer manager in the unsecured area (⑧ in Figure 2) operates the non-sensitive buffer pool, while the buffer manager in the secure area (⑨ in Figure 2) operates the sensitive buffer pool.

Storage Engine: Storage Engine (⑩ in Figure 2) performs general file I/O processing and store pages of buffer pool in the storage. Storage engine stores sensitive buffer pools and non-sensitive buffer pools in storage for persistent data. To protect the sensitive buffer pool, the sensitive data in the buffer pool needs to be encrypted before going to the unsecured area. The encryption is done by using the sealing (⑪ in Figure 2) function of Intel SGX, which is the encryption function using a secret key of the CPU. To extract the encrypted data in the secure area, unsealing (⑪ in Figure 2) function is provided.

We explain the page structure of the buffer pool and how pages of the buffer pool are stored in the storage. The page structure of the buffer pool is similar to that of PostgreSQL as shown in Figure 4. PostgreSQL holds fixed-size memory for every page and manages data with a record unit on a page. On a page, every record is placed from the end of the page back-to-back, and those pointers (P) indicating the location of every record are put one by one after the header information. Note that a page contains the page header (Header in Figure 4) and the record header (H_1, H_2, H_3 in Figure 4). The page structure of the proposed system is the same as PostgreSQL but the page of the sensitive buffer pool has only sensitive data of records and the page of the non-sensitive buffer pool has only non-sensitive data of records. When storing buffer pool pages, sensitive data must be encrypted by Intel SGX sealing before storing sensitive pages. The simplest method is to encrypt the entire sensitive page. However, since the length of the encrypted byte array for the entire page exceeds 8KiB (page size), it becomes impossible to manage the data by fixed-length pages in the storage. In fact, it is sufficient to encrypt only the sensitive data on the sensitive page. Thus, a method to encrypt only the sensitive data of each record is considered, but it

 ISBN: 978-1-68558-056-8

TABLE I
STUDENT TABLE FOR EVALUATION

columns	id	name	university	club
type	integer	char(100)	char(100)	char(50)
attribution	normal	sensitive	normal	sensitive

```
// Query-1
INSERT INTO STUDENT (id, name, university, club)
    values (1, "John", "NAIST", "soccer");
// Query-2
UPDATE STUDENT set name = "Mike" where id = 1;
// Query-3
DELETE from STUDENT where id = 1;
// Query-4
SELECT * FROM STUDENT;
```

Figure 5. Transaction for Evaluation.

requires encryption of the number of records on a page, thereby incurring extra overhead. Considering the above, the proposed system adopts the method of encrypting all records (including the header of each record) on a sensitive page at once. In this way, only one encryption per page is required.

Vacuum Process: Vacuum Process is a background process that periodically cleans up buffer pools becoming dirty as a result of repeated data processing. This process runs in the secure area and unsecured areas (⑫, ⑬ in Figure 2) to handle sensitive and non-sensitive buffer pools, respectively.

V. EVALUATION

As a security evaluation, it is necessary to show that the confidentiality of sensitive information is ensured. In this study, the confidentiality of sensitive information means that the sensitive information exists as plain text only in the secure area and is always encrypted in the unsecured area. In Section IV, we can see that all sensitive information is processed in the secure area and is always encrypted before being sent to the unsecured area, so confidentiality is ensured.

We evaluated the performance of the proposed system. The experimental environment used for performance evaluation was Ubuntu 20.04LTS OS, on Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz, 4 CPU cores, SODIMM DDR3-1600 8GiB memory, Samsung SSD 860 500GiB. In this experimental environment, the available secure area is limited to 128 MiB at the same time due to the Intel CPU version. The proposed system was implemented with C++ and Intel SGXSDK [8], a development tool for SGX applications. In the performance evaluation, we evaluated the secure area usage (the amount of peak stack and heap memory in the secure area) and the execution time (a period between receiving a query from a client and generating a reply to the client) for the transaction of Figure 5. Note that the value of each query and the right-hand value of the WHERE clause vary by transaction.

We compare the secure area usage and execution time for processing 1, 10, 100, and 1000 transactions respectively on the proposed system and a comparative system (same features as EnclaveDB), which processes all data in the secure area. Also,

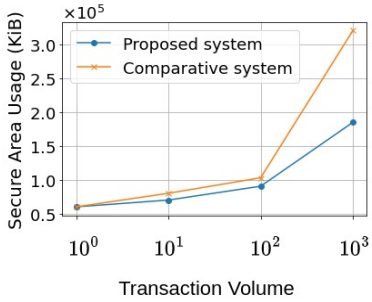

Figure 6. Peak Secure Stack and Heap Usage of Transactions.

