DBKDA 2021

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

ISBN: 978-1-61208-857-0

May 30th – June 3rd, 2021

DBKDA 2021 Editors

Malcolm Crowe, University of the West of Scotland, UK

Fritz Laux, Reutlingen University, Germany

Andreas Schmidt, University of Applied SciencesKarlsruhe Institute of Technology, Germany

Cosmin Dini, IARIA, EU/USA

DBKDA 2021

Foreword

The Thirteenth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2021), held between May 30 – June 3rd, 2021, 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 2021 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 2021. 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 2021 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that DBKDA 2021 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.

DBKDA 2021 Chairs:

DBKDA 2021 Steering Committee
Fritz Laux, Reutlingen University, Germany
Andreas Schmidt, Karlsruhe Institute of Technology / University of Applied Sciences
Erik Hoel, Esri, USA
Lisa Ehrlinger, Johannes Kepler University Linz, Austria / Software Competence Center Hagenberg GmbH, Austria
Peter Kieseberg, St. Pölten University of Applied Sciences, Austria

DBKDA 2021 Industry/Research Advisory Committee
Jerzy Grzymala-Busse, University of Kansas, USA
Filip Zavoral, Charles University Prague, Czech Republic

Konstantinos Kalpakis, University of Maryland Baltimore County, USA
Shin-ichi Ohnishi, Hokkai-Gakuen University, Japan
Thomas Triplet, Ciena inc. / Polytechnique Montreal, Canada
Stephanie Teufel, iimt - international institute of management in technology | University of Fribourg, Switzerland
Rajasekar Karthik, Oak Ridge National Laboratory, USA

DBKDA 2021 Publicity Chairs

Daniel Basterretxea, Universitat Politecnica de Valencia, Spain
Marta Botella-Campos, Universitat Politecnica de Valencia, Spain

DBKDA 2021

Committee

DBKDA 2021 Steering Committee

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

DBKDA 2021 Industry/Research Advisory Committee

Jerzy Grzymala-Busse, University of Kansas, USA
Filip Zavoral, Charles University Prague, Czech Republic
Konstantinos Kalpakis, University of Maryland Baltimore County, USA
Shin-ichi Ohnishi, Hokkai-Gakuen University, Japan
Thomas Triplet, Ciena inc. / Polytechnique Montreal, Canada
Stephanie Teufel, iimt - international institute of management in technology | University of Fribourg, Switzerland
Rajasekar Karthik, Oak Ridge National Laboratory, USA

DBKDA 2021 Publicity Chairs

Daniel Basterretxea, Universitat Politecnica de Valencia, Spain
Marta Botella-Campos, Universitat Politecnica de Valencia, Spain

DBKDA 2021 Technical Program Committee

Julien Aligon, Institut de Recherche en Informatique de Toulouse (IRIT) | Université Toulouse 1 Capitole, France
Alaa Alomoush, University Malaysia Pahang, Malaysia
AbdulRahman A. Alsewari, Universiti Malaysia Pahang, Malaysia
Emmanuel Andres, Hôpitaux Universitaires de Strasbourg, France
Zeyar Aung, Masdar Institute of Science and Technology, UAE
Gilbert Babin, HEC Montréal, Canada
Flavio Bertini, University of Bologna, Italy
Ali Boukehila, University of Annaba, Algeria
Zouhaier Brahmia, University of Sfax, Tunisia
Martine Cadot, LORIA, Nancy, France
Ricardo Campos, Polytechnic Institute of Tomar, Portugal
Sanjay Chaudhary, AhmedabadUniversity, India
Yung Chang Chi, National Cheng Kung University, Taiwan
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, Johannes Kepler University Linz, Austria / Software Competence Center Hagenberg GmbH, Austria
Amir Hajjam El Hassani, Université de Technologie de Belfort-Montbéliard, France
Gledson Elias, Federal University of Paraíba (UFPB), Brazil
Hannes Fassold, JOANNEUM RESEARCH - DIGITAL, Graz, Austria
Barbara Gallina, Mälardalen University, Sweden
Ana González-Marcos, Universidad de La Rioja, Spain
Luca Grilli, University of Foggia, Italy
Vibhuti Gupta, Meharry Medical College, USA
Robert Gwadera, Cardiff University, UK
Mohammed Hamdi, Najran University, Saudi Arabia
Hamidah Ibrahim, Universiti Putra Malaysia, Malaysia
Md Johirul Islam, Iowa State University, USA
Vladimir Ivančević, University of Novi Sad, Serbia
Ivan Izonin, Lviv PolytechnicNational University, Ukraine
Tahar Kechadi, University College Dublin (UCD), Ireland
Jam Jahanzeb Khan Behan, Université libre de Bruxelles (ULB), Belgium / Universidad Politécnica de Cataluña (UPC), Spain
Daniel Kimmig, solute GmbH, Germany
Sotirios I. Kontogiannis, University of Ioannina, Greece
Katrien Laenen, KU Leuven University, Belgium
Nadira Lammari, CEDRIC-Cnam, France
Aida Kamisalic Latific, University of Maribor, Slovenia
Friedrich Laux, Reutlingen University, Germany
Martin Ledvinka, Czech Technical University in Prague, Czech Republic
Yuening Li, Texas A&M University, USA
Tobias Lindaaker, Neo4j, Sweden
Chunmei Liu, Howard University, USA
Yanjun Liu, Feng Chia University, Taiwan
Francesca Maridina Malloci, University of Cagliari, Italy
Michele Melchiori, Università degli Studi di Brescia, Italy
Fabrizio Montecchiani, University of Perugia, Italy
Francesc D. Muñoz-Escoí, Universitat Politècnica de València (UPV), Spain
Roberto Nardone, University of Reggio Calabria, Italy
Nikola S. Nikolov, University of Limerick, Ireland
Joshua C. Nwokeji, Gannon University, Erie Pennsylvania, USA
Shin-ichi Ohnishi, Hokkai-Gakuen University, Japan
Taher Omran Ahmed, College of Applied Sciences, Ibri, Sultanate of Oman / Azzentan University, Libya
Moein Owhadi-Kareshk, University of Alberta, Canada
Shirish Patil, Sitek Inc., USA
Fabiano Pecorelli, University of Salerno, Italy
Elaheh Pourabbas, National Research Council | Institute of Systems Analysis and Computer Science

"Antonio Ruberti", Italy
Manjeet Rege, University of St. Thomas, USA
Peter Revesz, University of Nebraska-Lincoln, USA
Jan Richling, South Westphalia University of Applied Sciences, Germany
Peter Ruppel, CODE University of Applied Sciences, Berlin, Germany
Andreas Schmidt, Karlsruhe Institute of Technology / University of Applied Sciences Karlsruhe, Germany
Jaydeep Sen, IBM Research AI, India
Zeyuan Shang, Einblick Analytics, USA
Fatemeh Sharifi, University of Calgary, Canada
Ankur Sharma, Saarland University, Germany
Virach Sornlertlamvanich, Musashino University, Japan / Thammasat University, Thailand
Carmine Spagnuolo, Università degli Studi di Salerno, Italy
Günther Specht, University of Innsbruck, Austria
Sergio Tessaris, Free University of Bozen-Bolzano, Italy
Nicolas Travers, ESILV - Pôle Léonard de Vinci, Paris, France
Thomas Triplet, Ciena inc. / Polytechnique Montreal, Canada
Maurice van Keulen, University of Twente, Netherlands
Chenxu Wang, Xi'an Jiaotong University, China
Shaohua Wang, New Jersey Institute of Technology, USA
Shibo Yao, New Jersey Institute of Technology, USA
Damires Yluska Souza Fernandes, Federal Institute of Paraíba, Brazil
Feng Yu, Youngstown State University, USA
Qiang Zhu, University of Michigan - Dearborn, USA

Copyright Information

Table of Contents

A Survey on Algorithms for Big Data Analysis in Electromagnetics Scattering Problems

Christian O. Díaz-Cáez
Department of Electrical Engineering
and Computer Science
Howard University
Washington, D.C. USA
e-mail: christian.diaz@bison.howard.edu

Chunmei Liu
Department of Electrical Engineering
and Computer Science
Howard University
Washington, D.C. USA
e-mail: chuliu@howard.edu

***Abstract*— Computational Electromagnetics is a discipline that deals with the processing and modeling of multi-physics and electromagnetic problems. Thanks to the advent of computers and numerical methods, engineers today can develop algorithms and software to solve Maxwell's equations numerically. The electromagnetic scattering problem leads to a very large system of equations with millions or even billions of unknowns; traditional data analysis methods are oftentimes not efficient enough to handle the problem due to data volume. The field of Big Data has emerged from the need to process a massive amount of data and is a research area that facilitates the complex work of extremely large data sizes. Fast algorithms can be developed to efficiently manage the Big Data approach to support areas of science and engineering. In this paper, we explore an application of Big Data and algorithms in computational electromagnetics scattering problems.**

Keywords—Big Data; Computational Electromagnetics (CEM); Method of Moments (MoM); Fast Algorithms, Multilevel Fast Multipole Algorithm (MLFMA).

I. Introduction

We are currently in an era of digital information. This means that a great amount of information is generated daily. To manage, analyze and store this information, very powerful tools are needed. Big Data technology plays a very important role in this area. It allows large companies to optimize decision-making and obtain results optimally. Big Data is a term used to describe a set of data or combinations of sets of data whose size, complexity, and velocity of growth make it difficult to capture, manage, process or analyze using conventional technologies and tools, such as relational databases and conventional statistics or visualization package, within the time necessary for them to be useful [1]. Although there is no firmly defined size for determining whether a data set is Big Data, and the definition continues to change over time, professionals currently refer to Big Data to be datasets ranging from 30-50 Terabytes to several Petabytes [1].

For some problems, the data size may be so large that it does not fit in the main memory of a single machine. The need to process such a huge amount of data There is a need to process such a huge amount of data through efficient algorithms in machine learning, network traffic monitoring, scientific computing, signal processing, and other areas. Some well-known examples of such algorithms are numerical linear algebra algorithms for big matrices [2] (regression, low-rank approximation, matrix completion), dimensional reduction for reducing data dimension to conserve the geometric structure [3], compressed sensing for approximation recovery of sparse signals [4] and sparse Fourier Transform as fast algorithms for signals calculation in a frequency domain [5]. To better understand Big Data's difficulty, it is often broken down using five V's: Volume, Velocity, Value, Variety, and Veracity [8][9].

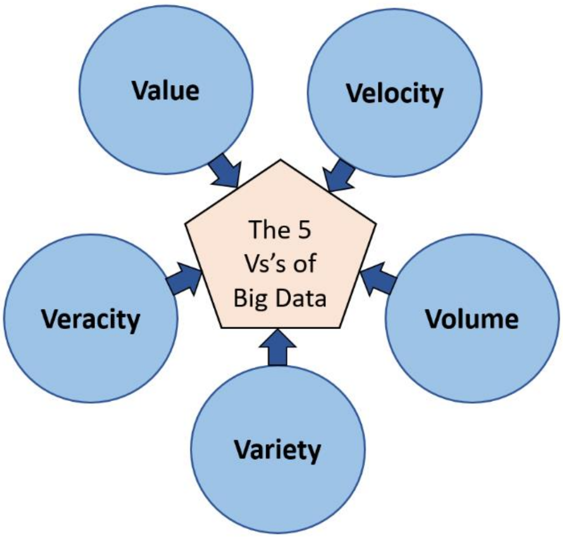


Figure 1. The 5 V's of Big Data.

The 5 V's of Big Data illustrated in Figure 1 can be defined as follows:

- Volume refers to the exponential increase in data resulting from new technologies, and the ease of generating digital data is a palpable reality. The volume means large size.
- Velocity is the rate of growth and how fast data is gathered for analysis.
- Value is indicative of substantial value, including the ability to understand the target better, accordingly, and optimize performance.
- Variety is information about the various types of data, such as structured, unstructured, semi-structured, etc.
- Veracity means the confidence established about the data to be used.

Big Data serves the purpose of converting data (information) into knowledge. Researchers have added more dimensions from 5 to 10 [6], covering terms such as validity, vulnerability, volatility, visualization, variability, and even more, which can be found in technology and data generation advances [7]. The rest of this paper is organized as follows: Section II describes Computational Electromagnetic (CEM) as an interdisciplinary field, Section III describes the Method of Moments (MoM) as a powerful numerical technique in CEM, Section IV addresses the algorithm techniques to exploit MoM, Section V describes the multilevel fast multipole algorithm and Section VI summarize some Big Data techniques implemented to solve different electromagnetic engineering problems. The conclusions close the article.

II. Computational Electromagnetic

Electromagnetic (EM) analysis is a discipline that solves Maxwell's equations to obtain a better understanding of complex systems. The advent of numerical methods and computers has changed the traditional ways of EM analyzing, and a field called Computational Electromagnetics has emerged [10]. It is a prominent EM research area that involves the modeling of the interaction of EM fields with physical objects, the study of electromagnetic compatibility between equipment in different environments, the design of antennas, the design of passive microwave circuits and components, the calculation of the Radar Cross Section (RCS) and Inverse Synthetic Aperture Radar (ISAR) images, the analysis of antennas embarked on complex structures, Doppler analysis, and radio propagation both indoors and outdoors.

When an EM problem is given for a practical application, we need to describe our problem mathematically based on EM physics to seek a numerical method. We can apply Partial Differential Equations (PDEs) and boundary conditions to define an equivalent boundary-value problem. Then, from our mathematical formulation, we can develop a numerical method effectively, and depending on the problem, we will need to decide to use an existing method or develop a new one addressing the problem. After a numerical method is selected or developed, it is necessary to develop an efficient computer program for implementation. Finally, after the computer program is validated, we can use it to solve the problem given by constructing a geometrical model and the specification of EM mediums (permittivity, permeability, and conductivity) [11]. All the steps previously discussed are summarized in Figure 2.

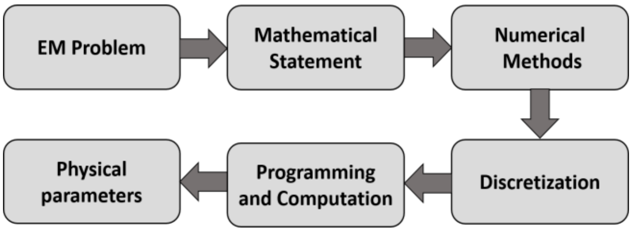


Figure 2. Numerical analysis steps for solving engineering problems.

As shown in Figure 3, CEM is a highly interdisciplinary field that combines physics, mathematics, and computer science to advance engineering applications.

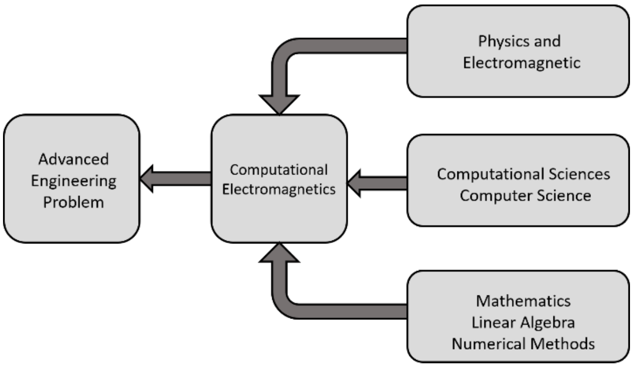

Physics and Electromagnetic
Advanced Engineering Problem
Computational Electromagnetics
Computational Sciences Computer Science
Mathematics Linear Algebra Numerical Methods

Figure 3. CEM is an interdisciplinary field for advancing engineering applications.

Today, numerical methods for EM scattering problems need to process a very large system of equations with millions or even billions of unknown variables [12]. Traditional methods are inefficient and fast algorithms in EM have been developed to solve this problem in an efficient manner [10]. As a common method, we can represent our system of unknowns as a hierarchical representation with a matrix system of N number of unknowns. Fast algorithms use *O(NlogN)* memory and approximately *O(N)* or even *O(logN)* time [12]. Traditional numerical methods usually require $O(N^2)$ memory and $O(N^2)$ time so in the scenario that N becomes very large, we can identify a huge discrepancy in memory and time between traditional and fast algorithms [10].

The next section describes an efficient algorithm for electromagnetic scattering problems that can be implemented in multicore-based and cluster architectures. Electromagnetics simulations are critically important in several application areas, such as antenna design for aircraft, satellites, and medical devices. We can reduce the numerical formulation cost by assuming time-harmonic solutions and reformulating Maxwell's equations to describe EM waves in terms of surface currents. The result of this approach is a numerical problem that can be solved on the surface of the object being studied.

III. Method of Moments

The method of moments is a very powerful numerical technique developed for solving complex EM problems. Compared to the Finite Element Methods (FEM), MoM also transforms the boundary-value problem into a matrix equation that can be solved on computers [13]. Mathematical-based MoM was proposed almost one century ago, but its applications did not arise until 1960s [14]. Today, it is one of the most important methods in CEM. MoM has been well studied on open-region electromagnetic problems, such as wave scattering and antenna radiation, and it is very efficient for problems involving either impenetrable or homogeneous objects [13]. Also, the capability of MoM has been improved by the development of fast algorithms that can deal with huge MoM matrix equations [12]. MoM forces the boundary conditions to be satisfied in an average sense over the entire surface.

We can see a system of equations to compute the surface currents as an inverse problem. Applying an iterative method, the inverse problem is converted to repeated solutions of the forward problem. For example, a basic problem in EM consists of computing an EM field, given the distribution of sources/charges. The forward model is well known to compute electrostatic potential

$$\Phi = \sum_{j=1}^{N} K(x, x_j)\; q_j \tag{1}$$

where q_j is the point charge at the location represented by x_j.The interaction between the field points and the charges is represented by the kernel $K(x, x_j)$ which is logarithmic in two dimensions and proportional to the inverse distance in three dimensions.

The corresponding scattering problem computes electric and magnetic fields **E** and **H** generated by surface currents on metallic objects, such as aircraft [21]. Avoiding its mathematical derivation, a simplified form is given by

$$E(r) = \int_{\partial\Omega} G(r, r')j(r') + \frac{1}{k^2}\, \nabla\, (G(r, r')\nabla \cdot j(r')\, dr' \tag{2}$$

where $\partial\Omega$ is the surface of the object, and for computer simulation, it is discretized, ***r* is** a point in the space and $G(r, r')$ is the *Green's function* representing a point source response [21]. Figure 4 shows a visualization of an example of a discretized unit sphere. To make this type of problem solvable by computers, we need to discretize the object in *N* number of pieces. We can represent the sources and fields of the surface current by a set of *basis functions* and corresponding coefficients to approximate the solution of the surface current [19]. After the discretization, we convert the problem to a matrix equation by intruding on another set of functions called *testing* or *weighing functions* [19].

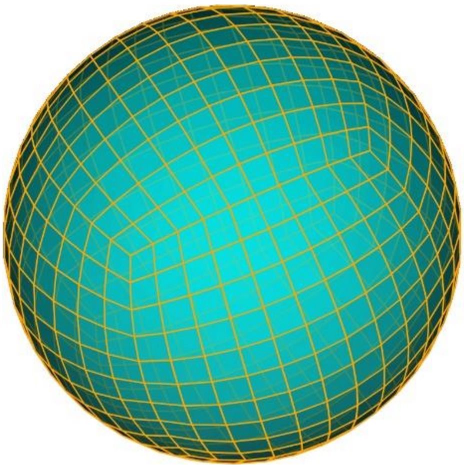


Figure 4. Discretization of a unit sphere in small patches.

It can be expressed in a compact form as

$$\sum_{j=1}^{N} Z_{ij} I_j = V_i \qquad i = 1,2,3 \ldots N \tag{3}$$

where Z_{ij} is the *N* x *N* matrix system with the unknown coefficients, I_j is the vector of unknowns, and V_i is the source vector.

IV. Fast Algorithms

Unlike FEM based on PDEs that yield to huge sparse matrix system, the method of moments, MoM, based on integral equations (IEs), produces a fully populated matrix system because of the applications of the Green's function. Now, the problem is the high complexity associated with methods for the full matrix solution. It becomes a limitation on the capability of MoM. In conventional methods for matrix solutions, such as Gaussian elimination or lower-upper (LU) decomposition, the time complexity is *O*(N^3) and the space complexity is *O*(N^2), where *N* is the matrix dimension. An iterative method can reduce the time complexity to *O*(N^2), but the memory remains the same for a direct method. The total time complexity is *O*(N_{iter} N^2) where N_{iter} is the number of iterations reaching a certain convergence. If N_{iter} is small, then an interactive process will be faster than LU decomposition just for the right-hand side of the equation, but the iterative solution must be repeated for every right-hand size [15][20], which makes MoM limited to one-, two- or three-dimensional problems.

A better understanding of the high computational complexity of traditional direct and interactive methods can be found in [15]. The complexities of *O*(N^3)and *O*(N^2) make the time and space increase dramatically with the increase of the number *N*, and it may exceed the capabilities computers have today. A technique used to reduce time and memory complexities for iterative methods, especially for large-scale problems, is called *fast algorithms*. We can broadly define fast algorithms as algorithms that can solve both matrix and integral equations that can be discretized in a matrix equation by MoM. More details regarding fast algorithms can be found in [10][14][15]. Some examples of fast algorithms are the Conjugate Gradient–FFT (CG-FFT) method, the Adaptive Integral Method (AIM), the Fast Multipole Method (FMM), and the Adaptive Cross-Approximation (ACA) method.

For this survey, we focus on FMM because it is the base for the technique presented in the next section of this paper. FMM divides the current elements into groups by their physical locations in space. A group is then defined as a collection of current elements near each other. Figure 5 illustrates an example of an arbitrary object with basis functions divided into groups, so the computation of far fields that is calculated indirectly in multiples steps is made fast, whereas near fields are computed directly (more quickly).

FMM integrates a new concept of decomposing the MoM matrix into near-and-far-interaction components. It makes a fast matrix-vector calculation possible by multipole or plane wave expansions and eventually reduces the computational complexity to *O(NlogN)* [18].

V. Multilevel Fast Multipole Algorithm

For a problem with *N* unknowns, we can divide them into *N/M* groups. For near-fields interactions, aggregations, and disaggregation, *O(NM)* operations are required, whereas the calculation of translation requires $O(N^2/\mathrm{M})$ operations. A small or large number of groups *M* will improve the complexity performance of the operation count for the calculation of near-fields interactions or translation calculation. An optimal choice of *M* is *M* proportional to $\sqrt{N}$ and the operation count in each calculation is balanced to $O(N^{3/2})$ [15]. We can apply FMM to each group; if we have small groups and each group has only a few basis functions, the calculation of near-fields interaction will be only *O*(*N*), and the same will be the case for aggregation and desegregation [15]. To reduce the translation calculation, when the groups are far from each other, we aggregate the field from the center of a group to another large group and designate the received field to the groups residing in the second larger group. This process reduces the translation counts, and this idea can be extended to multiple levels until there are no far-apart groups among the highest-level group. The algorithm that results from all this procedure is called Multilevel Fast Multipole Algorithm (MLFMA) [10].

In [10] and [15], the authors introduce a comparison example of a telephone communication scenario to understand how FMM and MLFMA work. We can consider a network with *N* telephones. Imagine that all the telephones are directly connected. In that case, we will need N^2 telephone lines. If we divide the telephones into groups according to their proximity to each other, and then connect all the telephones in the same group to a single hub, and then connect the hubs, we can reduce the number of telephones lines to $O(N^{3/2}log\ N);$ this is basically what FMM does. Now, imagine that we can establish a second level of hubs which can further reduce the number of telephone lines. If the number of telephone lines is very large, we can reduce the number of telephone lines to *O(NlogN)* by establishing multiple levels of hubs. Similarly, MLFMA reduces the operation counts and memory requirement of the FMM to *O(NlogN).*

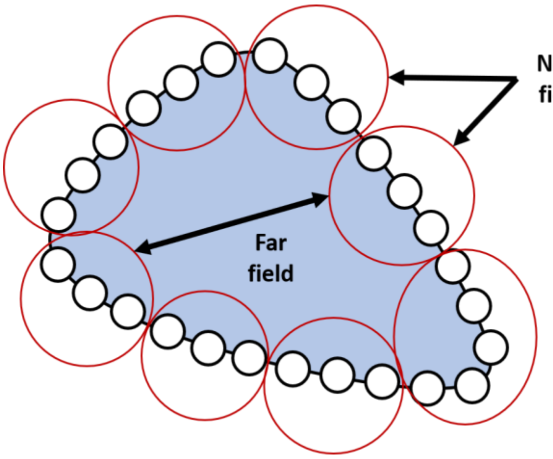


Figure 5. Basis functions are divided into groups for fast far-field computation.

Finally, as a real application example, in [22], there is a snapshot of the surface current on a card induced by a Hertzian dipole at 1.0 GHz and a snapshot of the surface current on an airplane induced by an incident plane wave at 2.0 GHz. The discretization of the airplane surface results in nearly 1 million unknowns. Storing it in its corresponding full MoM matrix would take around 8TB of storage memory. Using MLFMA, the memory storage requirement is reduced to 2.5 GB. In [18], the same airplane is simulated with approximately 10 million unknowns at 8 GHz. Also, in [12], we can find another example of surface current on an aircraft from a boundary element with approximately 2 million unknowns.

In addition to all the topics discussed above, there are parallelization approaches of the MLFMA on distributed memory computers. The most common parallelization approach is to partition the data as tree structures over computational nodes. To make this possible, we apply Message Passing Interface (MPI) based parallelization, such as [16][17].

This paper has briefly discussed the complexity performance for MoM and the accelerated versions with FMM and MLFMA as integral method solvers for EM problems in the frequency domain. A comparison of the complexity performance for these three algorithms using iterative solvers is summarized in Table 1.

TABLE I. MoM Based Fast Algorithm Complexities

Method	**Complexity**	
	Time	***Memory***
MoM	$O(N^2)$	$O(N^2)$
FMM	$O(N^{1.5})$	$O(N^{1.5})$
MLFMA	*O(NlogN)*	*O(NlogN)*

VI. Big Data Techniques in Electromagnetic Engineering Problems

The electromagnetic spectrum has shown four characteristics of Big Data, namely, Variety, Volume, Value, and Velocity [32]. One application of Big Data is reported in [32], where data mining is used to detect abnormal spectrum and abnormal positioning targets from massive EM data in real-time. Another application of Big Data in EM problems is Symbolic Regression (SR). This type of regression analysis is used to perform a search in an analytical expression that fits a large dataset [23] SR is classified as a Machine Learning technique and can be applied to derive a full-wave simulation-based analytical expression for the characteristic impedance Z_0 of microstrip lines using Big Data resulting from a 3D-EM simulation [23]. SR is considered a suitable algorithm for obtaining accurate analytical expressions where the interrelations within the data are highly complex in a very large dataset [23]. A different implementation of machine learning to manage the large size of data for design optimization in EM can be found in the literature, such as reinforcement learning for antenna configuration and design [33], deep learning for microwave filter and circuit design [34], EM inverse problems in oil and gas exploration, as well as microwave and optical imaging [30].

Big Data in EM is found in the design of tilted-beam antennas with aperiodic Partially Reflective Surfaces (PRS). To design antennas for beamforming and high gain wireless application, PRS are highly reflective metasurfaces considered well suitable for the design of antennas [25]. During the optimization process of the aperiodic PRS, a large data size is generated. An improved Hybrid Real-binary Bat Algorithm (HRBBA) is applied to optimize the aperiodic PRS [26]. Bat Algorithm (BA), inspired by the echolocation of microbats, efficiently and reliably process Big Data optimization problems [27]-[29]. In [30], a statistical approach is proposed based on the Markov Chain Monte Carlo (MCMC) for Large-Scale Georsteering inversion using directional electromagnetic logging measurements. Due to the high volume of data collection in the oil and gas industry, the proposed method in [30] addresses large-scale inverse problems.

Today, the convergence between Big Data Analytics and High-Performance Computing is considered a promising research area [35]. In CEM, training deep learning or running large-scale simulations can take a tremendous amount of time. For this reason, parallel and high-performance computing are essential to efficiently accelerate the convergence of an algorithm toward an accurate solution. An application of Big Data techniques in the EM scattering problem can be found in [31]. This work proposes a method to predict the number and location of scattering grating lobes produced by an array antenna. The method used implements the idea of decomposing the RCS of the array antenna into a multiplication of the array RCS factor and the element RCS factor.

Fast Algorithms such as MLFMA are developed to accelerate the algorithm execution. At the same time, they can reduce the complexity of the algorithm in terms of memory and time; especially, it considerably alleviates the memory requirement to store the matrix system that can store millions or billions of unknown's values. Parallel computing is implemented to reduce the computational time of the algorithm; in addition, it extends the usability of multiple threats for the mathematical operations in solving the problem. Applications of high-performance computing in EM engineering applications can be found in areas such as EM radiation, propagation and scattering, antenna analysis, RCS, analysis of Electromagnetic Compatibility (ECM) and Electromagnetic Interference (EMI), circuits modeling, microwave, analysis, nano-electronic devices among others [36]-[40].

VII. Conclusion

Big Data has become one of the most important fields for complex research related to engineering applications. We have seen that the term Big Data does not only mean a very large amount of data; it is also a concept considering several important factors, such as how we interpret data, how valuable it is, and even how variable the data could be, like the well-known 5 V's of Big Data. Besides, Big Data helps with the management of structured, unstructured, or misstructured data. Efficient algorithms exploit Big Data's potential by reducing its computational complexity in modern computers. High-performance computing supports efficient large-scale data-intensive processing to enable complex applications in different scientific and engineering fields.

In this survey, we have described what computational electromagnetics is and how highly multidisciplinary of a field it is. We have also described the numerical procedures of MoM and its application in EM scattering problems. MoM has been the base for fast algorithm implementations, such as FMM and MLFMA. It is important to state that MLFMA has been one of the most important advances in CEM in the last two decades. The development of numerical methods can be applied effectively across spatial, temporal, and frequency scales with the modeling and simulation of physical phenomena, such as circuits, heat transfer, and charge transport. This opens a new opportunity for computational electromagnetics research.

Acknowledgment

This work is supported by the US Science & Technology Center grant (CCF-0939370).

References

[1] R. Narasimhan and T. Bhuvaneshwari, “Big Data - A Brief Study,” International Journal of Scientific & Engineering Research, vol. 5, no. 9, pp. 350-353, 2014.

[2] T. Sarlos, “Improved Approximation Algorithms for Large Matrices via Random Projections,” In Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, pp. 143-152, 2016.

[3] C. Sorzano, J. Vargas, and A. Pascual Montano, “A survey of dimensionality reduction techniques,” arXiv preprint arXiv:1403.2877, 2014.

[4] D. Jin, Y. Yang, T. Ge, and D. Wu, “A Fast Sparse Recovery Algorithm for Compressed Sensing Using Approximate l_0 Norm and Modified Newton Method,” Materials, vol. 12, no. 8, pp. 1227, 2019.

[5] A. Gilbert, P. Indyk, M. Iwen, and L. Schmidt, “Recent Developments in the Sparse Fourier Transform: A compressed Fourier transform for big data,” IEEE Signal Processing Magazine, , vol. 31, no. 5, pp. 91-100, 2014.

[6] H. Shah, G. Badsha, A. Abbasi, and S. Salehian, “The 10 Vs, Issues and Challenges of Big Data,” Proceedings of 2018 the International Conference on Big Data and Education, pp. 52-56, 2018.

[7] N. Khan et al., “The 51 V's Of Big Data: Survey, Technologies, Characteristics, Opportunities, Issues, and Challenges,” Proceedings of the International Conference on Omni-Layer Intelligent Systems, pp. 19-24, 2019.

[8] R. Patgiri and A. Ahmed, “Big Data: The V's of the Game Changer Paradigm,” 2016 IEEE 18th International Conference on High-Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 17-24, 2016.

[9] A. Shanin, “Big Data Five V's Characteristics,” International Journal of Advances in Electronics and Computer Sciences, vol. 2, no. 1, 2015.

[10] W. C. Chew, “Introduction to Electromagnetic Analysis and Computational Electromagnetics,” Fast and Efficient Algorithms in Computational Electromagnetics, pp. 1-17, 2001.

[11] J. M. Jin, “Concluding Remarks on Computational Electromagnetics,” Theory, and Computation of Electromagnetic Fields, Ed. 2, pp. 651-671, 2015.

[12] E. Larsson et al., “Parallelization of Hierarchical Matrix Algorithms for Electromagnetic Scattering Problems, High-Performance Modelling and Simulation for Big Data Applications,” In High-Performance Modelling and Simulation for Big Data Applications, pp. 36-68, 2019.

[13] J. M. Jin, “The Method of Moments,” Theory and Computation of Electromagnetic Fields, Ed. 2, pp. 651-671, 2015.

[14] K. K. Mei and J. Van Bladel, “Scattering by perfectly conducting rectangular cylinders,” IEEE Trans. Antennas Propag., vol. 11, no. 2, pp. 185–192, 1963.

[15] J. M. Jin, “Fast Algorithms and Hybrid Techniques,” Theory and Computation of Electromagnetic Fields, Ed. 2, pp. 651-671, 2015.

[16] K. C. Donepudi, J. M. Jin, S. Velamparambil, J. M. Song, and W. C. Chew, “A higher-order parallelized multilevel fast multipole algorithm for 3D scattering,” IEEE Trans. Antennas Propag., vol. 49, pp. 1069–1078, 2001.

[17] J. Kurzak and B. Pettitt, “Massively parallel implementation of a fast multipole method for distributed memory machines,” Journal of Parallel and Distributed Computing, vol. 65, no. 7, pp. 870–881, 2005.

[18] W. C. Chew, J. M. Jin, E. Michielssen, and J. M. Song, “Fast and Efficient Algorithms in Computational Electromagnetics,” Norwood, MA: Artech House, 2001.

[19] D. B. Davidson, “A one-dimensional introduction to the method of moments: thin-wire modeling,” Computational Electromagnetics for RF and Microwave Engineering, pp. 118-145, 2005.

[20] D. B. Davidson, “The method of moments for surface modeling,” Computational Electromagnetics for RF and Microwave Engineering, pp 184-230, 2005.

[21] T. J. Cui and W. C. Chew, “Fast Forward and Inverse Methods for Buried Objects,” Fast and Efficient Algorithms in Computational Electromagnetics, pp. 347-424, 2001.

[22] J. M. Song, C. C. Lu, W. C. Chew, and S. W. Lee, “Fast Illinois solver code (FISC),” IEEE Antennas Propag. Mag., vol. 40, no. 3, pp. 27–34, 1998.

[23] P. Mahouti et al., “Symbolic regression for derivation of an accurate analytical formulation using "Big Data": An application example,” Applied Computational Electromagnetics Society Journal, vol. 32, no. 5, pp. 372-380, 2017.

[24] M. F. Korns, “Extremely accurate symbolic regression for large feature problems,” Genetic Programming Theory and Practice XII, Genetic and Evolutionary Computation, pp. 109-131, 2015.

[25] A. Hoorfar and C. Israel, “Nature-Inspired Optimization of Aperiodic Metasurfaces for Antenna Beam-shaping,” 2019 IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting, pp. 1035-1036, 2019.

[26] Y. F. Cheng et al., “Design of Tilted-Beam Fabry-Perot Antenna with Aperiodic Partially Reflective Surface,” Applied Computational Electromagnetics Society Journal, vol 32, no. 5, 2017.

[27] X. S. Yang, “A New Metaheuristic Bat-Inspired Algorithm,” In: J. R. Gonzalez et al. (eds), Nature Inspired Cooperative Strategies for Optimization (NISCO 2010), Springer, Berlin, pp. 65-74, 2010.

[28] A. H. Gandomi, X. S. Yang, A. H. Alavi, and S. Talatahari, “Bat algorithm for constrained optimization tasks,” Neural Computing and Applications, pp. 1239-1255, 2012.

[29] X. S. Yang, “Bat algorithm for multi-objective optimization,” International Journal of Bio-Inspired Computation, pp. 267-274, 2013.

[30] Q. Shen, X. Wu, J. Chen, and Z. Han, “Distributed Markov Chain Monte Carlo Method on Big-Data Platform for Large-Scale Geosteering Inversion Using Directional Electromagnetic Well Logging Measurements,” Applied Computational Electromagnetics Society Journal, vol. 32, no. 5, 2017.

[31] S. Zhang, X. Wang, L. Guo, and J. Ma, “Multiplication Theory for Prediction of the Scattering Grating-lobe of Array Antenna,” Applied Computational Electromagnetics Society Journal, vol. 32, no. 5, 2017.

[32] L. N. Liu, R. Shi, B. Hee, and M. Chen, “Detection on Abnormal Usage of Spectrum by Electromagnetic Data Mining,” In 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), pp. 182-187, 2019.

[33] X. Chai et al., “Reinforcement Learning Based Antenna Selection in User-Centric Massive MIMO,” In 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), pp. 1-6, 2020.

[34] J. Jing et al., “Recent Advances in Deep Neural Network Technique for High-Dimensional Microwave Modeling,” In 2020 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization (NEMO), pp. 1-3, 2020.

[35] S. Caíno-Lores, J. Carretero, B. Nicolae, O. Yildiz, and T. Peterka, “Toward High-Performance Computing and Big Data Analytics Convergence: The Case of Spark-DIY,” IEEE Access, vol. 7, pp. 156929-156955, 2019, doi: 10.1109/ACCESS.2019.2949836.

[36] M. J. Gander and S. Vandewalle, “Analysis of the parareal time-parallel time-integration method,” SIAM Journal on Scientific Computing, vol. 29, no. 2, pp. 556-578, 2007.

[37] C. Cong, X. C. Cai, and K. Gustafson, “Implicit space-time domain decomposition methods for stochastic parabolic partial differential equations,” SIAM Journal on Scientific Computing, vol. 36, no. 1, pp. C1–C24, 2014.

[38] M. Emmett, and M. Minion, “Toward an efficient parallel in time method for partial differential equations,” Communications in Applied Mathematics and Computational Science vol. 7, no. 1, pp. 105-132, 2012.

[39] A. J. Christlieb, C. B. Macdonald, and B. W. Ong, “Parallel high-order integrators,” SIAM Journal on Scientific Computing, vol. 32, no. 2, pp. 818-835, 2010.

[40] M. J. Gander, “50 years of time parallel time integration,” In Multiple shooting and time domain decomposition methods, Springer, Cham, pp. 69-113, 2015.

Information Integration using the Typed Graph Model

Fritz Laux
Fakultät Informatik
Reutlingen University
D-72762 Reutlingen, Germany
email: fritz.laux@fh-reutlingen.de

Malcolm Crowe
School of Computing
University of the West of Scotland
Paisley PA1 2BE, UK
email: malcolm.crowe@uws.ac.uk

***Abstract*—Schema and data integration have been a challenge for more than 40 years. While data warehouse technologies are quite a success story, there is still a lack of information integration methods, especially if the data sources are based on different data models or do not have a schema. Enterprise Information Integration has to deal with heterogeneous data sources and requires up-to-date high-quality information to provide a reliable basis for analysis and decision making. The paper proposes virtual integration using the Typed Graph Model to support schema mediation. The integration process first converts the structure of each source into a typed graph schema, which is then matched to the mediated schema. Mapping rules define transformations between the schemata to reconcile semantics. The mapping can be visually validated by experts. It provides indicators and rules to achieve a consistent schema mapping, which leads to high data integrity and quality.**

Keywords–data integration process; typed graph model; schema mapping; mapping rules; data quality.

I. Introduction

Information integration, integrating data from different sources, enables us to gain new knowledge and insights that help to make predictive analysis, coordinate complex processes and control systems. Depending on the application the mediated data can be materialized as in a Data Warehouse (DW) or each query can access the data sources to get an up-to-date result, which is called *virtual integration*. Many situations need up-to-date or near real-time information, which can only be achieved by virtual integration. This is the case not only in emergency situations like fighting epidemic, earthquake, and other disaster aid but also in industry production.

In Enterprise Information Integration (EII), the integration of heterogeneous data sources is usually achieved by a manually supervised process to ensure a high-quality mediated global schema. In this paper, we concentrate on *supervised* schema integration, i.e., the semantic data integration problem. Much research has been conducted to automatically match and map data sources [1][2], but the quality of the results are usually not sufficient for EII systems [3]–[6]. Thus, additional manual changes using context and other semantic information are necessary to build a high-quality global/mediated schema, i.e., an *engineered schema* [7].

The manual improvement and quality checking of the mediated schema should be supported by software that visualizes the semantics of the data integration and the impact and interplay of any changes.