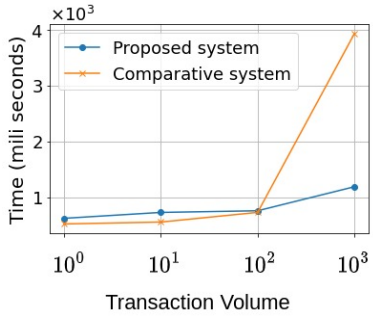

Figure 7. Execution Time of Transactions.

we ran the system three times for each transaction volume and used the average values of secure area usage and execution time as the evaluation values. The evaluation of secure area usage is shown in Figure 6, and the evaluation of execution time is shown in Figure 7. As Figure 6 shows, the proposed system uses less secure area than the comparative system for all 1, 10, 100, and 1000 transactions. As Figure 7 shows, the execution times of the proposed method and the comparative system are almost the same for 1, 10, and 100 transactions, but the overhead of the comparative system is about 3.3 times larger than that of the proposed method for 1000 transactions. Since the secure area available in the experimental environment is 128 MiB (96 MiB excluding reserved area), paging occurs very frequently for 1000 transactions of the comparative system, which uses much more than 96 MiB of the secure area. Thus, the proposed method improves execution speed over the comparative system when handling a large amount of sensitive data. In this evaluation, we used our own tables and transactions with transaction volumes of 10^0, 10^1, 10^2, and 10^3. It is future work to evaluate the results for larger transaction volumes and standard benchmark tests.

We explain the amount of change in secure areas and the number of communications between secure/unsecured areas as the number of tables, columns, and records increases. Since the database tables are managed in the unsecured area, the usage of the secured area does not increase even if the number of tables increases. When the number of sensitive columns or the number of records containing sensitive data increases, the usage of the secure area increases by the size of the secure area, but the number of communications between the secure/unsecured areas during query processing remains the same. However, if

 ISBN: 978-1-68558-056-8

TABLE II
SCORE TABLE

columns	id	name	score1	score2
type	integer	char(100)	integer	integer
attribution	non-sensitive	sensitive	non-sensitive	sensitive

TABLE III
SCORE DATABASE

id	name	score1	score2
1	John	86	68
2	Tom	95	98
3	Mike	58	99

the number of sensitive columns or records greatly increases, the secure area will be overutilized, and resource utilization efficiency will decrease. Thus, the proposed system can process even a large database consisting of multiple tables without incurring significant overhead if there is not a large amount of sensitive data.

VI. DISCUSSION

A. Extended RDBMS functionality

The current design ensures the confidentiality of sensitive data for queries that perform basic CRUD operations, but there are some queries that leak sensitive data. For example, suppose there is a table SCORE such as TABLE II and a database such as a TABLE III, and the query of Figure 8 is processed. In this query, Since `score1` is a non-sensitive column, `WHERE score1 < ···` is performed in the unsecured area. Thus, the return value of `SELECT score2 FROM WHERE id = 2` must be used in the unsecured area, leading to the leakage of sensitive data because `score2` is a sensitive column. However, since processes in the secure area can directly handle data in the unsecured area, we can compare sensitive data with non-sensitive data without storing the non-sensitive data in the secure area. It is necessary to process and evaluate such queries.

The system proposed in this study lacks important functions such as a transaction manager and a log manager, which are included in many RDBMS. It is necessary to evaluate whether or not the addition of such functions will ensure the confidentiality of sensitive data and whether or not the system can demonstrate practical performance.

B. Reduce Overhead due to communication between secure and unsecured area and Paging Reduction

As explained in Section III, communication between secure/unsecured areas and page swapping due to excessive use of secure areas incurs a large overhead. In the proposed system, these overheads are a serious problem when processing large amounts of sensitive data. As solutions to these problems, Intel SGX provides Switchless Call [9], which enables the communication between secure/unsecured areas without context switches and Eleos [10] enables paging within the secure area, thus reducing the paging overhead. The implementation of these techniques in our proposed system can improve performance.

```
SELECT * FROM SCORE
    WHERE score1 < (
        SELECT score2 FROM SCORE
            WHERE id = 2
    );
```

Figure 8. Sensitve Data Leakage Query.

C. Security Vulnerabilities

The secure area of Intel SGX is vulnerable to side-channel attacks, which can leak secret keys and internal registers [6] [7] [11] [12], but methods to mitigate these attacks significantly with little overhead are being studied [13] [14].

VII. CONCLUSION

In this paper, we proposed data protection for RDBMS that ensures data confidentiality while improving overall resource utilization efficiency using Intel SGX, a TEE with high-security features. The system efficiently utilizes the secure area by offloading only sensitive data and the processes that handle them to the secure area. In particular, in the case of handling both sensitive data and large amounts of non-sensitive data, the proposed system has improved both resource utilization efficiency and execution speed compared to the system processing all data in the secure area. We need to work on designing more query support and other important modules in the future.