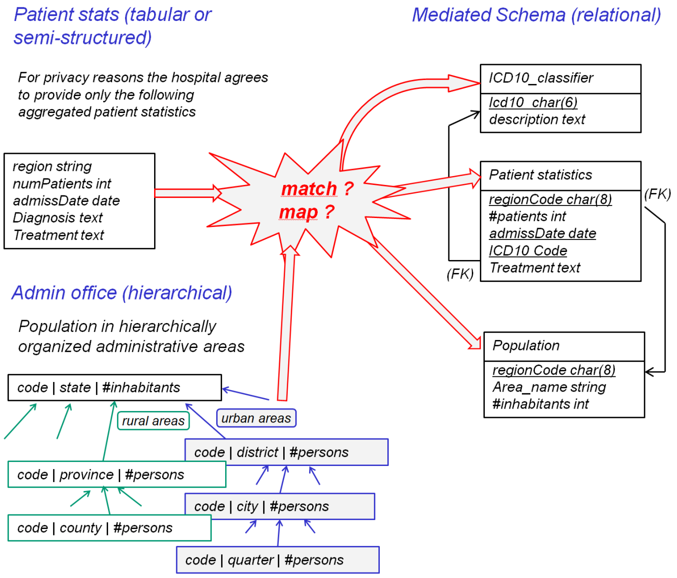

Patient stats (tabular or semi-structured)
Mediated Schema (relational)
For privacy reasons the hospital agrees to provide only the following aggregated patient statistics
region string
numPatients int
admissDate date
Diagnosis text
Treatment text
match ?
map ?
ICD10_classifier
Icd10 char(6)
description text
Patient statistics
regionCode char(8)
#patients int
admissDate date
ICD10 Code
Treatment text
(FK)
(FK)
Population
regionCode char(8)
Area_name string
#inhabitants int
Admin office (hierarchical)
Population in hierarchically organized administrative areas
code | state | #inhabitants
rural areas
urban areas
code | district | #persons
code | province | #persons
code | city | #persons
code | county | #persons
code | quarter | #persons

Figure 1. Data integration scenario (Running example)

Let us illustrate our approach with an example scenario presented in Figure 1. It is borrowed from Crowe et al. [8]. The World Health Organization reports on emergencies/epidemics, using data provided in many different formats by national or regional autonomous actors like national health authorities or hospitals. For simplicity, we only show one hospital providing aggregated *tabular data* on patients grouped by regions, admission date and diagnosis. For each group element the number of patients and their treatment are recorded. The statistics admin offices deliver demographic data in a *hierarchical structure* reflecting the administrative areas. We want to create an integrated/mediated schema that combines these sources using the International Classifier for Diseases (ICD). Given a *relational* Integration Schema, the question is “how do we find matches for the data items and transform them”.

We can identify five possible problems in this example.

From Figure 1, we need to identify the matching between *region string* of the hospital patient stats and the *regionCode* in the mediated schema despite the different coding (Problem 1). There might be different spelling or naming of *region* strings in different hospitals. These conflicts need to be resolved in the mediated schema (Problem 2). Preserving the semantics (Problem 3) requires the *ICD10_classifier* for the mediated schema to be obtained from the *diagnosis* text. The population data from administrative areas need to be mapped into relations preserving the hierarchy (Problem 4). And finally (Problem 5), we need consistency when multiple mapping paths exist like from *region* (hospital stats) to *regionCode* (Population), one via *code* (admin office) and the other either direct or via *regionCode* (Patient statistics).

In Figure 1, we used different graphical representations for the schemata to reflect the various data models and structures. But a consistent graphical representation would be better to solve the matching/mapping task. The Typed Graph Model (TGM) [9] represents a flexible model with enough expressive power to cover most data structures comprehensively and support several abstraction levels. This is why we have chosen the TGM for the schema mediation and data mapping.

Data integration has been intensively investigated from a theoretical point of view [10]–[12] leading to mainly two concepts, *Global As View* (GAV) and *Local As View* (LAV). As the terms indicate the global (resp. local) data are expressed as view of local (resp. global) data. The ACM digital library alone retrieves 163 matches for the key words "GAV or LAV". These papers are of great value to understand and specify the correspondences between two or more databases. But using description logic as formal specification gives little help to identify synonyms and homonyms or discover their semantics. Where sources of low quality overlap, it is important to remove redundancy while seeking to benefit from any extra information. The mapping of heterogeneous data sources to a specific target does not allow a fully automatic procedure if high data quality is required [5][3][13][14].

A. Contribution

Our idea is to use the TGM to support, formalize, and visualize data integration. The TGM helps to create a mediated target schema, in contrast to the less formalized Extract-Transform-Load process used in Data Warehousing. We propose a process, which divides the integration task into two phases: first, model all sources and the target using TGM, then, second, match and map the source models into the target model. The proposed process is well defined and combines for the first time *supervised* semi-automatic matching with mapping and merging of data. It provides rules for mapping and conflict resolution and defines criteria for quality control. We illustrate all integration steps using our running example.

B. Structure of the Paper

Section II briefly presents theoretical data integration work and practical experiences with emphasis on high-quality integration. In Section III the integration framework is presented. It consists of four activity steps:

1) Modeling the source data structures with TGM (Subsection III-C)
2) Defining the target schema using TGM (Subsection III-D)
3) Match/Map source with target data, resolve conflicts, and define necessary transformation (Subsection III-E)
4) Check and improve quality (Section IV).

Quality criteria and measures are developed in Section IV to help check and improve the integration quality. The paper ends with a summary of our findings and gives an outlook on ideas for future work.

II. Related Work

Since the middle of 1980s many papers on data integration have been published. In the following review, we restrict the focus to high-quality integration targeting EII.

A. Data Integration Overview

The papers of Sheth and Larson [15] give an introduction to federated database systems. Later textbooks of Öszu and Valduriez [16] and also Leser and Naumann [17] describe different integration methods and present general approaches for schema matching and mapping. Because our focus is on high-quality integration the experiences and considerations of Bernstein/Haas [3] and Halevy et al. [13] with real live projects are of high value to us. Both papers emphasize the need to integrate heterogeneous data sources including email, order documents, warranties, and other un- or semi-structured data.

Laura Haas states in her paper [6] that "despite the weighty body of literature, the information integration challenge is far from solved, especially in the enterprise context". As "Big I" challenges she identifies *entity resolution* and the lack of *theoretical and practical guidance* to make schema integration choices, noting the lack of a "broader framework" with quality control, which "considers the entire end-to-end integration process". This statement is confirmed by many papers that address the integration of Web data [18]–[20], XML data [20], or RDF data [21][22] in a declarative way but with little help on how to proceed in practice.

B. Data Integration using Graph Models

Some authors use a Graph Data Model (GDM) for data analysis and transformation. GRADOOP [23] and Pregel [24] are example prototypes for this approach. Most of the integration of graph databases is instance level based and uses graph transformations, see [25][2], but the question of how well the integration matches the original semantics and schemata of the sources is not addressed.

None of the papers really addresses how to match and map differently structured data elements by preserving the semantics of the sources. Practical guidance is available from de Sousa and del Val Cura [26] only for the mapping from the Extended Binary Entity Relationship (EB-ER) model to a Property Graph Model (PGM).

The only publication we are aware of that tries to improve data integration quality is Gelman [27]. In his paper he develops a theory that helps to produce accurate data integration output from multiple, overlapping, and inaccurate sources. He assumes that errors are not random and that "complementary" information helps to select the most accurate data. This approach could also help with the entity resolution problem.

Another problem with virtual data integration is that the global database may not be consistent with respect to integrity constraints. Bertossi and Bravo [28] recommend querying only the consistent part of the global database. They define a consistent answer to a query when the result is the answer to every "repair" of the global database, i.e., a maximal subset of the global database that satisfies the integrity constraints. The solution to the problem is quite general and conceptually clear, however an implementation is still missing. The authors demand that these issues must be “addressed in order to use those solutions in real database applications”.

III. The Integration Framework

The data integration framework uses the TGM as intermediate data model. This is why the TGM is shortly presented here. A detailed and formal introduction can be found in [9]. The data integration process consists of transforming the sources and target into Typed Graph Schemata (TGS) and then make the matching and mapping of schema elements.

A. The Typed Graph Model

The TGM combines schema support, complex data structures and abstraction using sub-graphs. Because of its rich semantics, we use it to capture the source and target semantics as accurately as possible. The source and target schemata act as a means of quality control.

The TGM informally constitutes a directed property hypergraph that conforms to a schema. Formally, the TGM is defined as quadruple $TGM = (N, E, TGS, \phi)$ 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 defined as tuple $TGS = (N_S, E_S, \rho, T, \tau, C)$, where N_S, E_S are vertices and edges of the graph schema, ρ defines the hyper-edges with its cardinalities τ. T is a set of data types used for vertices and edges, and finally C is a set of integrity constraints, which the graph database must obey.
- ϕ is a homomorphism that maps each node n and edge e of TGM to the corresponding type element of TGS,

As graphical representation for the TGS, we adopt the Unified Modeling Language (UML) to visualize the data model, which makes it useful for visualization tools to support the human controlled information matching and mapping. The visualization allows the modeling expert to validate the matching and mapping. The possible abstraction via subgraphs facilitates the overview and management of complex and large models.

A data model can act as a kind of **supermodel** if it is general enough to capture all popular models. Hull [29] and Atzeni et al. [30][31] describe such a supermodel that subsumes all popular data models including the Relational Model (RM), ERM, XML, Object Oriented Model (OOM), Object Relational (OR), and XSD. It consists of the following meta-constructs: *lexical, abstract, aggregation, generalization,* and *function*. The TGM can represent these meta-constructs in an information preserving way [32] by matching one-to-one lexical elements with properties, abstract constructs with nodes, aggregation and generalization constructs with the corresponding edge types, and functions with directed edges having multiplicity 1. This implies that the TGM is able to map the above-mentioned data models without loss of any information.

B. The Data Integration Process

The benefit of information integration is maximized when source data are integrated 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 data integration process consists of five tasks:

1) model sources as TGS $S_i \quad (i = 1, 2, ..., n)$
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

The full process is depicted in Figure 2 as a workflow modeled in BPMN. The 4^{th} task of the workflow is crucial for EII and other data integration projects, which demand highly accurate information quality. It might turn out that the resulting quality is not sufficient. As consequence the process might have to be iterated with different mappings in order to improve the quality. The last task is only necessary if the target schema is not a graph schema.

The first and second tasks consist of two alternatives depending on the pre-existence of source schemata, resp. target schema. If a schema already exists it is only necessary to transform it into a TGS. If a source has no schema, it is then necessary to collect structure and type information from a data expert or from additional information. This information is necessary before an appropriate TGS can be created. At least a minimal schema is required for every data source. This extra effort has the advantage to ensure a better data quality.

C. Model the Data Sources as TGS (Task 1)

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.

If the source is a database or other rigid structured data, the modeling of a TGS is rather simple and there are publications [33][34] that propose automatic schema conversion. The paper of Laux [9] gives some examples how to transform relational, object oriented, and XML-schemata into a TGS. If a schema already exists for a data source it may seem to be extra work to model it again as TGS. The benefit comes later when we look for matches because the schemata are all based on the same model. In addition, the quality of the matching can be checked formally with graph analysis.

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

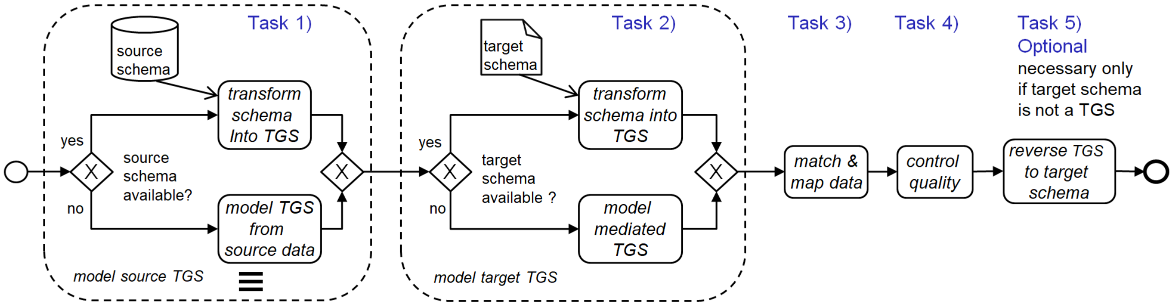
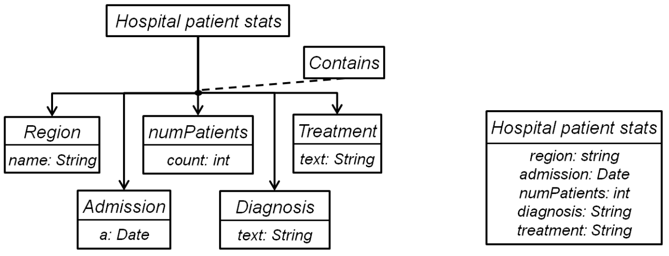


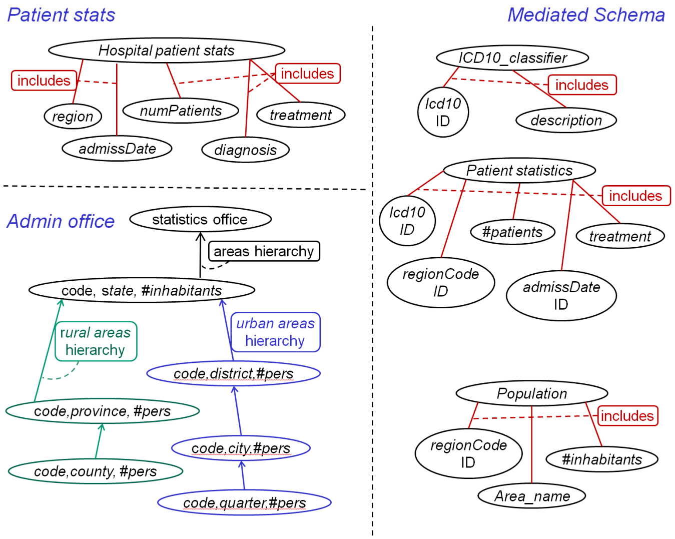
Figure 2. Workflow of information integration using the TGM



a) Detailed schema

b) Compact schema

Figure 3. TGS in UML notation for the patient statistics



Data sources as TGS

Target as TGS

Figure 4. Running example as TGS (data types not shown for simplicity)

is advisable to validate the matching with instance data or an information expert.

As example for semi-structured source data, we take the hospital patient statistics from our running example. The data for the statistics are entered in a form. We present two possible TGS in UML notation in Figure 3. Part a) shows a detailed schema where each data element is modeled as a node with its corresponding property and (simple) data type. In part b) the hospital patient stats are modeled as a complex data type. This little example demonstrates already the flexibility of the model in terms of detail and abstraction.

D. Model Mediated TGS (Task 2) and Target Schema (Task 5)

In general, many integration schemata are possible. In most cases the mediated schema tends to cover the union of the source schemata. Petermann et al. [35] present an automatic graph instance integration with the help of a Unified Metadata Graph (UMG). This method can be applied to generate a mediated graph TGS if the UMG is replaced by the data source schema graphs. Another approach for unstructured data is proposed by Buneman et al. [36] by union of the source graphs.

If a target schema T is given, but not already as TGS, it needs to be transformed to a TGS. In most cases a 1–1 transformation is possible. This can be done automatically using the same techniques as mentioned in III-C. This is the only case where Task 5) has to be executed. The TGS has to be reversed in this case to the target schema T again by applying the inverse transformation.

Let us return to our running example and model the two data sources and the mediated target as TGS. The result is shown in Figure 4. The nodes of the schemata are visualized with property names inside the ovals. The data types of the properties are suppressed, but the edge types (labels) are color coded and connected by dashed lines with the corresponding edges. In the next task, we have to decide on the matching and mapping of vertices.

E. Matching & Mapping Source TGS to Target TGS (Task 3)

Task 3 addresses problem 1 from the Introduction I. The matching step only identifies nodes and edges that are related, not necessarily one-to-one. In our example hospital statistics, apart from national language differences, the English names: cases, positive cases, reported cases, hospitalized, etc. could mean all the same or could mean different things. The TGM can help to identify, visualize, and match nodes and edges correctly using type and edge information (edge type, structure analysis, description). The matching can be supported by linguistic methods (name similarity, synonym and homonym list, thesaurus or ontology) and value analysis, e.g., using duplicates [37]. There are some publications that automate this process, see [1][2]. The use of TGM is flexible enough to support various data structures and visualizes the integration process, which helps to identify and resolve mapping problems manually.

In the next step, we define mappings between the source and the target nodes. The mapping is mainly a manual task and the integration schema designer has the responsibility to choose the best quality (freshness, reliability, precision) data, resolve conflicts if redundant data are available and to correct apparently incorrect names. These include merge operations with conflict resolution (Problem 2), e.g., entity deduplication, overlapping conflicting data, identification of global data, and (data type) transformation. It is important to preserve the semantics as far as possible (Problem 3). This requires the knowledge of meta-data (data type, structure analysis, description), which is supported by the TGM. The mapping may include data grouping, e.g., grouping patients according to cost factors (ABC analysis). The workflow sequence and the mapping function suggest following the more natural Global-As-View (GAV) approach when implementing the mappings. Even if the integration of new sources more complicated compared to the Local-As-View (LAV) approach, it reflects the reality of overlapping data with different coding or semantics, which has to be resolved in both approaches.

To complete our running example, we present in Figure 5 the result of the mapping of our running example. As in Figure 4 on the left side are two TGS representing the data sources and on the right side is the *Mediated Schema* which can be converted back to the relational target schema given in Figure 1. This corresponds to the final step 5 of the workflow. For illustration purpose the Foreign Keys (FK) are indicated by dashed blue lines. The two data source graphs are semantically connected (*related with*) via the region information. The matching and mappings (red arrows) between source schemata and target TGS exemplify only some of necessary matching and mappings. Between *Hospital patient stats* (source Patient stats) and *regionCode* (mediated schema) exist two commutative mappings: (1) via *Patient statistics* and (2) via *region*. Both include an isomorphism (iso) and a projection (π).

F. Example Patterns

In order to illustrate the modeling power and flexibility of the TGM, we present a series of typical mappings that arise during schema integration and mediation. Such mapping patterns reoccur often and have a standardized solution.

1) Merge Pattern: The Merge pattern solves problem 2) of how to merge two or more data nodes of similar semantics. The different data sources can provide additional and overlapping data. Multiple sources may produce conflicting values or duplicates, differ in scale and coding, and have different resolution. Duplicates must be removed. Rules need to be established for conflicting values, e.g., prioritize the most reliable value or calculate a mean value if all sources are of similar quality. In case of different coding use translation tables. For scale and unit mismatches define value transformations. The merge pattern is useful to reconcile data from different sources and to improve data quality if the data is redundant.

Figure 6 shows a typical merge situation of two sources (Hospital and Clinic) with different region coding. The mapping M_{12} between Hregion and Cregion is required and allows to merge the overlapping data. In case of a value conflict the Cregion gets precedence.

2) Homomorphism Pattern: A Homomorphism is a structure preserving mapping that helps to transform the source schemata into a target schema and thereby solves Problem 4) from the Introduction. It preserves edges but allows multiple nodes to map to the same target node. This can be used for data aggregation. If the mapping is injective (one-to-one), we have an Isomorphism. It transforms the source schemata into an equivalent target schema.

In Figure 7 the homomorphism f is a mapping that transforms all nodes and edges from schema S onto nodes and edges of target G. In this example the patient nodes are mapped to the numPatients node of the target by incrementing the numPatients value. The edges between patient$x$ ($x = 1, 2$) and Hospital are mapped (green arrows) to the same edge (PatientStats–NumPatients).

3) Commutative Mapping Pattern: When defining the mappings between two schemata special care has to be taken if a target node can be reached via different paths. This happens for example when in the source schema two data items are related and both items are mapped to the same target node. In this case, we have to use the Merge pattern to resolve the conflict. But, if we preserve edges like in our example in Figure 8, we should have mappings that commute. A function chain is called commutative if and only if the order of the functions does not matter, i.e., $f_2 \circ f_1 = f_1 \circ f_2$. If special mappings are used like projection g and isomorphism iso then chances are good that the mapping chain is commutative. Communicative mappings are an essential criterion for a consistent mapping, and it helps to solve problem 5) from the Introduction. Commutativity is important because it guarantees that we have the choice between alternative mapping paths and still end up with the same target data. In Figure 8 it is irrelevant if the projection g to region is done first or the isomorphic mapping iso_1 to patientCode. We always end up with the same regionCode.

IV. Quality Criteria (Task 4)

A bipartite graph is a graph where the nodes are separated in two disjoint subsets with no edges within the subsets. This is the case with graph matching if we remove (in our mind) the connections inside the local schemata. If we consider such a bipartite graph, we can define the following quality criteria: A matching between the source and target graphs is called *maximum matching* if the number of matched vertices is maximized. If all nodes are matched, we call this a *perfect matching*. This guarantees that all nodes (data elements) from the sources are matched with target nodes and this gives us a criterion of how well the matching covers the integration task.

The theorem of Hall [38] states that there exists a perfect matching if all possible subsets of the source nodes have at least as many links to target nodes as the cardinality of this target subset. More formally, let $V = (S \cup G, E)$ be a bipartite graph of two disjoint sets S, G, then there exists a perfect matching if $\forall R \subseteq S$ the inequality $d(R) \geq |R|$ holds where $d(R) := |\{g \in G \mid r \in R \wedge (r, g) \in E\}|$ is the number of nodes in G linked to R. The perfect matching is a general criterion for the coverage or completeness of a data integration. If no

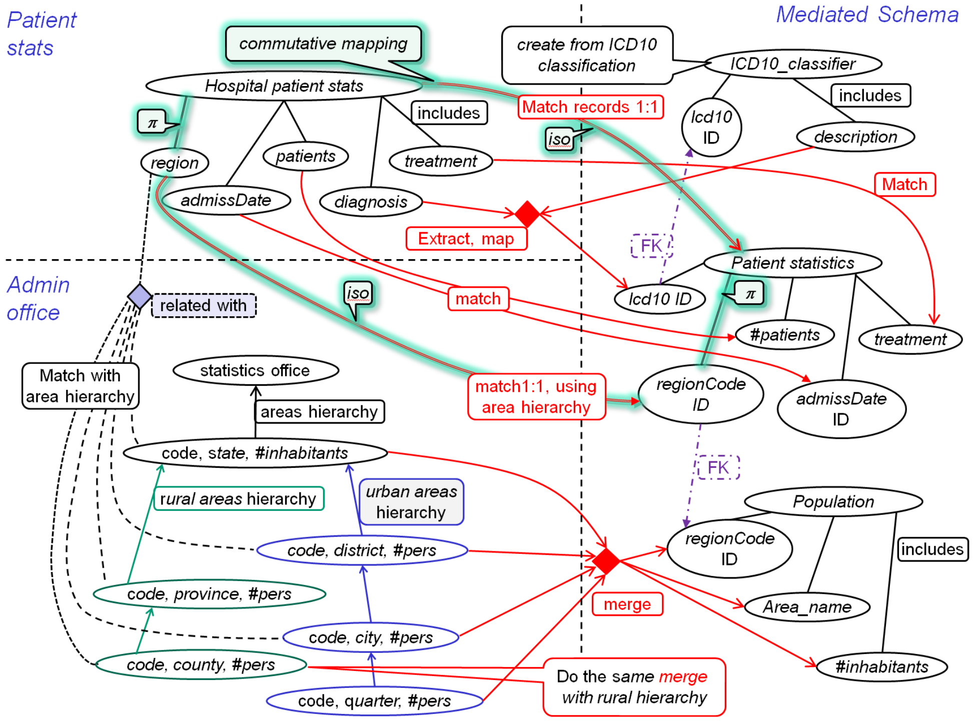


Figure 5. Matching and mapping result of the complete running example. To not overload the figure, only some of the matches and mappings (red arrows) are shown. The commutative mapping between *Hospital patient stats* and *regionCode* is highlighted (green glow).

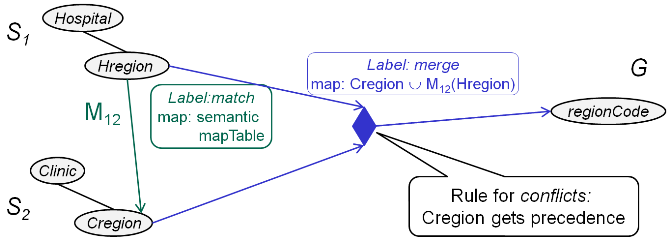


Figure 6. Merge example of two Patient stats sources (Hospital and Clinic) with map table and conflict resolution

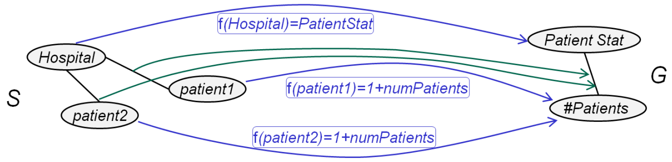


Figure 7. Example instance graph Homomorphism that sums patients

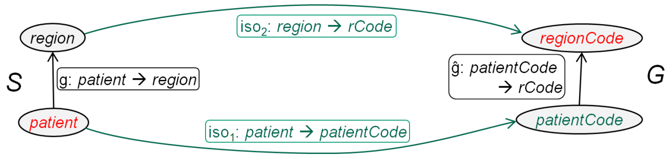


Figure 8. Commutative mapping from patient to regionCode

perfect match exists a merge conflict can arise and conflict resolution is necessary.

The theorem of Hall is only a formal quality criterion. In order to improve the semantic mapping quality, we may distinguish three cases on the instance level:

(1–1) Data are mapped 1–1 to compatible data types or via an enumeration list. This may include disjoint merge operations.

(n–1) Data are mapped to an aggregated value or a merge with redundant data.

(1–n) Data are distributed to multiple data elements. This can occur for address data that is split into separate fields or split up values that include tax.

A semantic mapping quality measure can be established by assigning 3 points to every 1–1 mapping, 2 points to every n–1 mapping. The n–1 mapping loses information compared to the 1–1 mapping; therefore, it receives a lower score. The 1–n mappings receive 1 to 3 points depending on the reliability of the split operation. Integration mappings with the highest sum represent the best match. There exist many other schema quality measures for completeness, correctness, minimality, etc. A nice overview on these measures is given in [39]. These measures are of great value to quantify the quality of a schema but give no direct help how to decide between mapping options.

Let us take Figure 8 as example and calculate the quality measure for the pictured mappings. The 1–1 mappings id_1, id_2, iso_1, and iso_2 have 3 points each and the projections g and $\hat{g}$ get 2 points each because of the n–1 mapping. The mapping chains $g \circ iso_2$ and $iso_1 \circ \hat{g}$ are commutative and yield the same score 5 (= 3 + 2), which confirms the equivalence of both mapping paths.

The matching and mapping in Figure 5 also show a pair of commutative mappings from *Hospital patient stats* to *regionCode* with equal quality scores (see arrows with green glow, *iso* = 3 points, projection $\pi = 2$ points). If the scores differ, it is recommended to only use the mapping with the higher score because it represents a higher mapping quality.

V. Conclusion and Future Work

This paper presented an information integration process using the TGM with emphasis on semantically high quality as required for enterprise or government applications. Due to the TGS with predefined and user-defined data types, the TGM improves the formal data quality compared to other data integration approaches. Integration patterns and quality criteria give guidelines for practical use of the matching and mapping task. The whole process was illustrated by a running example.

We conclude that high-quality integration with an engineered target schema is feasible. Some integration steps may be supported by automatic methods but still require human support with meta-data and context knowledge to achieve high-quality results. The development of software to support schema generation as well as the matching and mapping remains as future work. In particular, the current expert-based mappings in Task 3 could be enhanced with semi-automated suggestions to the user with possible mapping options. One approach for such a program could use some elements from the academic prototypes GRADOOP [23], IncMap [40] or Pregel [24] to reduce the development effort.

References

[1] E. Rahm and P. Bernstein, “A survey of approaches to automatic schema matching”, The VLDB Journal 10, pp. 334-350, 2001, DOI: 10.1007/s007780100057

[2] S. Melnik, H. Garcia-Molina, and E. Rahm, “Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching”, ICDE, pp. 117–128, 2002, DOI: 10.1109/ICDE.2002.994702

[3] P. Bernstein and L. Haas, “Information integration in the enterprise”, Communications of the ACM 51(9), pp. 72–79, 2008

[4] B. Golshan, A. Halevy, G. Mihaila, and W.-Ch. Tan, “Data Integration: After the Teenage Years”, ACM PODS 2017, pp. 101–106

[5] D. Strong and O. Volkoff, “Data Quality Issues in Integrated Enterprise Systems”, Proceedings of the MIT ICIQ Conference, no page numbers, 2005, [online] URL: http://mitiq.mit.edu/ICIQ/Documents/IQ Conference 2005/Papers/DQIssuesinIntegratedEnterpriseSystems.pdf [retrieved: 2021-04-24]

[6] L. Haas, “Beauty and the Beast: The Theory and Practice of Information Integration”, In T. Schwentick and D. Suciu (Eds.): ICDT 2007, pp. 28–43, Springer-Verlag, Berlin Heidelberg, 2007, DOI: 10.1007/11965893_3

[7] P. Bernstein and S. Melnik, “Model Management 2.0: Manipulating Richer Mappings”, Proceedings of the ACM SIGMOD International Conference on Management, pp. 1–12, 2007, DOI: 10.1145/1247480.1247482

[8] M. Crowe, C. Begg, F. Laux, and M. Laiho, “Data Validation for Big Live Data”, DBKDA 2017, pp. 30–36, ISBN: 978-1-61208-558-6.

[9] F. Laux, “The Typed Graph Model”, DBKDA 2020, pp. 13–19, ISBN: 978-1-61208-790-0.

[10] A. Doan, A. Halevy, and Z. Ives, Principles of Data Integration, Morgan Kaufmann, Elsevier, 2012, ISBN: 978-0-12-416044-6.

[11] M. Lenzerini, “Data Integration: A Theoretical Perspective”, Proceedings of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS) , pp. 233-246, 2002

[12] R. Fagin and Ph. Kolaitis, “Local transformations and conjunctive-query equivalence”, PODS 2012, pp. 179–190, DOI: 10.1145/2213556.2213583

[13] A. Halevy, N. Ashish, D. Bitton, M. Carey, D. Draper, J. Pollock, A. Rosenthal, and V. Sikka, “Enterprise information integration: successes, challenges and controversies” SIGMOD Conference, pp. 778–787, 2005

[14] D. Maluf and P. Tran, “Netmark: A schema-less extension for relational databases for managing semi-structured data dynamically”, ISMIS 2003, pp. 231–241, 2003

[15] A. Sheth and J. Larson, “Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases”, ACM Computing Survey 22(3), pp. 183–236, 1990

[16] M. T. Özsu and P. Valduriez, *Principles of Distributed Database Systems*, Fourth Edition, Springer Nature Switzerland AG, 2020, ISBN: 978-3-030-26252-5

[17] U. Leser and F. Nauman, *Informationsintegration - Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen* (Eng. Information integration - Architectures and methods for integration of distributed and heterogeneous data sources), dpunkt.verlag, 2006

[18] M. Friedman, A. Levy, and T. Millstein, “Navigational Plans for Data Integration”, AAAI/IAAI, pp. 67–73, AAAI Press, 1999

[19] D. Roman et al., “The Linked Data AppStore - A Software-as-a-Service Platform Prototype for Data Integration on the Web”, Mining Intelligence and Knowledge Exploration (MIKE); pp. 382–396; 2014

[20] L. Popa, Y. Velegrakis, R. Miller, M. Hernández, and R. Fagin, “Translating Web Data”, pp. 598–609, VLDB, 2002

[21] A. Langegger, “Virtual Data Integration on the Web: Novel Methods for Accessing Heterogeneous and Distributed Data with Rich Semantics”, iiWAS 2008, pp. 559–562, DOI: 10.1145/1497308.1497410,

[22] A. Adamou and M. d’Aquin, “Relaxing Global-As-View in mediated data integration from Linked Data”, In: *Proceedings of The International Workshop on Semantic Big Data* (SBD 2020), Article No.: 4, pp. 1-6, DOI: 10.1145/3391274.3393635

[23] M. Junghanns, A. Petermann, K. Gómez, and E. Rahm, “GRADOOP: Scalable Graph Data Management and Analytics with Hadoop”, Computing Research Repository (CoRR), vol. abs/1506.00548, no page numbers, 2015, [online] URL: https://arxiv.org/abs/1506.00548 [retrieved: 2021-04-24]

[24] G. Malewicz et al., “Pregel: a system for large-scale graph processing”, SIGMOD Conference 2010, pp. 135–146

[25] M. Kricke, E. Peukert, and E. Rahm, “Graph Data Transformations in Gradoop”. In T. Grust et al. (eds.) BTW 2019, pp.193–202, DOI: 10.18420/btw2019-12.

[26] V. de Sousa and L. del Val Cura, “Logical Design of Graph Databases from an Entity-Relationship Conceptual Model” iiWAS 2018, pp. 183–189

[27] I. A. Gelman, “A Theory of Complementarity for Extracting Accurate Data from Inaccurate Sources through Integration”, no page numbers, ICIQ 2005

[28] L. Bertossi and L. Bravo, “Consistent Query Answers in Virtual Data Integration Systems”, Inconsistency Tolerance 2005, pp. 42–83

[29] R. Hull, “Managing Semantic Heterogeneity in Databases: A Theoretical Perspective”, PODS ’97, pp. 51–61, ACM Press, 1997, DOI: 10.1145/263661.263668

[30] P. Atzeni and R. Torlone, “Management of Multiple Models in an Extensible Database Design Tool”, EDBT 1996, pp. 79–95

[31] P. Atzeni, P. Cappellari, R. Torlone, Ph. Bernstein, and G. Gianforme, “Model-independent schema translation”, VLDB Journal 17(6), pp. 1347–1370, 2008

[32] F. Laux, “The Typed Graph Model – a Meta-Language for Model Management and Data Integration”, International Journal On Advances in Software, vol. 14 no 1&2, 2021, ISSN: 1942-2628, unpublished

[33] R. De Virgilio, A. Maccioni, and R. Torlone, “Converting Relational to Graph Databases”, First International Workshop on Graph Data Management Experiences and Systems (GRADES), pp. 1–6, CWI/ACM, 2013, DOI: 10.1145/2484425.2484426

[34] R. Stoica, G. Fletcher, and J. Sequeda, “On Directly Mapping Relational Databases to Property Graphs”, Proceedings of the 13th AMW 2019, no page numbers, [online] URL: http://ceur-ws.org/Vol-2369/short06.pdf [retrieved: 2021-04-30]

[35] A. Petermann, M. Junghanns, R. Mller, and E. Rahm, “Graph-based Data Integration and Business Intelligence with BIIIG”, Proc. of the VLDB Endow, Vol. 7 no 13, pp. 1577–1580, 2014, DOI: 10.14778/2733004.2733034

[36] P. Buneman, S. Davidson, M. Fernandez, and D. Suciu, “Adding Structure to Unstructured Data”, 6th International Conference on Database Theory - ICDT, pp. 336–350, 1997

[37] A. Bilke and F. Naumann, “Schema Matching using Duplicates”, Proceedings of the 21st International Conference on Data Engineering (ICDE), pp. 69–80, 2005

[38] P. Hall, “On representation of subsets”, Journal of the London Math. Society, Vol. s1-10, Issue 1, pp. 26–30, 1935

[39] L. Ehrlinger and W. Wöß, “Automated Schema Quality Measurement in Large-Scale Information Systems”, in: H. Hacid, Q. Sheng, T. Yoshida, A. Sarkheyli, and R. Zhou (eds) Data Quality and Trust in Big Data (QUAT 2018), Lecture Notes in Computer Science, vol 11235, pp. 16–31, Springer, 2018, [online] URL: https://doi.org/10.1007/978-3-030-19143-6_2 [retrieved: 2021-05-02]

[40] Ch. Pinkel et al., “IncMap: A Journey towards Ontology-based Data Integration”, in Mitschang et al. (eds.) BTW 2017, pp. 145-164, [online] URL: https://dl.gi.de/20.500.12116/625 paper10.pdf [retrieved: 2021-04-24].

Visualization of Multi-Level Data Quality Dimensions with QuaIIe

Sheny Illescas Martinez
Johannes Kepler University Linz
Linz, Austria
k1257276@students.jku.at

Lisa Ehrlinger
Johannes Kepler University Linz
Linz, Austria, and
Software Competence Center Hagenberg GmbH
Hagenberg, Austria
lisa.ehrlinger@jku.at

Wolfram Wöß
Johannes Kepler University Linz
Linz, Austria
wolfram.woess@jku.at

***Abstract*—Data quality assessment is a challenging but necessary task to ensure that business decisions that are derived from data can be trusted. A number of data quality metrics have been developed to measure dimensions like accuracy, completeness, and timeliness. The tool QuaIIe (developed as part of our previous research) facilitates the calculation of different data quality metrics on both, schema- and data-level, and for heterogeneous information systems. However, to gain meaningful results from the automatically calculated metrics, it is key that humans *understand* the results of such metrics. This understanding is specifically important when contextual information needs to be considered, which is not encoded in the data. In this paper, we present a visualization approach to enable human-centered data quality assessment across multiple dimensions and arbitrary complex data sources. The approach has been implemented as graphical user interface in QuaIIe.**

Keywords*—*Data visualization; Data quality, Data quality dimensions; Graphical user interface.

I. Introduction

In a world where data is consumed and created on a daily basis, it has become a challenging task to determine whether the consulted data is of acceptable quality, especially for enterprises. The absence of Data Quality (DQ) assessment can have a severe financial impact on organizations and enterprises. A study made by the Data Warehousing Institute has shown that DQ problems cost American enterprises over 600 billion dollars on a yearly basis [1]. Determining the quality of data has therefore become a pivotal activity in the business field.

Data quality can be calculated with the help of DQ metrics, which are functions to assess different DQ dimensions in a quantitative manner [2]. Every DQ dimension evaluates a specific feature on the data or schema [3]. Assessing the DQ of an entire information system is not trivial, since tasks like specifying the dimensions of interest, determining the methods for assessing the quality of the selected dimensions, analyzing the results, and cleaning problematic data need to be carried out. Yet, the last two activities require one key aspect, which is to *understand* the DQ measurement results.

To effectively communicate and interpret DQ measurement results, visual representations are crucial. DQ visualization systems support the nonlinear analytical process of the comprehension of the qualitative state of the underlying information [4]. It is essential to understand and to know the quality of the data, in order to define goals and to determine the required data cleansing activities [5].

Enterprises and organizations often store their data in Integrated Information Systems (IISs). IISs typically manage and process data from different and heterogeneous information sources. In prior research, we developed the DQ tool QuaIIe (Quality Assessment for Integrated Information Environments), which allows to calculate 15 different DQ metrics for automated data- and schema-quality assessment of IISs [5]. With *multi-level*, we refer to the different aggregation-levels of an IIS, where DQ measurement can be carried out: on attribute-, concept-, data-source-, and IIS-level. Real-world settings in enterprises require automated assessment of numerous and complex data sources, where the navigation and identification of quality issues through humans experts is not trivial. Thus, we contribute in this paper with a human-centered approach to facilitate DQ assessment across multiple data sources and DQ dimensions.

This paper is organized as follows: in Section II, we discuss existing DQ visualization tools and distinguish them from our work. Section III introduces the visualization approach, which was developed in this research and Section IV covers its implementation in QuaIIe. In Section V, we demonstrate the applicability of the approach and conclude with an outlook on future work in Section VI.

II. Related Work

In recent years, DQ has been gaining importance both in research and in industry. However, the exploration of visualization techniques in the context of DQ is sparse. In this section, we discuss ongoing research and DQ tools, which support the visualization of DQ-related aspects.

Kandel et al. [6] implemented Profiler, a tool that uses data mining methods and type inference to target DQ issues in relational Databases (DBs). Profiler provides visual assistance and automatically suggests visualizations for identifying problematic data. In contrast to QuaIIe, the default implementation of Profiler supports only relational tables and other types of data need to be integrated manually. Moreover, Profiler requires the user to choose between multiple suggested visualizations. Such a decision requires domain knowledge of the analyzed data, since the choice for one type of visualization affects the conclusions drawn about the data quality [6].