REFERENCES

[1] C. Priebe, K. Vaswani, and M. Costa, "Enclavedb: A secure database using SGX," in *IEEE S&P 2018*, pp. 264–278, IEEE Computer Society, 2018.
[2] D. Vinayagamurthy, A. Gribov, and S. Gorbunov, "Stealthdb: a scalable encrypted database with full SQL query support," *Proc. Priv. Enhancing Technol.*, vol. 2019, no. 3, pp. 370–388, 2019.
[3] Y. Wang, Y. Shen, C. Su, J. Ma, L. Liu, and X. Dong, "Cryptsqlite: Sqlite with high data security," *IEEE Transactions on Computers*, vol. 69, no. 5, pp. 666–678, 2019.
[4] M. Taassori, A. Shafiee, and R. Balasubramonian, "Vault: Reducing paging overheads in sgx with efficient integrity verification structures," in *ASPLOS'18*, pp. 665–678, 2018.
[5] O. Weisse, V. Bertacco, and T. Austin, "Regaining lost cycles with hotcalls: A fast interface for sgx secure enclaves," *ACM SIGARCH Computer Architecture News*, vol. 45, no. 2, pp. 81–93, 2017.
[6] Y. Xu, W. Cui, and M. Peinado, "Controlled-channel attacks: Deterministic side channels for untrusted operating systems," in *IEEE S&P 2015*, pp. 640–656, 2015.
[7] J. Götzfried, M. Eckert, S. Schinzel, and T. Müller, "Cache attacks on intel sgx," in *Proceedings of the 10th European Workshop on Systems Security*, pp. 1–6, 2017.
[8] "Intel Software Guard Extensions (Intel SGX) SDK for Linux OS." https://download.01.org/intel-sgx/sgx-linux/2.13/docs/Intel_SGX_Developer_Reference_Linux_2.13_Open_Source.pdf.
[9] H. Tian *et al.*, "Switchless calls made practical in intel sgx," in *SysTEX'18*, pp. 22–27, 2018.
[10] M. Orenbach, P. Lifshits, M. Minkin, and M. Silberstein, "Eleos: Exitless os services for sgx enclaves," in *EuroSys'17*, pp. 238–253, 2017.
[11] S. Shinde, Z. L. Chua, V. Narayanan, and P. Saxena, "Preventing page faults from telling your secrets," ASIA CCS '16, (New York, NY, USA), p. 317–328, Association for Computing Machinery, 2016.
[12] F. Brasser *et al.*, "Software grand exposure: Sgx cache attacks are practical.," in *WOOT*, pp. 11–11, 2017.
[13] M.-W. Shih, S. Lee, T. Kim, and M. Peinado, "T-sgx: Eradicating controlled-channel attacks against enclave programs.," in *NDSS*, 2017.
[14] A. Rane, C. Lin, and M. Tiwari, "Raccoon: Closing digital {Side-Channels} through obfuscated execution," in *24th USENIX Security Symposium (USENIX Security 15)*, pp. 431–446, 2015.

 ISBN: 978-1-68558-056-8

Enriching the Knowledge of a Domain Expert in a Recommendation System Based on Knowledge-Graph via Integrating a Domain-Specific Ontology

Sivan Albagli-Kim, Dizza Beimel
Department of Computer and Information Sciences
Ruppin Academic Center
Emek Hefer 4025000, Israel
email: sivana@ruppin.ac.il, dizzab2ruppin.ac.il

Sivan Albagli-Kim, Dizza Beimel
Dror (Imri) Aloni Center for Health Informatics
Ruppin Academic Center
Emek Hefer 4025000, Israel
email: sivana@ruppin.ac.il, dizzab2ruppin.ac.il

***Abstract*—This work in progress expands a previous study, where we proposed a decision-making framework designed to address the following need: an end-user contacts a domain expert to help him solve a problem. The end-user and the domain expert establish an interaction between them, consisting of questions and answers. This interaction is required to be effective, and to this end, the number of questions must be limited. The purpose of the framework is to suggest the next question to the domain expert, while he interacts with the end user. The framework consists of inference algorithms, making use of the domain expert's knowledge, which is structured into a knowledge graph. During the interaction, the end-user provides data that is fed into the graph as evidence and serves the inference algorithms to refine the next recommended question to the domain expert. The proposed extension refers to the addition of an existing ontology, describing the relevant domain, to the framework's base of knowledge. In particular, we want to take advantage of the knowledge, existing in the ontology (i.e., concepts and their relations), to enrich the framework's ability to offer a greater and more accurate range of questions. In the paper, we describe the proposed extension, followed by a case study.**

Keywords-knowledge graph; semantic reasoning; medical diagnostic; decision support systems; ontologies.

I. INTRODUCTION

The world of "big data" produces many challenges [1], One of them refers to the integration of big data in the technological realm dealing with decision-making processes to leverage these processes. Considering different needs, there are several types of decision-making processes, each requiring a suitable setup [2].

Our ongoing research [3] focuses on decision-making processes with the following setup: the process involves two entities - an *end user* (which is also the process initiator) and a *domain expert* (which assists the end-user to solve a problem); the entities establish an interaction, consisting of questions and answers, and is required to be as limited as possible (in time, the number of questions, money, etc.).

Given the above setup, we propose a semantic technology-based framework, which assists the domain expert in solving the end-user's problem, by suggesting a set of questions (inferred from the integrated big data) for the end user, such that the cycles of questions and answers will be reduced.

Our framework includes three components: (a) a formal representation of the relevant domain expert's knowledge using semantic technology, specifically a *knowledge graph*, which has emerged as a natural way of representing connected data [4], (b) an interactive set of algorithms, using the knowledge graph, and initial knowledge provided by the end user. The framework suggests relevant questions to the end user, while his/her answers advance the domain expert in the decision-making process and become input for the next iteration. The iterations will stop once the domain expert is satisfied, and a decision is made; (c) a domain-specific *ontology*, which is integrated into the knowledge graph. The ontology enriches the knowledge graph, thereby expanding the set of questions the domain expert can ask the end-user. The larger the question space, the more accurate decisions the domain expert can make.

The framework can support several domains that comply with the required setup; to demonstrate this, we chose to focus on the medical domain. To that end, we built a knowledge graph, which consists of two types of nodes representing *diseases* and *symptoms*. The directional edges, going from a symptom node to a disease node, represent a symptom that characterizes the disease. It is possible that a specific symptom can characterize several diseases. The goal of the decision-making process is to assist the domain expert to decide on a *diagnosis* (i.e., provide an explanation for a given set of symptoms based on analyzing available data).

The terms: disease, symptom, and diagnosis can be generalized, thus being used to represent other domains. For instance, in the domain of appliance repairs: the symptom represents a problem, the disease represents a malfunction, and the diagnosis is a fault identification.

The rest of the paper begin with reviewing knowledge representations (Section 2). We then briefly introduce the proposed framework (Section 3) and its new extension. Then, we provide further details on the KG enrichment by using the Symptoms Ontology (Section 4). In Section 5, we compare the previous version of our work with the current one. lastly, in section 6, we discuss our contribution and future work.

II. BACKGROUND: KNOWLEDGE REPRESENTATION

According to Davis [5], a Knowledge Representation (KR) serves five roles: as 1) a surrogate to enable an entity to determine the consequences of a plan or idea; 2) a set of ontological commitments about how and what to see in the world; 3) a fragmentary theory of intelligent reasoning; 4) a medium for efficient computation; and 5) a medium for human expression.

 ISBN: 978-1-68558-056-8

In this section, we review methods of KR: knowledge graphs, ontologies, and semantic technology.

A. Knowledge Graph

Knowledge Graphs (KG) represent information by converting data into a coded form, in particular by formulating relationships between entities into graph structures. KGs, also known as semantic graphs, generate interest among academic and industrial researchers, who deal with a wide variety of topics that all have the need to represent knowledge in common.

KGs have the property of providing semantically structured information. This property enables KGs to provide creative solutions for important tasks, such as answering questions [6], recommendation systems [7] and information retrieval [8]. Knowledge graphs are also considered to hold great promise for building smarter machines. KGs are also considered to offer great promise for building more intelligent machines.

B. Ontology

An ontology [9] is an explicit, machine-interpretable specification of a conceptualization—that is, the entities, or concepts, that are presumed to exist in some area of interest, their attributes, and the relationships amongst them. Ontology defines a common vocabulary for humans and machines that need to share information in a domain. The key reasons to develop ontologies includes [10]: 1) to enable the sharing of common understanding about the structure of information, among people or software agents; 2) to allow reusing of domain knowledge; and 3) to analyze domain knowledge.

C. Semantic Technology

Semantic technology represents a family of technologies that seek to derive meaning from information. That is, manage knowledge and join different data streams to perform inference. Representing knowledge is naturally done using the domain ontologies, and since the ontologies are based on a graph model, it is common to use a graph model to represent and store the data. By using graph representation for both the data and the domain knowledge, graph algorithms are used in order to infer new insights.

III. THE FRAMEWORK

In this section, we briefly introduce the proposed framework in [3], which includes a collection of algorithms and the flow between them. Then, we describe our extension to the framework, which is our current work.

We aim for interaction-based decision-making processes. The interaction is between a domain expert and an end-user, and results in a limited number of iterations consisting of questions that the framework suggests the domain expert ask the end-user. The decision-making process will progress according to the end-user's answers.

When we analyzed these types of processes, we concluded that they can be generically modeled as a collection of symptoms and diseases. Eventually, the process goal is to assist the domain expert to decide on a diagnosis (i.e., provide an explanation for a given set of symptoms based on analyzing available data). Questions that may arise during the diagnosis process are of the type: Does the end-user have a particular symptom?

The above terms (i.e., symptoms, diseases, questions, and diagnoses) produce a jargon that can naturally be used in the medical diagnostic domain, yet it is also suitable for other domains, such as appliance repairs: the symptom represents a problem, the disease represents a malfunction, the diagnosis is a fault identification, and a typical question can be: Does the end-user have a particular problem with his appliance?

In the rest of this section, we describe the framework presented in [3] along with its algorithms, and the extension of this work.

We start with building a knowledge graph from raw data, which will assist in exploring the relationships between diseases and symptoms. Following this, we use the Louvain hierarchical clustering [11] on the KG (Algorithm 1) to find communities (i.e., clusters of diseases that have similar symptoms). Then, given the symptoms reported by the end-user (called evidence symptoms), we find the possible diseases that are compatible with the evidence symptoms using inference on the KG (Algorithm 2). At this point, we infer the most probable community to include the end-user disease and suggest to the domain expert a question (symptom) that indicates this community (Algorithm 3). Lastly, we find the best diseases and symptoms that the end-user might have, to suggest to the domain expert (Algorithm 4), to address the improvement of the diagnostic process.

The whole framework is divided into two main parts: the first part, the pre-processing part, is carried out once the framework is launched; while the second part, the processing part, is carried out each time a new request arrives in the framework. This current work expands on our previous work. As mentioned, it semantically enriched the knowledge maintained within the framework. In particular, the new addition expanded the pre-processing part, in step 2 (See Figure 1 for the new architecture of the pre-processing part).

A. Pre-Processing Part

Input: A list of diseases and their symptoms.
Step 1: Construct a knowledge graph (KG) of diseases and symptoms. The left hand side of Figure 4 exhibits an example of a such KG.
Step 2: Enrich the KG with symptoms Ontology [12]. (See Section IV for more details).
Step 3: Cluster the diseases into groups (called communities), according to their symptoms: diseases with similar symptoms will be in the same community (Algorithm 1, from [3]).