Bors et al. [7] present a visual approach for analyzing DQ aspects and integrate functions for customizing and creating DQ metrics. Automatically computed DQ metrics are used for determining the DQ of a given data set. Their tool MetricDoc allows to assess DQ on tabular data sets through interaction techniques and distinct views, which provide quality-related information on different aspects. Similar to Profiler, MetricDoc considers only tabular data by default.

Abedjan et al. [8] made a tutorial on metadata discovery, i.e., data profiling. The authors present a series of techniques for profiling activities and stress the importance of visualization when transmitting and interpreting data profiling results. They present a visual representation of functional dependencies with a sunburst diagram, but do not provide detailed explanation of this chart. The hierarchical structure of the diagram is used to express dependencies between sets of attributes and the color is used to represent these sets. In our work, we present a different use of the sunburst diagram, as we make explicit use of the hierarchical order to represent the aggregation levels of an information system. Additionally, we assign semantic meaning to the colors in the chart, to inform the user about the qualitative state of an entire IIS with a simple glance on the diagram.

Xie et al. [9] present two different techniques to visualize and transmit DQ information for multivariate data. In one approach, a given data set is extended with quality measures as new data dimension. In the second approach, DQ information is integrated into visual attributes of existing multivariate visualizations. Both approaches are evaluated with different visualization techniques, such as, parallel coordinates, scatterplots, matrices, and star glyphs [9]. Based on the experiments in [9], the authors conclude that the selection of the visual attributes is the main success factor for the visualization. They noted that hue has a stronger capacity of transmitting quality information in parallel coordinates when compared to other visual attributes (e.g., the size). A possible explanation for this effect is that hue has a high degree of preattentive processing and does not need extra space [9]. This argument supports our design decision of using color, i.e., hue for transmitting quality information.

Gratzl et al. [10] present an interactive visualization technique for rankings. Rankings are typically influenced by one or more attributes, which can have multiple dependencies to other attributes. The visual tool LineUp introduced in [10] aims to assist the user in analyzing and comparing multiple rankings, as well as, on how changes in the attribute combination can affect the end-ranking. The visualization uses mainly bar charts for representing different ranking-related information, such as the individual attribute categories. Attributes are mapped to a normalized value and the sum of these values reflects the obtained ranking. The user is able assign a weight to the different attributes and gets visual feedback when the weights are changed.

Furmanova et al. [11] introduce a scalable visualization technique for tabular data called Taggle. The tool is open-source and can be accessed via a web application [11]. It consists of two main components: the tabular panel and the data selection panel. Columns from the tabular panel can be altered by means of aggregation, filtering, sorting, etc. They can be further inspected in the detailed data selection panel, which provides a visual summary of the data in form of histograms for suitable data sets [11]. The authors provide different mechanisms for aggregating and grouping data.

Blumenschein et al. [12] present a visual approach for analyzing high-dimensional data called SMARTexplore. Their tool uses a table-based technique that supports the identification and examination of clusters, patterns, and correlation in high-dimensional data. Rows represent single records or a collection of records (i.e., record groups) and columns represent different dimensions [12]. Record groups are aggregated and visually encoded by color, making it possible to compare them across dimensions. The identification of patterns is facilitated by automatically sorting groups and dimensions.

III. Visualization Approach

Our work represents a visual extension of QualIe, a DQ tool, which was originally introduced in [5]. QualIe performs automated DQ assessment of multiple data sources within an IIS [5]. During the quality assessment process, various DQ measurements are computed on different IIS aggregation levels and are stored in a *DQ report*.

This DQ report contains quality information for several DQ dimensions at the attribute-, concept-, data-source-, and IIS-level. In short, a data source represents a single schema in an IIS and it can contain an arbitrary number of concepts or associations (i.e., schema elements). A concept is a representation of a real-world object, e.g., in case of a relational DB, it can be a table. An association is a relationship between two or more concepts. An attribute is a property of a concept or an association. The internal arrangement of the DQ report resembles the hierarchical structure of the components of the IISs under inspection. More information on the schema representation in QualIe is provided in [5] and [13].

The quality computations on the four levels (attribute-, concept-, data-source-, and IIS-level) can contain a wide range of information. Displaying all this information at once could overload the user and thus, hamper the extraction of important information. Therefore, we split the information into two views: an overview (see Section III-A), in which we decided to use the sunburst diagram, and a detailed view (see Section III-B), which allows to dig deeper on demand by clicking on particular IIS elements in the overview.

A. Overview: Sunburst Diagram and Scoring Function

The data investigated with QualIe has a hierarchical structure and the relationships between the data items build a tree network [5]. A common visual representation of trees is a graph (consisting of nodes and edges), since it displays the structure of the data. However, it misuses space, e.g., by displaying edges that provide little information [14].

After analyzing the state-of-art for hierarchical data visualization and taking into account their advantages and disadvantages, we decided to explore and evaluate the potential of the sunburst diagram for the overview. An essential aspect for this decision was that the sunburst diagram is a compact visualization (in contrast to graphs) and overcomes a key drawback of other representations such as the tree-maps, since it preserves the structure of the tree hierarchy. A sunburst diagram indicates hierarchy through a series of rings, which are typically sliced based on the number of nodes within the hierarchy level [15]. The slices of the inner circles have a hierarchical relationship to the segments of the outer circles [16]. These relationships are formally translated as the parent-child relationship.

Figure 1 shows the overview on the Graphical User Interface (GUI), which was developed for QualIe. The sunburst diagram in sector (B) can be interpreted as follows: the white circle in the center represents an entire IIS, the slices of the first ring represent the different information sources of the IIS, e.g., DBs, ontologies, Comma-Separated Values (CSV) files, and the slices of the second ring represent the concepts (e.g., tables or classes) of the data sources. The sunburst diagram is ideally suited to represent large trees while preserving the hierarchical tree-structure [14]. This is an advantage over the traditional tree map, where the hierarchical structure of the tree cannot be easily detected due to its representation [14].

A common technique when employing the sunburst diagram is to use color (hue) for hierarchical grouping or for assigning categories to the schema elements. In the context of DQ, color can be used for transmitting information related to the qualitative state of the elements. To ease the learning effort for the user, we limit the number of colors and provide a color palette that indicates their meaning. To determine the color for each element in the IIS, a mechanism for assessing their quality rating is required, i.e., a *categorization function*. Due to the categorization function, a user can easily identify elements that require attention (i.e., have low quality) by inspecting the diagram.

With QualIe, multiple DQ dimensions can be analyzed simultaneously, which, in turn, can be assessed by one or more DQ metrics. We refer to the value computed by a metric as its *rating*. To visually transmit the quality information from the DQ report efficiently and effectively, a mechanism to summarize the data from the report is required. In our approach, we summarize DQ ratings by DQ dimension. Having a single (numerical) value per dimension allows to compute a quality rating for each element that contains DQ measurements. The quality rating is computed according to

$$rating_s = \frac{\sum_{i=1}^{n} dim_{si}.w_i}{n}, \tag{1}$$

$$dim_{si} = \frac{\sum_{j=1}^{m} r_j}{m}, \tag{2}$$

where $rating_s$ is the quality rating of an element *s*, w_i the weight of the dimension *i*, dim_{si} the dimension average of element *s* and dimension *i*, and r_j a rating computed with metric *j*. We contemplate the possibility to specify weights for the different dimensions and thus, to define the degree to which each DQ dimension contributes to the overall quality of the different elements. By default this value is set to one, assuming that all DQ dimensions are of equal importance.

After computing the quality rating, we can determine a category for it, that is, poor, fair, good, or excellent. The computed quality ratings are normalized values between zero and one, and thus, adhere to the DQ metric requirement (1) by Heinrich et al. [2]. This range is divided into four intervals, where each interval is assigned to a specific category. In combination with the quality rating, we determine the quality category of an element *s* as follows:

$$category_s = \begin{cases} poor, & \text{if } rating_s < 0.25 \\ fair, & \text{if } 0.25 \geq rating_s < 0.5 \\ good, & \text{if } 0.5 \geq rating_s < 0.75 \\ excellent, & \text{if } rating_s \geq 0.75 \end{cases}$$

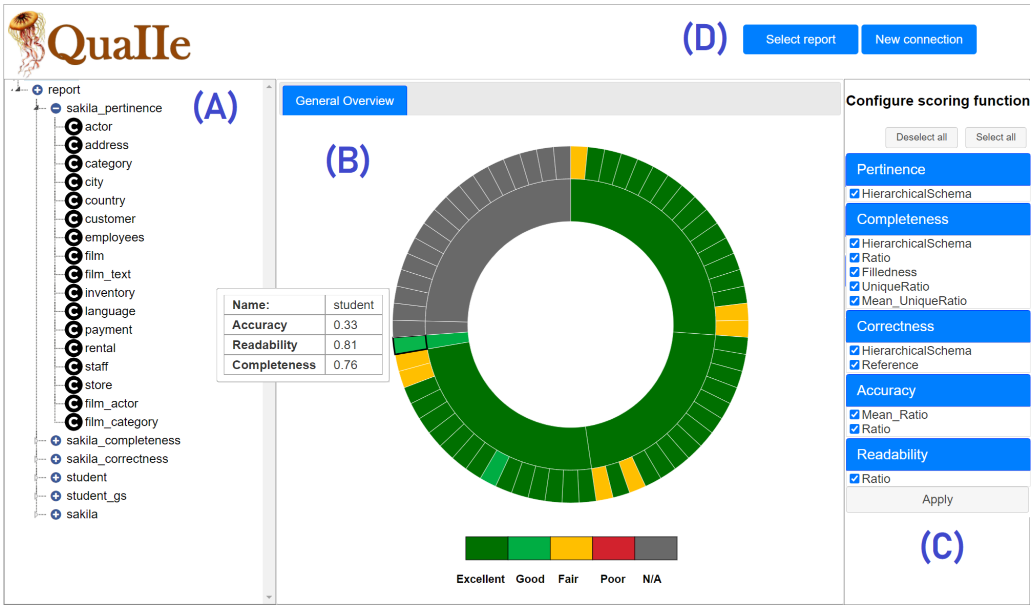
This function allows to determine a quality state for each element, which in turn allows to determine the overall quality state of IISs.

In addition, the sunburst diagram is enriched with tooltips that display extra information for each source, e.g., its name and computed quality category. Further, a filter (cf. sector (C) in Figure 1) allows to select DQ metrics and dimensions of interest. This feature allows to specifically analyze the impact of selected metrics and dimensions with respect to the calculated DQ category.

B. Detailed View

The purpose of the detailed view is to present the entire quality information from the DQ report in an organized and understandable way, including the information that was hidden in the overview. We assume that the data analyst is not interested in reviewing all DQ information at once, but rather has a goal in mind, e.g., an element with low quality detected through the sunburst diagram. This view can be activated by double-clicking on an element of the diagram or from the tree view (sector (A) in Figure 1). Figure 2 shows an excerpt of the detailed view of the concept `student`. It can be seen that the detailed view is organized in three sections.

The first section provides a summary of the quality dimensions of the selected element. Additional (textual) quality information provided by `QuaIIe` is displayed in the annotations table. The second section contains all DQ ratings measured by the different metrics for each dimension. Every metric is represented by a bar chart and the height of the bar represents the measured value. The third section displays all DQ information of the attributes, if available. Hence, the third section is only visible if the selected element contains attribute information. In Figure 2, we can notice that the attribute `id` has further quality information. The dimensions key attribute and completeness were measured by different metrics namely, “Pseudo_Boolean”, “UniqueRatio”, and “Filledness” (cf. [17]



Name:	student
Accuracy	0.33
Readability	0.81
Completeness	0.76

Pertinence
☑ HierarchicalSchema

Completeness
☑ HierarchicalSchema
☑ Ratio
☑ Filledness
☑ UniqueRatio
☑ Mean_UniqueRatio

Correctness
☑ HierarchicalSchema
☑ Reference

Accuracy
☑ Mean_Ratio
☑ Ratio

Readability
☑ Ratio

Fig. 1. Data quality visualization of an IIS

for details on the calculation). As can be seen in the chart, the dimension “Key_Attribute” was only evaluated by the metric “Pseudo_Boolean”, whereas the dimension “Completeness” was measured by the two remaining metrics. All three metrics computed a value of 1 for their respective dimension.

IV. User Interface Implementation for QuaIIe

The GUI presented in this paper was developed as a Java Server Pages (JSP) web application. Figure 1 displays the surface, which consists of four sections: (A) tree view of the IIS components, (B) sunburst diagram, (C) filter panel, and (D) resource loading and configuration.

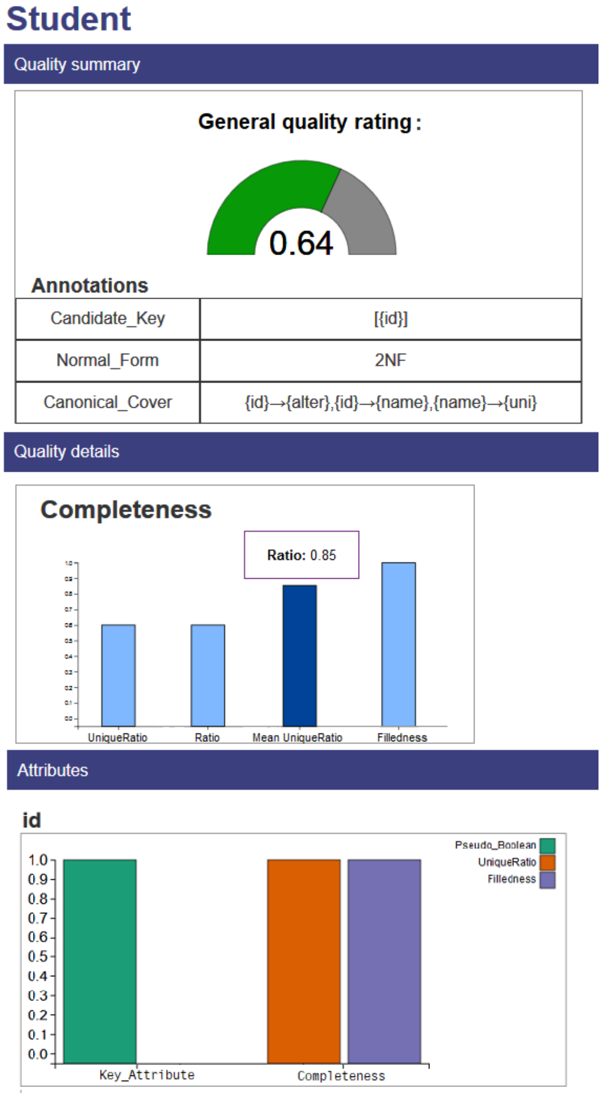
The sunburst diagram in the center provides on the one hand an overview on the entire IIS under inspection and allows on the other hand an easy identification of focal points. In Figure 1, we can quickly identify that there are seven elements with lower quality within our IIS, namely the elements with the yellow color. We are aware that using colors to distinguish graphical elements is not barrier-free, especially for users with red-green color blindness or complete color blindness. Although accessibility was no core requirement for this application, we plan to increase the customization by letting the user choose the respective colors and optionally use patterns to distinguish between the categories. These and other possibilities how graphical charts can be made accessible are discussed by Altmanninger and Wöß in [18].

The sunburst diagram was also extended with tooltips, which are displayed when the user hovers over the element of interest. In Figure 1, the tooltip for the data source `student` is highlighted, which contains a summary of its assigned quality dimensions. The tooltip indicates that the DQ dimensions accuracy, readability and completeness have a rating of 0.33, 0.81, and 0.76 respectively. Since the accuracy is notably lower than the readability and completeness, it subsequently reduces the total rating. This explains why the quality measurement for student falls within the quality range [0.5, 0.75[and is therefore rated as `good`.

The filter panel allows to configure the scoring function used to calculate the quality rating, and where the results are represented by the colors in the main chart. The user can select the DQ metrics and dimensions that should be considered when computing the quality rating of the elements. In this way, the user is able to analyze the impact that specific dimensions and metrics have on the quality state of the IIS. In addition, the filter panel also allows inspecting single DQ dimensions individually.

Resources are loaded/configured in the upper right section of the GUI (D). Existing DQ reports can be loaded and immediately visualized. Section (D) further allows to configure new data sources and provides a process for triggering new quality calculations. During the quantification process, special constraints are taken into account. For example, certain metrics require a *gold standard*, that is, a point of reference against which the available data source can be compared. If no such gold standard is provided, metrics that require one are hidden in the application.

To communicate with the back-end application (QuaIIe), several Java Servlets were created, where each serves a specific purpose, e.g., the configuration of a new MySQL data soure connection. The current version of the QuaIIe GUI supports the configuration of data sources and gold standards of the following types: ontologies, MySQL DBs, and CSV files. Configuration for all other data sources supported by QuaIIe (e.g., Cassandra wide-column stores) are planned for future work. Since QuaIIe creates an abstraction layer on all data sources, different types can be used in combination (cf. [5]). For example, a MySQL data source can have an ontology



Candidate_Key	[{id}]
Normal_Form	2NF
Canonical_Cover	{id}→{alter},{id}→{name},{name}→{uni}

Fig. 2. Detailed view of the QuaIIe GUI

as gold standard. After the configuration of the data sources, the DQ of all their elements can be measured. In this step, the seven quality dimensions supported by QuaIIe (accuracy, correctness, completeness, pertinence, minimality, readability, and normalization) can be analyzed. Each of these dimensions can be quantified using one or several metrics, which can be calculated on the schema- or data-level. When the user submits his/her selection of DQ dimensions and metrics, the DQ for the current data source is assessed. After the process is finished, the user can configure a new data source or can trigger the generation of a DQ report. When the DQ report is created, it is parsed into a JSON format suitable for different visualizations used in the application (sunburst diagram, bar charts, donut chart, etc.) and sent to the front-end application for its processing. All the visualizations of this application are created with the D3.js library [19].

V. Demonstration: DQ Assessment of Sakila DB

To demonstrate the capabilities of the visualization approach in the QuaIIe GUI, we present a case study, where we evaluated the DQ of the Sakila DB [20], which models a DVD rental store. For demonstration purposes, we also included a simple data set representing students information. Prior to DQ assessment, the connections to the data sources need to be established. For the demo, we created the following:

- `students_gs`: a CSV connection representing a file with data about students (name, address, academic title, matriculation number, etc.),
- `students`: a CSV connections where arbitrary attributes were altered/deleted from the original students data,
- `sakila_gs`: a MySQL connection, which represents the original Sakila DB and that is used as gold standard,
- `sakila_pertinence`, `sakila_complet- eness`, `sakila_correctness`: three ontology connections in Data Source Description (DSD) notation, which are representations of the original Sakila DB, but were modified to artificially downgrade the respective quality dimension.

The DSD notation is introduced in [13] and modification details of the three files are provided in [5]. For repeatability, the DSD files for all four data sources are published on the QuaIIe project website [21].

During the configuration process, a user can select an option to create a data source connection as the gold standard. When this option is selected, the connection is internally stored as a gold standard and can be reused for any other open connection.

To create an ontology or a CSV connection, the respective DSD file is uploaded and a name is assigned to the new connection. In the DQ form, the user selects the dimensions of interest and which metrics should be used for the quality assessment process. Each metric shows a list of all concepts. By selecting a concept, that metric is calculated for that concept at the concept level. In addition, the user can specify whether the calculation should be carried out at the schema level. In this use case for the Sakila ontologies, only one DQ dimension per connection is evaluated, e.g., relevance for `sakila_pertinence`. The process is repeated for all three files. In the case of the CSV connection, we analyze the dimensions accuracy, completeness, normality and readability in the schema and (whenever possible) also in the concept level. Details of the performed evaluation and DQ calculations are provided in [17].

Finally, the DQ report is created and its visualization is displayed immediately. Figure 1 shows the resulting sunburst diagram. When viewing the diagram, the gray color for `sakila_gs` and `students_gs` indicates that no DQ measurements have been performed for the gold standards. In addition, seven elements with lower quality (in yellow) stand out. By investigating these elements through the detailed view, the annotations indicate that all seven elements have been marked as *extra elements*. Extra elements are elements that exist in the investigated data source, but have no correspondence

ISBN: 978-1-61208-857-0

in the gold standard [5]. Thus, they impact the DQ dimensions pertinence (describing the prevalence of unnecessary elements) and correctness.