B. Processing part

This part is presented in detail in [3].
Input: k evidence symptoms
Step 1: Find the most probable diseases: the possible diseases that are compatible with evidence symptoms (Algorithm 2).
Step 2: Infer and suggest to the domain expert (repeatedly as required) a question (symptom) that indicates the most probable community to include the end-user disease (Algorithm 3).

 ISBN: 978-1-68558-056-8

Step 3: Infer and suggest to the domain expert a list of diseases the end-user might have and their related questions (symptoms), sorted by relevance (Algorithm 4).

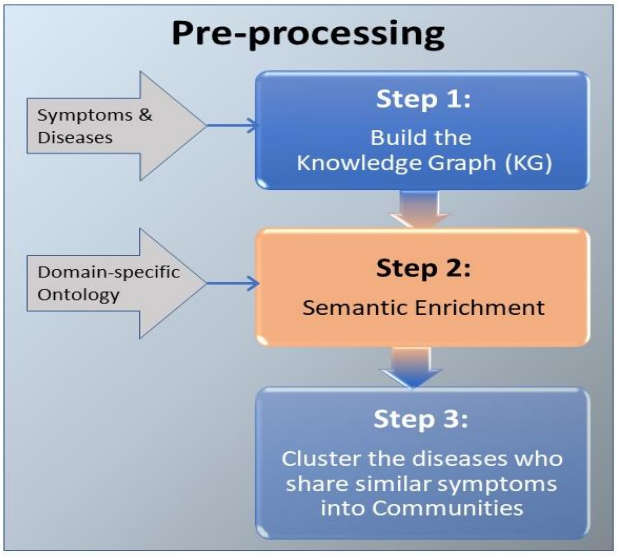

Figure 1. The architecture of the new pre-processing part

C. *Contribution of the Semantic Technology Extention*

Adding the ontology to the KG, as part of the *Pre-Processing* part (see Figure 2), has a main role in enriching the semantic knowledge of the domain expert. The KG is data driven knowledge, based on the historical examination of domain experts [13], and does not consist of the structure of the symptoms themselves (hierarchy). Adding this knowledge to the KG assists the recommendations process by inferring new relations, and thus inferring new relevant diseases to the domain expert.

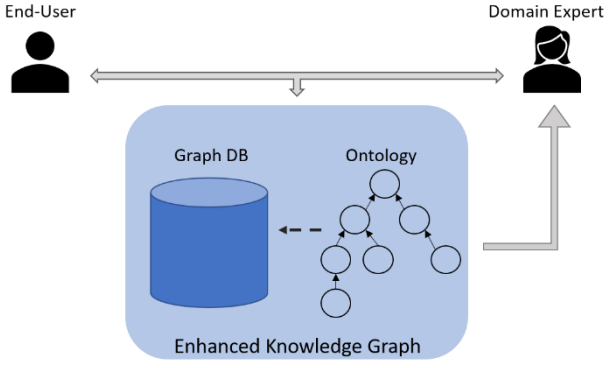

Figure 2. The Semantic Technology Architecture

IV. GRAPH ENRICHMENT ALGORITHM

In this section, we present Step 2 of the *Pre-Processing part*: the enrichment of the KG using the Symptoms Ontology [12]. The Symptoms Ontology (SYMP) consists of nodes representing the symptoms, and edges representing an isA relation between symptoms. Thus, the ontology represents the hierarchy of the symptoms. The right-hand side of Figure 4 exhibits an example of such ontology. After constructing the KG (step 1 of the *Pre-Processing* part), and storing it using Neo4j Graph Database, the ontology SYMP is added to the database, and then we perform the following procedures:

A. *Add Symptom Nodes to the KG*

- For all edges $e = (s_i, s_j)$ in SYMP, such that $s_j \in KG$ and $s_i \notin KG$:
  - Add s_i as a symptom node to KG.
- For all edges $e = (s_i, s_j)$ in SYMP, such that $s_i \in KG$ and $s_j \notin KG$:
  - Add s_j as a symptom node to KG

B. *Add isA Relations between Symptoms in the KG, according to the Ontology*

- For all edges $e = (s_i, s_j)$ in SYMP, such that $s_j \in KG$ and $s_i \in KG$:
  - Add the edge (si, sj) to KG, labeled *isA*.

Figure 3 presents the legend we use in Figure 4 and Figure 5. Figure 4 and Figure 5 present the construction and the integration of the ontology into the KG.

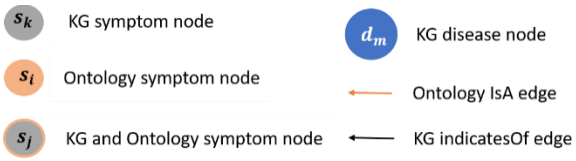

Figure 3. Legend

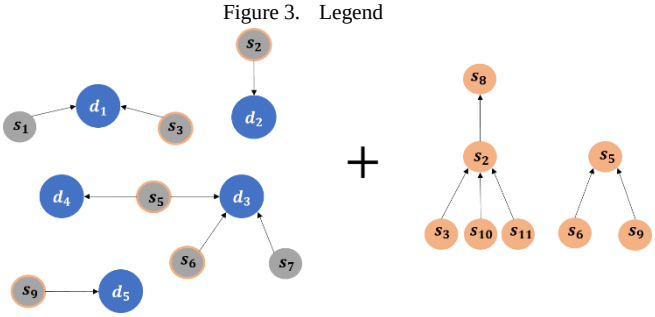

Figure 4. On the left side the KG, on the right side the Ontology

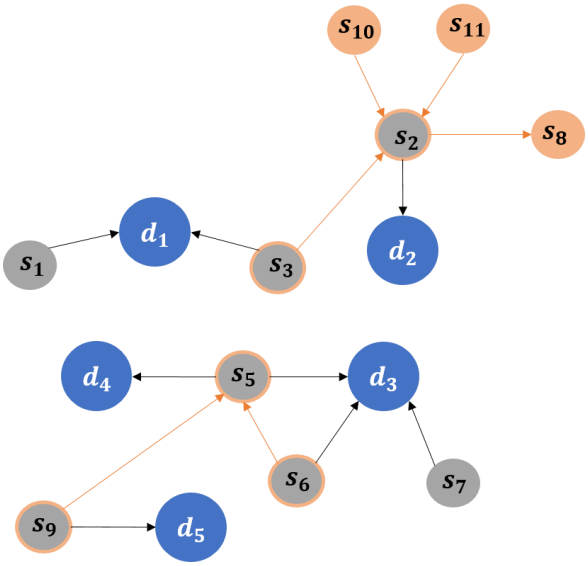

Figure 5. The Knowledge Graph Enrichment

V. COMPARING THE PREVIOUS WORK WITH THE CURRENT

The new edges that were added to the knowledge graph (the ontological edges - painted in orange), semantically enriched the space of questions, and are used by the domain expert. In the previous work, the selected symptoms were those that reinforce the most probable disease (independent of the other symptoms the patient has). In the current work, the symptoms that will be examined are those that will strengthen the most probable disease, and are also semantically related to the other symptoms of the patient.

To illustrate the influence of this semantic enrichment, let's consider the following scenario: A patient arrives with the following two symptoms: s_1, s_2 and s_5. These are our evidence symptoms. Therefore, the patient's probable diseases are d_1, d_2, d_3 and d_4. Since s_2 is an evidence, and s_3 isA s_2, the symptom s_3 is more likely to be considered to the domain expert in the hypothesis that d_1 is the patient's disease. In addition, since s_5 is an evidence, and s_6 isA s_5, the symptom s_6 is more likely to be considered to the domain expert in the hypothesis that d_3 is the patient's disease.

VI. CONCLUSION AND DISCUSSION

This section summarizes our results including our contribution, and present our future work.

A. Summary

In most areas of life, one can find decision-making processes, which makes this topic interesting for relevant research. At the same time, since these processes are found in many worlds of content, there is a wide and rich variety of decision-making processes, characterized by different needs. Therefore, in any attempt to support this topic, we must focus on a specific subtopic, characterized by specific requirements.

In the current (and ongoing) work, we focus on decision-making processes with the following configuration: an end-user and a domain expert are involved in the process, which establishes an interaction between them, consisting of questions and answers, to address a problem of the end-user. The domain expert uses the suggested framework to make the interaction as limited as possible (in time, the number of questions, money, etc.).

B. Contribution

In our previous work [3], we introduced for the first time the framework we built, including a detailed description of the algorithms that we developed as part of the framework, which enable inference of big data. The innovation of [3] stems from the use of semantic technologies, including a graphical data model, combined with unique algorithms.

In the current work, we introduce an extension to our framework, such that a domain-specific ontology is integrated into the knowledge graph, and hence expands the space of questions the domain expert can ask, resulting in a more accurate inference algorithm.

C. Future work

We want to develop the current research, in particular, to explore the contribution of the ontology to the decision-making process, and to run a case study on the knowledge graph we created in the previous study, after incorporating the ontology into that graph.

In addition, we wish to explore the possibility of using weighted edges in the knowledge graph for representing the cost of each question.

REFERENCES

[1] I. A. T. Hashem, et al., "The rise of "big data" on cloud computing: Review and open research issues," Inf. Syst. 47, pp. 98–115, 2015.

[2] D. J Power, "Decision Support Systems: Concepts and Resources for Managers," Greenwood Publishing Group: Westport, CT, USA, 2002.

[3] S. Albagli-Kim and D. Beimel, "Knowledge Graph-based Framework for Decision-making Process with Limited Interaction," Mathematics | Special Issue: From Edge Devices to Cloud Computing and Datacenters: Emerging Machine Learning Applications, Algorithms, and Optimizations. 10 (21),pp. 3981,2022.

[4] I. Robinson, J. Webber, and E. Eifrem, "Graph Databases: New Opportunities for Connected Data," O'Reilly Media, Inc.: Middlesex County, MA, USA, 2015.

[5] R. Davis, H. Shrobe, and P. Szolovits, "What is a Knowledge Representation?", AI Magazine, vol. 14, no. 1, pp. 17-33, 1993.

[6] A. Gashkov, A. Perevalov, M. Eltsova, and A. Both, "Improving Question Answering Quality through Language Feature-Based SPARQL Query Candidate Validation," The Semantic Web. ESWC 2022. Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13261. https://doi.org/10.1007/978-3-031-06981-9_13.

[7] Q. Guo, et al., "A survey on knowledge graph-based recommender systems," IEEE Trans. Knowl. Data Eng. 34, pp. 3549–3568, 2020.