Next, we decided to specifically analyze the dimension completeness by specifying this in the filter. As expected, `sakila_completeness` has the lowest rating (compared to the other data sources) of 0.75. Nevertheless, the data source reaches the minimum for the rating `excellent` ($\geq$ 0.75). After further analysis and with the help of the chart, it is found that all concepts have a high rating for the dimensions completeness. Thus, another reason for the lower rating could be that complete concepts are missing from the data source. This assumption can be confirmed by comparing the number of concepts of `sakila_completeness` to its gold standard `sakila_gs` (e.g., with the help of the tree view). The comparison shows that the original database has 16 concepts, but `sakila_completeness` only has 14.

VI. Conclusion

In this paper, we presented an approach to visualize arbitrary complex IISs for exploring their DQ measurements across different DQ dimensions and metrics. The approach was implemented as a GUI for the DQ tool QuaIIe, originally introduced in [5]. The GUI supports a data analyst in selecting DQ dimensions and metrics of interest and in inspecting the qualitative state of several data sources. For ongoing and future work we plan the following:

- Averages can blur potential outliers, which might be important for users. Thus, the calculation of the quality rating will be extended with different aggregation functions, where the user can select the most appropriate one for a given use case.
- The user experience of the presented visualization will be evaluated to determine how easy and efficient the execution of DQ measurement tasks are perceived.
- The performance of the GUI should be evaluated with more real-world data.
- During the implementation of the GUI, the DQ tool QuaIIe has been extended with DQ monitoring capabilities. Thus, we additionally plan to extend the GUI to support the visualization of continuous DQ measurements over time.

Acknowledgment

The research reported in this paper has been funded by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG.

References

[1] W. W. Eckerson, “Data Quality and The Bottom Line Achieving Business Success through a Commitment to High Quality Data,” 2002. [Online]. Available: https://tdwi.org [Retrieved: 04, 2021]

[2] B. Heinrich, D. Hristova, M. Klier, A. Schiller, and M. Szubartowicz, “Requirements for Data Quality Metrics,” *Journal of Data and Information Quality*, vol. 9, no. 2, pp. 12:1–12:32, January 2018.

[3] R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, March 1996.

[4] J. M. B. Josko and J. E. Ferreira, “Visualization Properties for Data Quality Visual Assessment: An Exploratory Case Study,” *Information Visualization*, vol. 16, no. 2, pp. 93–112, 2017. [Online]. Available: https://doi.org/10.1177/1473871616629516

[5] L. Ehrlinger, B. Werth, and W. Wöß, “QuaIIe: A Data Quality Assessment Tool for Integrated Information Systems,” in *Proceedings of the Tenth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA 2018).* Nice, France: International Academy, Research and Industry Association, 2018, pp. 21–31.

[6] S. Kandel, R. Parikh, A. Paepcke, J. Hellerstein, and J. Heer, “Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment,” in *Advanced Visual Interfaces*, 2012. [Online]. Available: http://vis.stanford.edu/papers/profiler

[7] C. Bors, T. Gschwandtner, S. Kriglstein, S. Miksch, and M. Pohl, “Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics,” *Journal of Data and Information Quality*, vol. 10, no. 1, pp. 3:1–3:26, May 2018. [Online]. Available: http://doi.acm.org/10.1145/3190578

[8] Z. Abedjan, L. Golab, and F. Naumann, “Data Profiling: A Tutorial,” May 2017, pp. 1747–1751.

[9] Z. Xie, S. Huang, M. O. Ward, and E. A. Rundensteiner, “Exploratory Visualization of Multivariate Data with Variable Quality,” in *2006 IEEE Symposium On Visual Analytics Science And Technology*, October 2006, pp. 183–190.

[10] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit, “LineUp: Visual Analysis of Multi-Attribute Rankings,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2277–2286, 2013.

[11] K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresova, A. Lex, and M. Streit, “Taggle: Scalable Visualization of Tabular Data Through Aggregation,” *Information Visualization*, vol. 19, no. 2, pp. 114–136, 2019. [Online]. Available: https://taggle.caleydoapp.org [Retrieved: 04, 2021]

[12] M. Blumenschein, M. Behrisch, S. Schmid, S. Butscher, D. R. Wahl, K. Villinger, B. Renner, H. Reiterer, and D. A. Keim, “Smartexplore : Simplifying high-dimensional data analysis through a table-based visual analytics approach,” in *IEEE Conference on Visual Analytics Science and Technology (VAST) 2018*, 2018.

[13] L. Ehrlinger and W. Wöß, “Semi-Automatically Generated Hybrid Ontologies for Information Integration,” in *SEMANTiCS (Posters & Demos)*, ser. CEUR Workshop Proceedings, vol. 1481. Aachen: RWTH, November 2015, pp. 100–104.

[14] G. Wills, *Visualizing Hierarchical Data*. Boston, MA: Springer US, 2009, pp. 3425–3432. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_1380

[15] M. Schermann, “A Reader on Data Visualization,” 2019. [Online]. Available: https://mschermann.github.io/data_viz_reader [Retrieved: 04, 2021]

[16] F. Aps, “Sunburst Diagram,” 2015. [Online]. Available: https://datavizproject.com/data-type/sunburst-diagram [Retrieved: 04, 2021]

[17] L. Ehrlinger, B. Werth, and W. Wöß, “Automated Continuous Data Quality Measurement with QuaIIe,” *International Journal on Advances in Software*, vol. 11, no. 3 & 4, pp. 400–417, December 2018.

[18] K. Altmanninger and W. Wöß, “Accessible Graphics in Web Applications: Dynamic Generation, Analysis and Verification,” in *International Conference ICCHP (Computers Helping People with Special Needs)*, ser. Lecture Notes in Computer Science, vol. 5105. Berlin Heidelberg: Springer-Verlag, 2008, pp. 378–385.

[19] M. Bostock, “D3 – Data-Driven Documents,” 2020. [Online]. Available: https://d3js.org [Retrieved: 04, 2021]

[20] Oracle, “Sakila Sample Database,” 2006. [Online]. Available: https://dev.mysql.com/doc/sakila/en [Retrieved: 04, 2021]

[21] L. Ehrlinger, “Automated and Continuous Data Quality Measurement,” 2016. [Online]. Available: http://dqm.faw.jku.at/ [Retrieved: 04, 2021]

An Interaction Profile-based Classification for Twitter Users

Jonathan Debure
AIRBUS & CNAM
Paris, France
jonathan.debure@airbus.com

Stephan Brunessaux
AIRBUS
Paris, France
stephan.brunessaux@airbus.com

Camelia Constantin
Sorbonne University
Paris, France
camelia.constantin@lip6.fr

Cédric Du Mouza
CNAM
Paris, France
dumouza@cnam.fr

***Abstract*—Social networks have become a primary communication tool and are used by hundreds of millions of users daily. They bring together a wide variety of people, individuals, companies, public figures, media, influencers, etc. Users have different behaviours on social networks, such as different publication frequencies, number of followers or different user interactions. In the Twitter social network, for instance, users do not reply, quote or use mentions in the same way. Our intuition is that these interactions may characterise different user types and we consequently present in this work a non-supervised classification method based on interaction scores. We propose and experimentally compare different score estimations, leading our experiments to confirm the relevance of our approach.**

***Index Terms*—Social Network; Clustering; Behaviours; PageRank.**

I. Introduction

Nowadays, Online Social Networks (OSN) are omnipresent. There are different kinds of OSN, providing different services. Among the most famous ones, we can mention Twitter, which allows users to share short messages, media (videos and photos) and private messaging, Facebook, which proposes to share with friends photos, videos, messages and even to sell items or services, LinkedIn, which targets professional users and proposes a recruitment service and YouTube that hosts entertainement videos that users can stream. For several years, these social networks have been analyzed in different contexts. For instance, the content of some messages is analyzed to deduce users sentiments regarding a company or a product for marketing purposes. Recommendation systems use sociological studies in attempt to understand users behaviour for advertising purposes or recommendations of friends connexions. Other applications analyze content and/or user connections to detect inappropriate content or criminal activity.

Community detection in social network analysis is also gaining increasing attention as shown by the current tremendous amount of researches in this area. Existing approaches generally rely on the underlying social network graphs and attempt to group highly connected or frequently communicating users. Our goal here is quite different since we group people according to their type of profile: individuals, media, influencers, etc. The underlying assumption of this classification is that users react differently to messages contents, depending on their profile. We make use of interaction analysis to classify user accounts and to automate users classification.

The rest of the paper is structured as follows. In the Section II, we present a state of the art on online social networks researches. In the Section III, we present our datasets. Then, in Section IV, we explain our data model. In the Section IV, we present an analysis of our model on a global relational graph from our dataset and the obtained results. Then, in the Section VI, we present the same analysis, but this time on specific relational graphs and the obtained results. Finally, in the Section VII, we summarise the different analysis and future work.

II. Related work

Twitter users classification is mostly based on messages content: some studies use linguistic content [11] to classify users by their political orientation [7] or ethnicity. [3] proposes six different approaches to classify tweets content based on different symbols, keywords, categories or interacting messages into different groups, such as “Information”, “Conversation”, “Broadcast”, or “Other”. Content analysis is also employed by [2] which identifies and classifies users in three categories: “Bot”, “Human” or “Cyborg” based on message structures of entities such as URL, images, mentions, etc. [13] and [1] identify users sub-graphs (*i.e.*, community detection) by using a PageRank-based clustering that spreads computation scores through a random walk computation on the graph structure. Network structure-based users clustering and community detection are proposed by [10] and by [8]. The detected communities mainly reflect users’ connectivity and messages spreading across the network.

Several approaches have been proposed to perform community detection in social graphs, based on the follower/followee graph. Users exchange information in a privileged way inside the detected communities. Some existing methods determine a measure of users authority inside a social network based on node degrees [9]. Other approaches are based on the betweenness centrality measure proposed by [4]. They compute node authority depending on the distance between nodes, therefore highlighting users who are in the middle of the network. There are also approaches which consider recommendation scores provided by a PageRank-like algorithm that considers incoming links of nodes and that takes into consideration user centrality. A node with an high score of PageRank is a popular user with a high probability to propagate messages.

ISBN: 978-1-61208-857-0

III. Presentation of our dataset

To build our dataset, we use the Twitter API Stream that allows us to collect 1% of all tweets published on the platform. We collected tweets during a 5-month period of observation. We filter them to obtain two datasets: the first dataset gathers tweets about COVID, and the second dataset is composed of all tweets about NBA (National Basketball Association). Our final datasets consist of around 24 millions tweets.

NBA Dataset

The NBA dataset consists of 5M tweets produced by 2M unique users. From this 5M tweets, we identified 4.9M interactions (`Retweet`, `Quote`, `Reply` and `Mention`). It is important to note that not all tweets correspond to interactions while, at the same time various tweets may contain several interactions (such as retweets and/or quotes). To build the interaction graph used in our experiments, we only kept users that performed at least two interactions. Then, we computed the largest connected component (we used the NetworkX Python library [5]). This pre-processing step avoids to get a small sub-graph with isolated nodes that can reduce the global PageRank score of graph nodes.

The main characteristics of the NBA dataset are presented in Table I.

COVID Dataset

The COVID dataset consists of 21M tweets which allow to build an interaction graph of 6 million unique users and 17 million interactions. The extraction of the largest connected component during the pre-processing step produces a graph with 2,789,316 users.

The main characteristics of the COVID dataset are presented in Table II.

IV. The data model

We introduce in this section our notations and our data model. We consider the Twitter platform and its underlying directed graph of interactions $\mathcal{G} = (\mathcal{U}, \mathcal{E})$ where $\mathcal{U}$ denotes the set of nodes, *i.e.* users, $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{U}$ is the set of edges, such that $(u_1, u_2) \in \mathcal{E}$ means that user u_2 performed an action on the tweets of user u_1. We denote $\mathcal{A}$ the set of possible interactions that a user can execute on another user tweet. In the following, we consider that $\mathcal{A} = \{a_{rt}, a_{qt}, a_{rp}, a_{mt}\}$, which corresponds respectively to the actions of `Retweet`, `Quote`, `Reply` and `Mention`.

The restriction of the interaction graph $\mathcal{G}$ to a given action $a \in \mathcal{A}$ denoted $\mathcal{G}_a$ is the graph $\mathcal{G}_a = (\mathcal{U}_a, \mathcal{E}_a)$ with $\mathcal{U}_a \subseteq \mathcal{U}$ and $\mathcal{E}_a \subseteq \mathcal{E}$ such as $(u_1, u_2) \in \mathcal{E}_a$ if u_2 performed an interaction of type a on a tweet of u_1.

Obviously, all edges from an interaction graph do not represent the same level of interaction between users. Some interactions may happen frequently, while others may happen rarely. To capture this notion, we define an interaction weight ω as follows:

Definition 1 (Global interaction weight): The global interaction weight ω is a function $\omega : \mathcal{E} \to \mathbb{R}$ that takes into account all interactions between user couples.

Definition 2 (Specific interaction weight): The specific interaction weight ω for an action $a \in \mathcal{A}$ is a function $\omega : \mathcal{E} \times \mathcal{A} \to \mathbb{R}$. This score is mainly based on the interactions of a type $a \in \mathcal{A}$ and gives less (or no) importance to other types of interaction.

Finally, we assume the existence of a function $count$: $\mathcal{E} \times \mathcal{A} \to \mathbb{N}$, such as $count((u_1, u_2), a)$ is the number of interactions of type a that u_2 performed on tweets of u_1.

V. Global interaction score-based clustering

This approach is a little different from our original goal, which aimed to identify users with the same "profile" (role) within the social network using different interactions. Our intuition is, that clustering based on diff Better interactions between users provide more relevant clusters.

A. Global interaction occurrences-based clustering

Definition 3 (Occurrences-based global interaction score): The occurrences-based global interaction score σ_u^g for a user u is defined as:

$$\forall v \in \mathcal{U}, \omega(v, u) = \Sigma_{a \in \mathcal{A}} \; count((v, u), a)$$
$$\sigma_u^g = \log\left(\frac{\Sigma_{v \in \mathcal{U}} \; \omega(v, u)}{max_{w \in \mathcal{U}}(\Sigma_{v \in \mathcal{U}} \; \omega(w, v))}\right) \tag{1}$$

Note the normalization of the score and the usage of the $\log$ function to smooth the differences between accounts.

According to this interaction score, since data is not tagged, we decided to use a non-supervised clustering. More precisely, we chose the K-Means clustering algorithm for its scalability and because it is known to give good clustering results. To determine the number of K-clusters, we rely on the Silhouette Score [12].

For the occurrences-based global interaction score approach, we observe that the Silhouette score is increasing with the number of clusters (see Fig. 1). It illustrates that no clusters number appears to be better than another (except maybe clusters with a single user). Moreover, the manual analysis of a clustering, for example with $K = 4$ or $K = 5$ reveals that the clusters obtained contain very heterogeneous classes of users.

B. Global interaction PageRank-based clustering

It has been shown that PageRank can accurately compute influence ranks since it is not influenced by the number of followers but by the user interactions [6]. Consequently, we expect that a PageRank-based global interaction score will provide a better user classification. The PageRank score for a user $u_i \in \mathcal{U}$ is estimated by the following formula:

$$PR(u_i) = (1 - \alpha) + \alpha \sum_{u_j \in In(u_i)} \frac{PR(u_j)}{Out(u_j)} \tag{2}$$

TABLE I
NBA DATASET: STATISTICS

	Followers	Friends	# Tweets	# Quotes	# Retweets	# Mentions	# Replies
Value Count	882494	882494	1935124	561041	472376	644758	211985
Mean	4000.47	1096.57	2.49	1.25	1.93	2.27	0.45
Median	328	445	1	1	1	1	0
Std Dev	200434.26	4632.68	10.78	3.21	6.69	7.66	2.32
Min	0	0	1	0	0	0	0
Max	87244738	1480293	9171	1394	1478	2002	748

TABLE II
COVID DATASET: STATISTICS

	Followers	Friends	# Tweets	# Quotes	# Retweets	# Mentions	# Replies
Value Count	2789316	2789316	6278280	1783237	1699905	1945609	588131
Mean	3832.08	1128.18	2.64	1.31	2.49	2.01	0.37
Median	287	435	8348	1	1	1	0
Std Dev	128751.36	4644.14	8.27	3.09	8.76	6.74	2.37
Min	0	0	1	0	0	0	0
Max	85941911	1907480	8768	929	7910	2823	1480

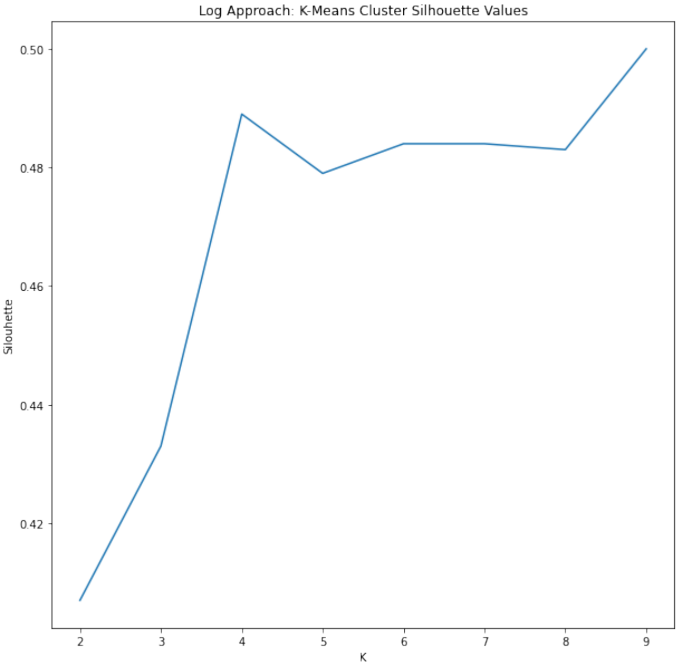


Fig. 1. Log Basic Approach: K-Means silhouette

where $In(u_i)$ denotes the set of users that have an interaction with u_i (*i.e.*, $\{u_j \in \mathcal{U}, (u_i, u_j) \in \mathcal{E}\}$), $Out(u_j)$ the out-degree of user u_j, α is a dumping factor.

We take into consideration multiple occurrences of the same interaction between two users by supposing that they illustrate a strong interaction between those users. This is modeled by the edges weights of the interaction graph $\mathcal{G}$. The weight of an edge between two users is the total number of interactions between them. In order to compute a PageRank score using edges weight, we use the Weighted PageRank (WPR) Algorithm [14]. WPR assigns higher scores to more important nodes instead of dividing the score between their neighbours. Nodes will get a value proportional to their number of in-interactions (interactions a user had with his tweets) and out-interactions (interactions a user had with tweets of other users). Consequently, we adopt the following definition:

Definition 4 (Interaction weights):
The in-interactions weight $\mathcal{W}^{in}_{(i,j)}$ and out-interactions weight $\mathcal{W}^{out}_{(i,j)}$ for an edge $(u_i, u_j) \in \mathcal{E}$ are estimated as:

$$\mathcal{W}^{in}_{(i,j)} = \frac{\sum\limits_{a \in \mathcal{A}} count((u_j, u_i), a)}{\sum\limits_{v \in In(u_i)} \sum\limits_{a \in \mathcal{A}} count((v, u_i), a)}$$

$$\mathcal{W}^{out}_{(i,j)} = \frac{\sum\limits_{a \in \mathcal{A}} count((u_i, u_j), a)}{\sum\limits_{v \in Out(u_i)} \sum\limits_{a \in \mathcal{A}} count((u_i, v), a)} \tag{3}$$

Finally, we adapt the Weighted PageRank proposed in [14] to take into consideration interaction weights on edges.

Definition 5 (Weighted interaction PageRank score):
Using the previous PageRank formula and the interaction weights defined above, we estimate:

$$WPR(u_i) = (1 - \alpha) + \alpha \sum_{p_j \in In(u_i)} WPR(u_j) \times \mathcal{W}^{in}_{(j,i)} \times \mathcal{W}^{out}_{j,i)} \tag{4}$$

VI. INTERACTION PROFILES-BASED CLUSTERING

By considering all interactions as similar, we are masking differences in the user behaviours. Indeed, the analysis of a few accounts seems to reveal that some users appear to favour certain interactions over others and this could be a relevant classification criterion. To verify this intuition, we built an interaction profile for the users that is defined as follows:

Definition 6 (Interaction profile): The interaction profile of a user u is a quadruplet $\sigma_u^p(\sigma_{rt}(u), \sigma_{qt}(u), \sigma_{rp}(u), \sigma_{mt}(u))$ where each dimension σ_a is the specific interaction score determined on the graph restriction $\mathcal{G}_a$.