[8] L. Dietz, A. Kotov, and E. Meij, "Utilizing knowledge graphs for text-centric information retrieval," In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, pp. 1387–1390, July 2018.

[9] T. R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," Intl J Human-Computer Studies, vol. 43, no. 5-6, 1995.

[10] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Medical Informatics Technical Report SMI-2001-0880 2001.

[11] H. Lu, M. Halappanavar, and A. Kalyanaraman, "Parallel heuristics for scalable community detection," Parallel Computing, 47. pp. 19-37, 2015

[12] Symptom Ontology, Ontology Lookup Service (OLS), https://www.ebi.ac.uk/ols/ontologies/symp, retrieved: 02,2023

[13] Kaggle, https://www.kaggle.com, retrieved: 02,2023

 ISBN: 978-1-68558-056-8

An Efficient Ensemble of Deep Neural Networks for Detection and Classification of Diabetic Foot Ulcers Images

Basabi Chakraborty
Research and Regional Cooperation Division
Iwate Prefectural University, Takizawa, Japan
Computer Science and Technology
Madanapalle Institute of Technology and Science
AP, India
email: basabi@iwate-pu.ac.jp

Suma Sailaja Nakka
Software and Information Science
Iwate Prefectural University
Takizawa, Japan
email: g231t026@s.iwate-pu.ac.jp

Takahisa Sanada
Software and Information Science
Iwate Prefectural University
Takizawa, Japan
email: takahisa_s@iwate-pu.ac.jp

***Abstract*—Classification of Diabetic Foot Ulcers (DFU) wounds using computerized methods is becoming an important research area due to development of machine learning and deep learning algorithms for image classification. In this work an efficient ensemble of several deep neural networks has been proposed for classification of DFU images. Simulation experiments with publicly available Diabetic Foot Ulcers Grand Challenge (DFUC 2021) data set has been done to justify the proposal. The performance of the ensemble has been studied and it is found that the ensemble produced a classification accuracy of 91% with a reasonable computational cost which is considered higher compared to the existing approaches.**

***Index Terms*—Diabetic foot ulcers classification, deep neural network, ensemble classifier**

I. Introduction

Deep Neural Networks (DNN) are increasingly used in medical image analysis as a part of the development of computer aided diagnosis systems. Automatic detection and classification of Diabetic Foot Ulcers (DFU) wounds using deep learning tools is one of such applications. DFU is one of the major complications of diabetes. DFU with infection and ischemia can be a serious threat to the patient, leading to death. Early detection and classification of wounds according to the presence of infection, ischemia or both is needed for successful treatment. A comprehensive assessment of several techniques for detection of DFU based on Diabetic Foot Ulcers Grand Challenge (DFUC) 2020 dataset is reported by Yap et al. [1]. The DFUC 2020 classification results reported in various literature so far are summarized by Zhang et al. [2]. Most of the research works on binary classification of DFUC 2020 by deep networks are based on Convolutional Neural Network (CNN) architecture and its several variants. Goyal et al. proposed DFUNet based on CNN [3] and an ensemble of CNN and SVM [4] for binary classification of normal and DFU images and infection vs non-infection, ischemia vs non-ischemia, respectively. Xu et al. [5] used vision transformer model and Das et al. [6] used ResNet for binary classification of infection vs non-infection and ischemia vs non-ischemia classes. Diabetic Foot Ulcers Grand Challenge (DFUC) 2021 results are summarized by Cassidy et al. [7]. DFUC 2021 data set contains four classes of DFU wounds i.e., with infection, with ischemia, with both and none. Several deep learning models and their ensembles are proposed by challengers and the best F1 score on test data were reported as 0.63%.

In this work, several deep neural network models are used for four class classification task of DFUC 2021 data set. The performance of each model has been assessed using several metrics like classification accuracy, precision, recall and F1 score. Finally, an ensemble of top 3 individual neural network models is proposed for the classification task and the performance study by simulation experiments has been done. In section 2, a brief description of the data set and deep neural networks are presented. Section 3 describes the simulation experiments and results followed by section 4 containing conclusion and future work.

II. Data set and Deep Neural Networks

A. Data set

The Diabetic Foot Ulcers data set (DFUC2021) reported by Yap et al. [8] is a pathology analysis data set focusing on infection and ischemia. The final release of DFUC2021 consists of 15,683 DFU patches, with 5,955 training, 5,734 for testing, and 3,994 unlabeled DFU patches. The data set consists of images of four distinct classes which are infection, ischemia, both infection and ischemia, none. Among 5955 training samples, 2,552 samples are without ischemia and infection, 2,555 samples are infection only, and 621 samples are both infection and ischemia, 227 samples are ischemia only. As the number of samples in 4 classes are not balanced, image augmentation techniques (oversampling) are used to increase the number of samples in the classes having less number of samples. Simulation experiment is performed on the both imbalanced and balanced data sets of DFU images.

B. Deep Neural Networks

In this study, popular deep neural networks for various image classification tasks are used. They are VGG16, VGG19,

 ISBN: 978-1-68558-056-8

TABLE I
RESULTS FOR IMBALANCED DATA

Model	Accuracy	Precision	F1 Score	Time
VGG16	0.69	0.50	0.52	20:12
VGG19	0.70	0.54	0.57	15:50
ResNet50	0.53	0.35	0.34	23:22
EfficientNetB0	0.42	0.18	0.25	29:08

ResNet50, DenseNet121, InceptionV3, EfficientNetB0 and EfficientNetB7.

III. SIMULATION EXPERIMENTS AND RESULTS

A. Simulation Experiments

Simulation experiments are done with training set images. All the images are used as RGB images and were resized to a size of $224 \times 224 \times 3$ and normalized. All the network models used in this study are initially pretrained with ImageNet data set. The pre-processed DFU images are then fed to different network models with ImageNet weights. Each model is trained for 20 epochs and batch size of 32 for both imbalanced and balanced data using train-test-validation-split of 70%,20%,10% for training, testing and validation of images respectively. Oversampling of minority classes of the data set is done to increase the number of samples in minority classes equal to majority sample classes. Several image processing techniques like rotating the training images by random rotation angles, horizontal and vertical flips, contrast and brightness, shearing, and zooming in and out of the images have been done as oversampling. Finally, 5-fold Stratified K-fold Cross Validation with k=5 has been done with Stochastic Gradient Descent (SGD) as optimizer for 50 epochs for the balanced data set. To avoid the overfitting of the models we used dropout rate of 50% in all the layers.

B. Simulation Results

Table I represents the performance of some of the models in terms of classification accuracy, precision, F1 score and computational time in min:sec with original data set without balancing with a train-test-validation-split of 70%,20%,10%. VGG19 seems to be the best model according to classification accuracy. After using balanced data sets, the classification accuracy increased to 0.76 for VGG16, 0.79 with VGG19, 0.61 for ResNet50, 0.59 with ResNet101 and 0.54 for EfficientNetB0.

Table II represents the performance of all the models in terms of classification accuracy, Precision, F1score and computational time in hours:min after fine tuning of the models and using Stratified 5-fold Cross Validation technique with 50 epochs over balanced data set. It is found that the top three performing deep network models are DenseNet121, InceptionV3 and VGG19 respectively. An ensemble of the above three models is proposed in this study in order to achieve better results. The initial layers of individuals models are locked with weights pretrained by ImageNet, the last layers are trained by training set followed by the full connected layers for averaging the output prediction of the three independent models. Finally, a softmax activation layer is used for the output class. The classification performance of the ensemble model is reported in the last line of Table II. It is found that the classification accuracy of the ensemble model is higher than the individual models.

TABLE II
RESULTS FOR BALANCED DATA

Model	Accuracy	Precision	F1 Score	Time
VGG16	0.779	0.794	0.776	3:54
VGG19	0.842	0.844	0.841	3:20
ResNet50	0.785	0.786	0.785	3:55
DenseNet121	0.891	0.891	0.891	3:06
EfficientNetB7	0.544	0.469	0.503	3:57
EfficientNetB0	0.433	0.242	0.276	3:58
InceptionV3	0.863	0.865	0.863	3:17
Ensemble	**0.916**	0.917	0.916	4:04

IV. CONCLUSION AND FUTURE WORK

Automatic classification of DFU wounds help doctors in early detection of severity of the disease and save time for treatment. This paper examines several deep neural network models for their effectiveness in four class classification of DFUC 2021 data set and an ensemble of top ranking models is proposed. It has been found that the classification accuracy of the ensemble model is higher than the individual models and reported research works in this problem. But the computational cost is higher than the individual models. The model should be tested for other DFU data sets and compared to other existing high performing models as the future work.

ACKNOWLEDGMENT

This research was partly funded by Japan Society of Promotion of Science (JSPS) KAKENHI Grant Number JP 20K11939.

REFERENCES

[1] M. H. Yap et al., “Deep learning in diabetic foot ulcers detection: A comprehensive evaluation,” Computers in Biology and Medicine, Vol. 135, 104596, Aug. 2021.

[2] J. Zhang et al., ”A comprehensive review of methods based on deep learning for diabetes-related foot ulcers”, Front Endocrinol (Lausanne). 2022 Aug 8;13:945020.

[3] M. Goyal, et.al.,”Dfunet: Convolutional neural networks for diabetic foot ulcer classification. IEEE Trans Emerg Topic Comput Intell vol.4(5), pp. 728– 739, 2018.

[4] M. Goyal, et.al. “Recognition of ischaemia and infection in diabetic foot ulcers: Dataset and techniques,” Computers in Biology and Medicine, vol. 117, no. 103616, 2020.

[5] Y. Xu, et.al., ”Classification of diabetic foot ulcers using class knowledge banks”, Front Bioengineer Biotechnol, Vol 9, 2021. doi:10.3389/fbioe.2021.811028

[6] S. K. Das, P. Roy and A. K. Mishra, ”Recognition of ischaemia and infection indiabetic foot ulcer: a deep convolutional neural network based approach. Int J. Imaging Syst Technol Vol. 32(1),pp.192–208, 2022.

[7] B. Cassidy et.al., ”Diabetic Foot Ulcer Grand Challenge 2021: Evaluation and Summary”, Diabetic Foot Ulcers Grand Challenge: Second Challenge, DFUC 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. Springer-Verlag, Berlin, Heidelberg.

[8] M. H. Yap, B. Cassidy, J. M. Pappachan, C. O’Shea, D. Gillespie and N. D. Reeves, “Analysis Towards Classification of Infection and Ischaemia of Diabetic Foot Ulcers,” arXiv:2014.03068, 2021.

 ISBN: 978-1-68558-056-8