As for the global approach, we compare the straightforward approach with specific interaction scores estimated with the number of interactions of the corresponding action, and the PageRank one.

A. Occurrences-based interaction profiles

For this approach, we consider that a specific interaction weight for an interaction a is estimated on the restricted graph $\mathcal{G}_a$ as:

$$\forall (u, v, a) \in \mathcal{U}^2 \times \mathcal{A}, \omega(u, v, a) = count((u, v), a) \quad (5)$$

Consequently, our interaction profile scores are estimated as follows:

Definition 7 (Occurrences-based interaction profiles scores): The scores for the occurrences-based interaction approach σ_u^p for a user u is defined as:

$$\forall a \in \mathcal{A}, \sigma_a(u) = \log\left(\frac{\Sigma_{v\in\mathcal{U}}\omega((v, u), a)}{max_{w\in\mathcal{U}}(\Sigma_{v\in\mathcal{U}}\omega((w, v), a)}\right) \quad (6)$$

Once these scores are computed, we perform the K-Means non-supervised clustering. To evaluate the clustering quality, we have performed a human validation which consists of manually analyzing a sample of 50 accounts randomly chosen inside each cluster.

We observe that, with the occurrences-based interaction profile approach, our clusters remain heterogeneous, as well as with the occurrences-based global approach: all kinds of users are present in each cluster. This phenomenon can be explained by the fact that this method only uses in-degree values. However, we aim at classifying users based on interactions on their messages. It has been shown that messages shared by popular or central users of the graph can be spread efficiently [15].

B. PageRank-based interaction profiles

Instead of a straightforward interaction scores computation, based on the number of occurrences, we estimate them using a PageRank approach. Therefore, we considered the graph restriction $\mathcal{G}_a$ of each interaction and performed the weighted PageRank algorithm to compute the associated dimension of the interaction profile. Our intuition is that capturing the "influence" of a user on a given interaction (*i.e.* his capacity to generate a given interaction on the network) better characterizes a user behaviour.

Definition 8 (PageRank-based interaction profiles scores): The scores for the PageRank-based interaction approach σ_u^p for a user u is defined as:

$$\forall a \in \mathcal{A}, \sigma_a(u) = WPR_{\mathcal{G}_a}(u) \quad (7)$$

TABLE III
WEIGHTED PAGERANK: CLUSTERS SUMMARY

	Weighted PageRank Value
Cluster 1	Reply PageRank is in average **9.10% smaller**, Retweets PageRank is in average **26.56% smaller**, Quote PageRank is in average **29.39% smaller**.
Cluster 2	Reply PageRank is in average **70.84% greater**, Mentions PageRank is in average **20.30% smaller**, Quote PageRank is in average **13.99% smaller**.
Cluster 3	Retweet PageRank is in average **520% greater**, Quote PageRank is in average **230% greater**, Mention PageRank is in average **204% greater**.
Cluster 4	Reply PageRank is in average **182% greater**, Mention PageRank is in average **181% greater**, Retweet PageRank is in average **84.71% greater**.
Cluster Outliers	Reply PageRank is in average **493% greater**, Retweet PageRank is in average **3155% greater**, Quote PageRank is in average **3857% greater**.

where $WPR_{\mathcal{G}_a}(u)$ is the Weighted PageRank score computed for u on the graph $\mathcal{G}_a$, and $\mathcal{G}_a$ is the reduction of the graph $\mathcal{G}$ for the interaction a.

Once these scores are computed, we also perform the K-Means non-supervised clustering. The clustering produces clusters with very different characteristics (see Table III). Globally, we see that `Reply` actions are what is mostly done by real individual users. On the contrary, entities (companies, media, etc) generate more `Retweet` actions. `Mention` can be generated by both human and entities. As there is also a correlation between popularity and `Retweet` actions, we can consider `Quote` as a kind of `Retweet`. Using weighted PageRank on the different interaction graphs to estimate the interaction scores allows to demonstrate the importance of the nodes that interact with the user. This is why we obtain homogeneous clusters with a large majority of similar users.

As for the occurrences-based interaction profile approach, we evaluated clustering quality with a human validation by manually analyzing (i.e Read users timelines, descriptions, photos) a sample of 50 accounts randomly chosen inside each cluster. Since the clusters are more homogeneous, it was possible to qualify the different classes of users we identified. Table IV presents the results of our analysis where Types are defined from our manual analysis of each account.

Finally, we perform a last experiment to validate our clustering. We consider 100 new users we manually "tagged" with a cluster id, according to the cluster composition we observed with our initial dataset. Then, we use our clustering algorithm to allocate them in a cluster. For these new users, we obtain that 96% of them were tagged with the good cluster id, which means that the clusters we obtained correspond to well-identified classes of users.

TABLE IV
WEIGHTED PAGERANK: CLUSTERS COMPOSITION

	Size	**Composition**	**Types**
Cluster 1	92.63%	100% composed from common users	Common users
Cluster 2	5.44%	55% composed from common users and 45% popular users (more than 4000 followers)	Moderately popular users, local celebrities, doctors, media specialists and active community users
Cluster 3	0.59%	55% composed from entities and 45% human users but mainly above 10 000 followers	Entities, professional users, brands, hospital, city and feed/news accounts
Cluster 4	0.66%	60% composed from popular user more than 4000 followers and 35% users with more than 10 000 followers	Influencers, writers, journalist, attorneys
Cluster Outliers	0.68%	60% human users, 40% entities. With 45% users with more than 100 000 followers and 40% with more than 10 000 followers	Celebrities, international news, politicians and brands

VII. CONCLUSION

This article presents a method to cluster Twitter users based on the interactions on their tweets. Based on interaction graphs and Weighted PageRank computation, we determine the user interaction profiles. Then, we perform a K-means non-supervised clustering which groups users with similar interaction profiles. Our experiments and manual validation confirm that this approach provides relevant clusters. As future work, we intend to study the parameters that influence the differences between the number of followers within the same cluster. We will also consider the graph dynamicity to propose an adaptive cluster re-computation on a sliding window.

REFERENCES

[1] R. Andersen, F. Chung, and K. Lang. Local partitioning for directed graphs using pagerank. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 166–178. Springer, 2007.

[2] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30, 2010.

[3] S. Dann. Twitter content classification. *First Monday*, 2010.

[4] L. C. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[5] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. In G. Varoquaux, T. Vaught, and J. Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[6] B. Hajian and T. White. Modelling influence in a social network: Metrics and evaluation. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 497–500. IEEE, 2011.

[7] O. Hanteer, L. Rossi, D. V. D'Aurelio, and M. Magnani. From interaction to participation: The role of the imagined audience in social media community detection and an application to political communication on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 531–534, 2018.

[8] B. S. Khan and M. A. Niazi. Network community detection: A review and visual survey. *arXiv preprint arXiv:1708.00977*, 2017.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600, 2010.

[10] S. Nandanwar and M. N. Murty. Structural neighborhood based classification of nodes in a network. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1085–1094, 2016.

[11] M. Pennacchiotti and A.-M. Popescu. A machine learning approach to twitter user classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 2011.

[12] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[13] S. A. Tabrizi, A. Shakery, M. Asadpour, M. Abbasi, and M. A. Tavallaie. Personalized pagerank clustering: A graph clustering algorithm based on random walks. *Physica A: Statistical Mechanics and its Applications*, 392(22):5772–5785, 2013.

[14] W. Xing and A. Ghorbani. Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314. IEEE, 2004.

[15] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern. Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, volume 104, pages 17599–601. Citeseer, 2010.

Analysing CV Corpus for Finding Suitable Candidates using Knowledge Graph and BERT

Yan Wang
Capgemini Engineering DRI
Vélizy-Villacoublay, France
Email: yan.wang2@altran.com

Yacine Allouache
Capgemini Engineering DRI
Vélizy-Villacoublay, France
Email: yacine.allouache@altran.com

Christian Joubert
Capgemini Engineering DRI
Vélizy-Villacoublay, France
Email: christian.joubert@altran.com

***Abstract*—Recruiter and candidate are the two main roles in the process of employment. Even though there is an abundance of job openings and a scarcity of qualified candidates to fill those openings, the objective is to offer only the profiles that fit the requirements of clients. Bidirectional Encoder Representations from Transformers (BERT) have been proposed in 2018 to better understand client searches. The challenges today are the frequent evolution of the experiences in Curriculum Vitae (CV) and the need of adaptive data for the specific staffing tasks of BERT. In this paper, we present an approach of ranking candidates based on competence keywords. There are four stages. First, we use Term Frequency–Inverse Document Frequency (TF-IDF) Vectorizer to calculate the score of matching between a competence keyword and a corpus of CVs. Second, we apply the Weighted Average Method to calculate a global score of CV based on two types of competence keywords – function and specialty. Third, we construct a Knowledge Graph (KG) from the structured Competence Map (CMAP), which can classify the relationships of bidirectional association and aggregation. At last, we propose to use the Named-Entity Recognition (NER) and Masked Language Modeling (MLM) of BERT to better identify tokens from the input inquiries of the client. The experiments are using the CVs from the HR (Human Resource) management system of Altran.**

Keywords—CV; CMAP; Knowledge Graph; BERT; TF-IDF Vectorizer; NER; MLM; HR-analytics; Job Matching.

I. INTRODUCTION

HR-analytics is always essential for companies to improve workforce performance. It allows many companies to exploit the immense amounts of data collected from their customers, their markets, social networks, real-time applications, and even the cloud. There is a coherent connection between engagement, performance and profit. It is imperative to generate performance results at all levels of the organization in order to take a position in the market and to stimulate growth. Recruiting talented candidates is not enough. However, it is important that people are assigned to specific roles where their talents will have the greatest impact on achieving company goals and where they are most likely to remain fully engaged. Job matching involves defining superior performance of each position and using objective criteria to determine who gets employed. Traditional hiring methods that only use a job description and a list of desirable candidates, technical, educational and professional experiences as filters, with a favorable interview are not working effectively. The process of job matching goes beyond conventional employment methods to create the most comprehensive definition of the job. This makes clients choose the right person for the job that suits them best. The result is a person who is happier at work and who makes good progress in meeting performance goals. Job Matching issues always exist in the center of operational and staffing concerns.

The similarity calculation is the mechanism that consists in evaluating the distance between two objects. Similarity measurements make it possible to solve problems from various fields such as text mining, image recognition, Nature Language Processing (NLP), computer vision, speech processing, image processing, and so on. In recent years, the recruiting process, Chatbot, search engines and recommendation systems have brought these technologies up to date. In this paper, we focus on the search engine for recruitment and staffing whereas Chatbot and recommendation systems are studied in future work.

Most staffing search engines do not consider about the evolution of the experiences. The objective of this paper is to use the NER of BERT [1] to extract relevant competence keywords from candidate experiences and client requests. To achieve this, the model of BERT that we use is fine-tuned by the vocabulary of competence keywords using the MLM of BERT. A Knowledge Graph is proposed for a semantic structure to help users find more precise results. However, the vector BERT assigns to a word is a function of the entire sentence. The vectors can be different with the same word. To deal with this problem, TF-IDF Vectorizer and Weighted Average Method are proposed to calculate the score based on the static vector.

The paper is organized as follows. Section 2 provides the related work of information parser, job matching and BERT. Section 3 explains the approach of extracting competence keywords and constructing the KG for recommendations. Section 4 shows the experiment of comparing the tool BERT with Linx. The conclusion and perspective are in Section 5.

II. RELATED WORK

Most of staffing software for recruiting and staffing in information extraction from CVs is either privatized by companies as an in-house tool or sold as a commercial product, for instance the private HR management system Linx in Altran. Artificial intelligence has contributed to great success in this field, but due to the confidential rule, some works are not well seen by the research community. Resume parsers have been proposed to extract information from the Internet [2], Github [3] and PDF [4], as well as from the general non-LinkedIn formats [5]. In case of Linx, only pre-processing is required instead of the parser.

For the purpose of improving the matching rate between jobseekers and available jobs, an ontology based expert system [6] can improve the accuracy of this matching. [7] proposes to extend the matching to multiple slots available to accept contracts. [8] presents a survey of exact string matching and approximate string-matching algorithms. Machine learning can continue to improve the performance of matching rate such as unsupervised feature extraction [9], using Convolutional Neural Network (CNN) [10] and deep Siamese network [11]. At last, by combing the knowledge graph and recommendation system, we expect to improve the matching quality with the help of the relationships between entities, for example, by using embedding-based methods [12] or path-based methods [13]. Considering the user preference [14] can also be interesting. In this paper, we consider the matching based on the competence keywords.

BERT is known as a more powerful and efficient technique than the other NLP tools like RNN, CNN and LSTM, which understands inquiries better than ever before. The embedding models word2vec [15] and GloVe [16] have been presented to be less effective in documents recognitions. The performance also differs from section to section [17]. In recent years, BERT has been used for sentence classification based on the suitability of job description [5] and semantic search over the corpus [18]. BERT will become more and more important in the staffing software.

III. THE PROPOSED APPROACH

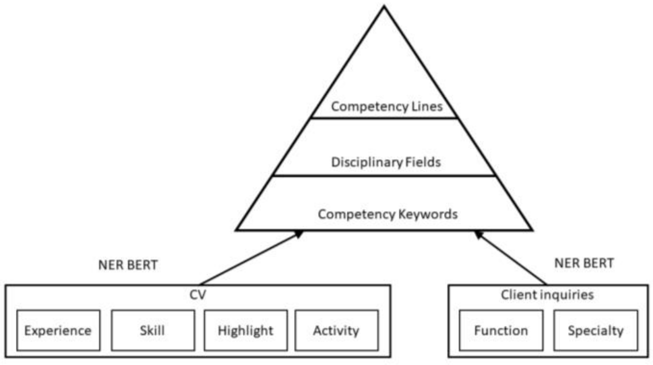


Figure 1. Structure of proposed approach to construct the KG.

The structure of the proposed approach is shown in Figure 1. In Section III.A, we define the competency keywords as the vocabulary. Section III.B and III.C will explain how to calculate the matching score. The pyramid represents the structure of the knowledge graph in Section III.D. Section III.E proposes to use MLM and NER of BERT to identify the features in client inquiries. The process of this approach is first to extract the keywords using NER of BERT, and then return a list of the ranked candidates from Knowledge Graph.

A. *Competency keyword*

The structure of a CV from Altran contains several sections. In this paper, we focus on four sections – experience, core skills, skill keywords and activity keywords. Competency keywords are used as a vocabulary for a client to search for candidates. Competency keywords are collected from the core skills, skill keywords and activity keywords sections, where they are stored by the HR system of Altran. But they are not standardized and refined. Accordingly, some keywords may have the same meaning in the semantic competency. In order to deal with the confusing keywords, a pre-processing of data is required. The purpose of cleaning is to remove ambiguous data, for example “JS” represents “Java Script”. The competency keywords are also classified into two types – function and specialty. Function refers to the position of the job coming from activity keywords such as “software developer” and specialty refers to the specific skill coming from skill keywords such as “mobile application”.

B. *TF-IDF Vectorizer for Matching Score*

The TF-IDF method is used at first to represent text as a vector of dimension D such that D is the number of words in the vocabulary - competency keywords. So, for each word of the vocabulary, we compute its TF-IDF in each section of the document – experience, core skills, skill keywords and activity keywords. To compute this score, we proceed in two steps. The first step is to compute the TF, which is the number of occurrences of the word in the document. Then, we calculate the IDF, which is a metric of the importance of the word. The intuition behind the definition of IDF is if a term is present in more documents, then it is more important.

$$w_{i,j} = \mathrm{t}f_{i,j} \times \mathrm{idf_i} \tag{1}$$

$$\mathrm{id}f_i = \log\left(\frac{1+N}{1+df_i}\right) \tag{2}$$

$\mathrm{w}_{i,j}$ = score of term i in document j
$\mathrm{t}f_{i,j}$ = number of occurrences of term i in document j
$\mathrm{d}f_i$ = number of documents containing the term i
N = total number of documents

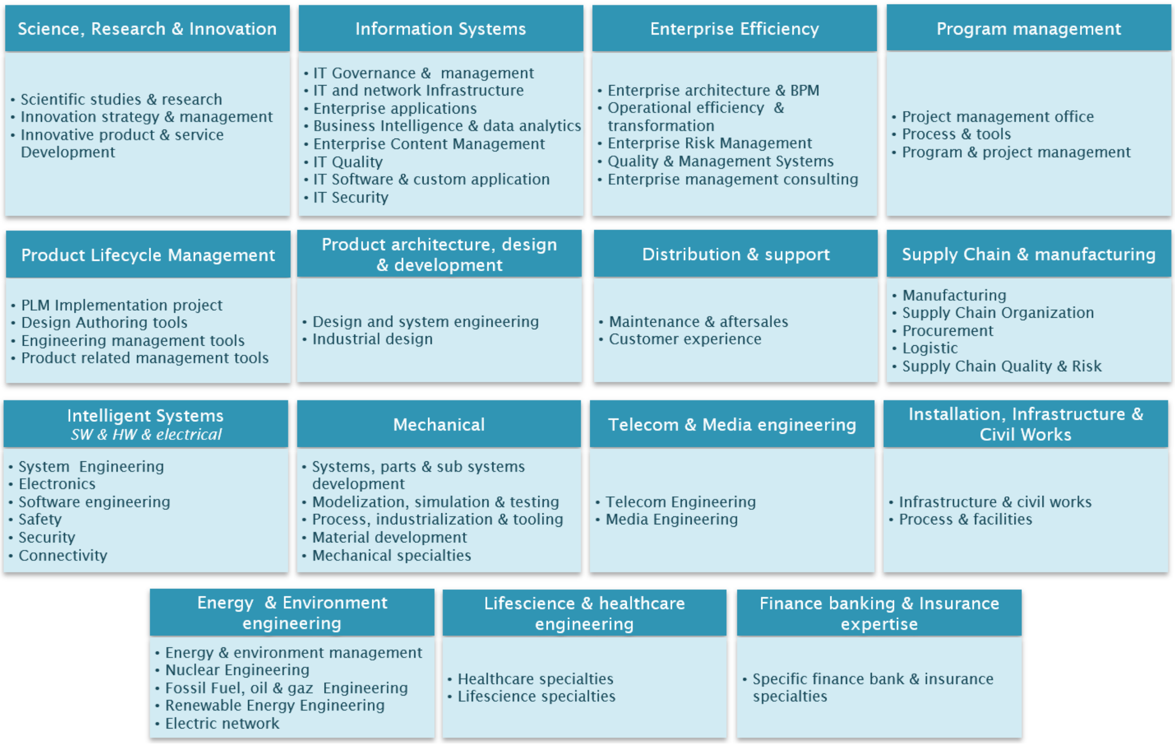
TF-IDF Vectorizer allows us to have an explainable representation, as each dimension *i* of our vector represents the score of word *i* in a document *j* or in the client query.

C. *Weighted Average Method for Global Score*

The text of each section in a CV – experience, core skills, skill keywords and activity keywords can be given a matching score. However, the strength of the connection in each of the four sections may be different. A Weighted Average Method has been proposed to calculate a global score concerning all four sections. The weighted average method is defined by the following equation:

$$\mathrm{S} = \frac{\sum \mu(s)\cdot s}{\sum \mu(s)} \tag{3}$$

where s is the score of matching in each section and µ represents the possibility of each section. The possibility of each section is pre-defined, but the value may change according to the types of competence keywords. For example, if the competence keywords are in the type of function like “software developer”, the possibility of activity keywords is higher than core skills and skill keywords. Conversely, the same is true for the type of specialty like “mobile application”.



Science, Research & Innovation
- Scientific studies & research
- Innovation strategy & management
- Innovative product & service Development

Information Systems
- IT Governance & management
- IT and network Infrastructure
- Enterprise applications
- Business Intelligence & data analytics
- Enterprise Content Management
- IT Quality
- IT Software & custom application
- IT Security

Enterprise Efficiency
- Enterprise architecture & BPM
- Operational efficiency & transformation
- Enterprise Risk Management
- Quality & Management Systems
- Enterprise management consulting

Program management
- Project management office
- Process & tools
- Program & project management

Product Lifecycle Management
- PLM Implementation project
- Design Authoring tools
- Engineering management tools
- Product related management tools

Product architecture, design & development
- Design and system engineering
- Industrial design

Distribution & support
- Maintenance & aftersales
- Customer experience

Supply Chain & manufacturing
- Manufacturing
- Supply Chain Organization
- Procurement
- Logistic
- Supply Chain Quality & Risk

Intelligent Systems
SW & HW & electrical
- System Engineering
- Electronics
- Software engineering
- Safety
- Security
- Connectivity

Mechanical
- Systems, parts & sub systems development
- Modelization, simulation & testing
- Process, industrialization & tooling
- Material development
- Mechanical specialties

Telecom & Media engineering
- Telecom Engineering
- Media Engineering

Installation, Infrastructure & Civil Works
- Infrastructure & civil works
- Process & facilities

Energy & Environment engineering
- Energy & environment management
- Nuclear Engineering
- Fossil Fuel, oil & gaz Engineering
- Renewable Energy Engineering
- Electric network

Lifescience & healthcare engineering
- Healthcare specialties
- Lifescience specialties

Finance banking & Insurance expertise
- Specific finance bank & insurance specialties

Figure 2. The competency lines and disciplinary fields in candidate environment of CMAP

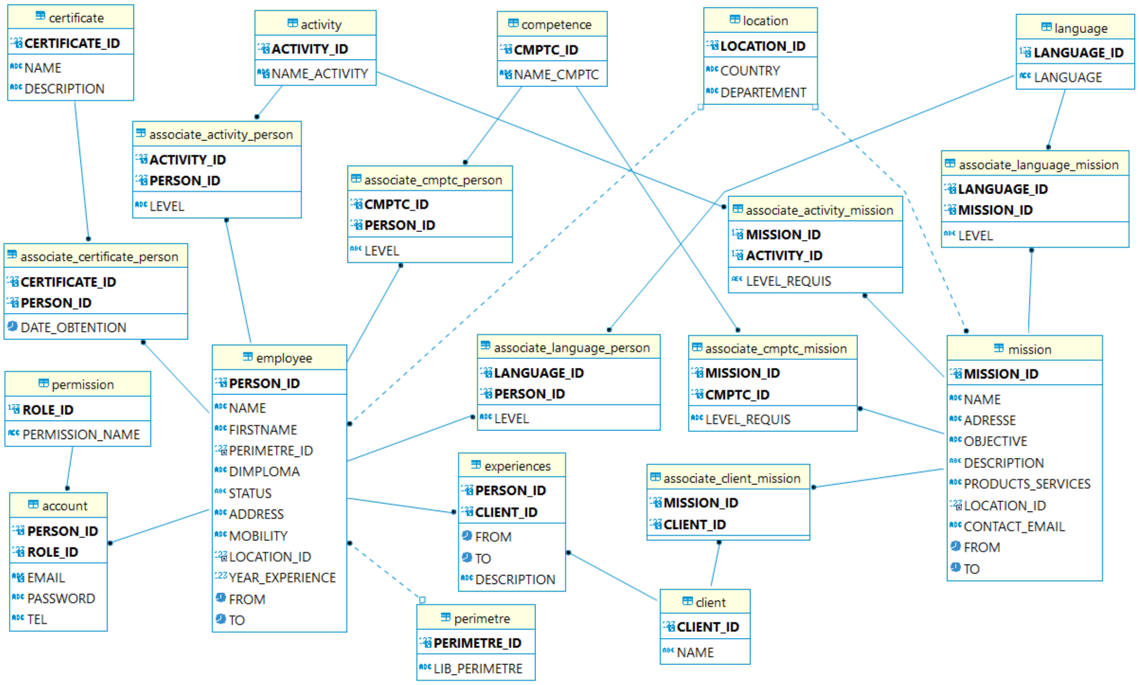


Figure 3. ER diagram of relational database for the profil of candidate and mission.

D. Knowledge Graph from CMAP

CMAP has been proposed as a homogeneous description of the competencies in Altran. The purpose is to allow and support competency-based management. It is not only a knowledge management system, but also a strategic workforce planning. CMAP contains a client environment and a candidate environment. Figure presents the structure of the candidate environment. The keywords in dark blue belong to the competency lines and the keywords in light blue belong to the disciplinary fields.

In this paper, a knowledge graph is proposed to ensure search results are contextually relevant to requirements. The first version of the KG with structed data is stored in the graph database [17]. The Neo4j Graph Platform that we choose for KG is an example of a tightly integrated graph database and algorithm-centric processing, optimized for graphs. There are three labels in the KG, where the label marks the node as the part of a group. The first label is constructed by the competency lines and the second label is generated from the disciplinary fields. The relationship of aggregation is identified such that the first label contains the second label based on CMAP and each second label contains several competency keywords represented by the third label. In addition, two competency keywords can have a relationship of bidirectional association which indicates a close connection between each other. The first version of the labels in the Knowledge Graph is based on structured data. It is a tree and sparse graph as well as bipartite graph distinguished between function and specialty.

E. MLM and NER using BERT

BERT [1] is a feature extractor based on deep Neural Probabilistic Language Models proposed in 2018. MLM is a fill-in-the-blank task, where a model uses the whole context words to predict the masked word. In our approach, BERT is first trained by the MLM method to modify the model distribution to be specific to our domain. This method consists of masking the tokens of a sequence with a masking token <**MASK**> and asking the model to fill this mask with the appropriate token. These tokens come from the keywords of Knowledge Graph, such that if the word belongs to the Knowledge Graph, we mask it with a probability ***P***. If not, we mask it with a probability ***1-P***. A threshold of ***P > 0.5*** is set to allow the model to focus on the competency keywords. This allows the model to be attentive to both the right context - tokens on the right of the mask, and the left context - tokens on the left of the mask.

Secondly, we continue the training of BERT with the NER method in order to extract the competency keywords from the profiles of candidates and the requests of clients. NER can classify tokens based on a class, for example, identifying a token as a person, an organization, or a location. In this approach, we use this technique to classify tokens as FUNC, SEP or NONE, such that *FUNC* means function, *SEP* means specialty, and *NONE* means a simple token. Therefore, we can further enrich our Knowledge Graph with the new competency keywords detected by BERT, which allows us to create a cycle of both refining BERT with the Knowledge Graph and developing the Knowledge Graph with BERT.

IV. EXPERIMENT

A. Dataset

In the HR management system of Altran - Linx, there is a searching engine for CVs. Location, keyword, availability, industry sector, competence domain, activity, certificates and years of experience are the options for searching. In order to evaluate our proposed approach, we make a search of “engineer” and “available” in “France”. In total, 106 CVs are exported in Excel, and 45 competence keywords are collected after pre-processing of removing confusing keywords. The interesting part of these CVs are filtered and then stored in the relational database MySQL shown in Figure 3. In this Entity Relationship (ER) diagram, the table employee stores the main information of the candidate as well as the availability time and years of experiences. For privacy, we keep name and personal contact information of the candidate as empty. Another table stores the content of experience. Five key tables store the keywords of certificate, function, specialty, location and language, respectively. At last, the table of permission is used to define the authentication of the user to login the system. The tool BERT implemented by the approach proposed in this paper is executed as a Micro-Service connected with the frontend.

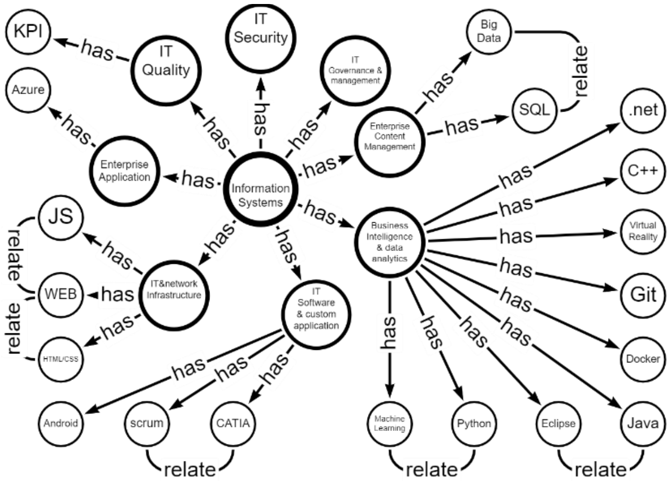


Figure 4. Knowledge Graph in the label of Information System.

The Knowledge Graph is constructed in Neo4j using 45 competency keywords and CMAP. To be different, MySQL is used to store the CVs and other original data and Neo4j is used to construct the KG for recommendations of competence keywords. Figure 4 presents the Knowledge Graph in the label of Information System. All the eight disciplinary fields surrounding this label have the aggregation relationship of “has”, for example “Information System” has a relation of “has” with “IT Quality”, “IT Security”, etc. 19 competency keywords are attached to the corresponding disciplinary field. Among the competency keywords, some of them can have a bidirectional association relationship of “relate”, for example “Java” is related to “Eclipse”. The relationship of “relate” can recommend the related keyword with explanation based on the request of the client.

B. *Results*

The CVs extracted from Linx are written in both English and French language. A multi-language tokenizer is used for pre-processing at first. To perform a search using BERT, we extract the competency keywords from the profiles of candidates, which allows us to represent a candidate as a list of competency keywords. The profile is represented by ID and the competence keyword is represented by KEY. Then, we apply the TF-IDF model to calculate the scores of KEY in each section of the profile and then Weighted Average Method to calculate the global stores of KEY. The global scores are stored by the index of ID:KEY and the inverted index KEY:ID in a Redis database. Secondly, for the query of clients, we have as input a list of words that represent the criteria searched by the client. This query is passed to the NER tool, especially multi-language DistilBERT base model [20], to extract the competency keywords. The model of DistilBERT that we choose can understand both English and French language. Since the inverted index contains KEY:ID, Redis can return a list of profiles containing these competency keywords sorted in decreasing order.

The NER tool is implemented using the model of DistilBERT-base-multilingual-case and is pre-trained by MLM. This tool is able to understand the language of both English and French. The objective of this tool is to use BERT methods to automate the recognition and extraction of keywords from the context.

At last, we compare the results obtained with Linx and BERT. However, Linx is a tool where the search principle is only based on the competency keywords presented in the platform and we are limited on how to express our query such that it must be written in a Boolean way containing only the competency keywords of the skills that we want. For this reason, we have implemented this tool using BERT in order to express our needs with natural language. So, for our query “Java” we chose to search the first 3 consultants that are proposed by our models. With the limitation of the pages in this paper, we choose three top ranking candidate results from both Linx and BERT in Table 1. The sections Experience, Core Skills and Skill keywords contain original data that we use for analysis. It is obvious to see that Linx does not pay attention to the experiences of candidates and the results of Linx only focuses on the sections of core skills and skill keywords, while BERT considers both the experiences and the skills. All the three candidates from BERT have experiences of Java and this skill is also highlighted in the sections of core skills and skill keywords. Especially, the first candidate from BERT has the experience of “Eclipse” which is related to “Java” in the knowledge graph. The volume of CVs and the latency are not considered in this experiment as we focus on the measurement of semantic similarity.

TABLE I. THREE TOP RANKING CANDIDATE RESUTLS IN BOTH LINX AND BERT

Linx				**BERT**			
ID_ candidate	***Experience***	***Core Skills***	***Skill Keywords***	***ID_ candidate***	***Experience***	***Core Skills***	***Skill Keywords***
16874	Developer at ENGIE - France Developer within the Genesys team then WattsOn at GEM IS Consultant at SFR - France Project manager / IT Engineer MOE - Scalable and corrective maintenance of various applications. IS Consultant at SOCIETE GENERALE - France IT MOE Engineer - Scalable and corrective maintenance of several applications	Test automation, Analysis & development of software requirements, Software design	DDD, Microsoft Visual Studio 2010, Entity Framework 4.0, C# 4.0, Java TDD	346575	Developer at AIRBUS - France: Creation of a Platform for Icing studies Developer at AIRBUS - France: Prerevolution Software Developer at DASSAULT AVIATION - France: Eclipse RCP based Verification Tool development for SCADE ENSO	Modelling, Model-Based Systems Engineering, IT Test & Validation, Automation	JUNIT, Maven, EMF, Tortoise Git Java 8
17071	Study engineer at BNP PARIBAS - France Universal Plug application - migration from Exadata to AIX, optimization Technical consultant at SOCIETE GENERALE - France I2R - Performance optimization of the Oracle Exadata database; migration from Oracle 11gR2 to 12c Technical consultant at SOCIETE GENERALE - France Optimization of the AGORA-AIR application database	DBA study, Database development, System and database	Oracle PL/SQL, Oracle Exadata, Oracle 12c, Oracle SQL Developer, Java 8	67747	Developer at ALTRAN - France: Clinuikali project - Java Application development Developer at ALTRAN - France: Python Application development Testing & Validation Engineer at ALTRAN - France: ACS - Automatic Testing within Continuous Integration	Software design, Marketing studies and strategy, Integration Validation Verification & Qualification,	Python, Software Development, Java, CSS 3, Html5
15868	Tester at HP ENTERPRISE SERVICES COMMUNICATION & MEDIA SOLUTIONS - France System Tests and Functional Tests on a virtualized system of the 4G network	Functional testing and validation, Test and technical validation,	Collabnet Svn, Teamforge Svn, Microsoft Office,	67711	Developer at ACS - France: OA: The system quantifies the fatigue at work for an employee during his working hours. Developer at ACS - France:	Application WEB, Core network mobile circuits,	HTML, JavaScript, AngularJS, Angular, Java

Linx				**BERT**			
ID_ candidate	***Experience***	***Core Skills***	***Skill Keywords***	***ID_ candidate***	***Experience***	***Core Skills***	***Skill Keywords***
	core, 2G / 3G environment (MAP, Diameter, AAA, EIR protocols) Integrator at HP ENTERPRISE SERVICES COMMUNICATION & MEDIA SOLUTIONS - France Integration of software solutions - Pre-Integration Tests Management of off-shore teams Industrialization & production engineer at TOTAL - France Outsourcing control on data transfer applications Definition of new flows; Dedicated projects with high business impact	Collaboration and networking	HP ALM, Collaborative Tools		LBS: Tracking the movement of physical assets on indoor and outdoor topology, by scanning barcode labels attached to assets or using smart labels, such as LORA or Antiote labels, which broadcast their location. Developer at ACS - France: OA: The system quantifies the fatigue at work for an employee during his working hours.	Product design and development	

V. Conclusion

This paper presents a keyword-based search engine for recruitment and staffing using knowledge graph and BERT. NER of BERT pre-trained by MLM can better recognize the competence keywords from the corpus of CVs and the natural language of client inquires. The knowledge graph composed of CMAP and competency keywords can recommend good results in the neighborhood domain. The proposed approach provides a way to use BERT for this specific task. The experiment based on BERT shows a better performance on finding good candidates than Linx.

The future work contains three parts. A method of using BERT to compute the score of the word embedding is required to replace the TF-IDF Vectorizer; the first version of KG that we propose in this paper is based on structed data. A richer KG is needed with the properties of each node and weighted relation for the dynamic management in case of a new competence; in Figure 2, the information about mission is also stored in the relational databases. A recommendation system with KG embedding method is needed for the matching between the profile of a candidate and the description of a mission. Unifying knowledge graph learning and recommendation is highly suggested to improve the matching efficiency.

Acknowledgment

This paper is sponsored by Direction of Research and Innovation (DRI) of Capgemini Engineering in France.

References

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[2] C. Mang, Online job search and matching quality. Ifo Working Paper, 2012.

[3] C. Hauff and G. Gousios, Matching GitHub developer profiles to job advertisements. IEEE/ACM 12th Working Conference on Mining Software Repositories. IEEE, pp. 362-366, 2015.

[4] J. Chen, L. Gao, and Z. Tang, Information extraction from resume documents in pdf format. Electronic Imaging, pp. 1-8, 2016.

[5] V. Bhatia, P. Rawat, A. Kumar, and R. R. Shah, End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT. arXiv preprint arXiv:1910.03089, 2019.

[6] V. Senthil Kumaran and A. Sankar, Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (EXPERT). International Journal of Metadata, Semantics and Ontologies, pp. 56-64, 2013.

[7] S. D. Kominers and T. Sönmez, Matching with slot ‐ specific priorities: Theory. Theoretical Economics, pp. 683-710, 2016.

[8] S. I. Hakak et al., Exact string matching algorithms: Survey, issues, and future research directions. IEEE Access, pp. 69614-69637, 2019.

[9] Y. Lin, H. Lei, P. C. Addo, and X. Li, Machine learned resume-job matching solution. arXiv preprint arXiv:1607.07657, 2016.

[10] C. Zhu et al., Person-job fit: Adapting the right talent for the right job with joint representation learning. ACM Transactions on Management Information Systems (TMIS), pp. 1-17, 2018.

[11] S. Maheshwary and H. Misra, Matching resumes to jobs via deep siamese network. Companion Proceedings of the The Web Conference 2018. pp. 87-88, 2018.

[12] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W. Y. Ma, Collaborative knowledge base embedding for recommender systems. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 353-362, 2016.

[13] H. Zhao, Q. Yao, J. Li, Y. Song, and D. L. Lee, Meta-graph based recommendation fusion over heterogeneous information networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 635-644, 2017.

[14] Y. Cao, X. Wang, X. He, Z. Hu, and T. S. Chua, Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. The world wide web conference. pp. 151-161, 2019.

[15] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[16] J. Pennington, R. Socher, and C. D. Manning, Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532-1543, 2014.

[17] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, PROSPECT: a system for screening candidates for recruitment. Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 659-668, 2010.

[18] A. A. Deshmukh and U. Sethi, IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles. arXiv preprint arXiv:2007.12603, 2020.

[19] M. Needham and A. E. Hodler, Graph Algorithms: Practical Examples in Apache Spark and Neo4j. O'Reilly Media, 2019.

[20] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

A Framework for Improving Offline Learning Models with Online Data

Sabrina Luftensteiner
Data Science
Software Competence Center Hagenberg GmbH
Hagenberg im Muehlkreis, Austria
email: sabrina.luftensteiner@scch.at

Michael Zwick
Data Science
Software Competence Center Hagenberg GmbH
Hagenberg im Muehlkreis, Austria
email: michael.zwick@scch.at

***Abstract*—The usage of available online data is rising as machines get equipped with more sensors to control and monitor processes. The produced data can be used to directly fit existing prediction models to enhance their accuracy, adapt to alterations within the environment and avoid the training of new models. During the online learning step, which is used for the adaptation of the models using online data, catastrophic forgetting of already learned tasks may occur. We propose a new framework that utilizes several state-of-the-art methods in deep learning, as well as machine learning to minimize catastrophic forgetting. The methods range from memory-based approaches to methods for loss calculation and different optimizers, whereat the framework also provides possibilities to compare the methods and their impact with each other. The proposed framework is specifically tailored for regression problems, focusing on industrial settings in the experiments section. It is able to cope with single and multi-task models, is expandable and enables a high variety of configuration possibilities for adaptation to a given problem.**

***Index Terms*— Online Learning; Catastrophic Forgetting; Regression; Domain Adaption.**

I. Introduction

In manufacturing, the usage of sensors and microprocessors on machines and their produced data is continuously increasing. This trend is part of Industry 4.0 and enables a huge source of structured and unstructured streaming data [1]. The produced data stream is used to achieve a higher level of operational efficiency, as well as productivity and furthermore enables a higher level of automatization and flexibility [2] [3]. Another important aspect regarding Industry 4.0 is the customization of applications with small batch sizes enabling flexible adaptions and optimizations [2].

At the moment, machine learning, especially deep neural networks, are very effective in solving various tasks, including classification and regression problems [4]. Industry 4.0 is able to utilize such machine learning techniques to create self-learning and adaptive systems for predictions, predictive maintenance, outlier detection and various other evaluations [2] [3]. The used techniques have to support domain adaptation to provide models with the needed adaptability for expanding and alternating environments, which is especially useful in custom process industry.

Many of the current learning approaches are based on batch settings, meaning the complete training data set has to be available prior to the learning task [5]. As sensors in Industry 4.0 settings continuously produce new data [1] [2], which are used to optimize models, the so-called offline learning is often not sufficient enough. Online learning, on the other hand, is able to deal with such continuous data streams and dynamically changing environments and additionally enables domain adaptation [6]. Online models have the ability to gain new knowledge over time while retaining previously gained knowledge to a certain extent [7]. The main issue of online learning models is catastrophic forgetting, meaning the forgetting or fading of previously learned knowledge due to the stability-plasticity-dilemma [7] [8]. Researchers have worked on solving this dilemma and came up with various solutions, ranging from memory-based approaches [9] to parameter-specific approaches, like *Learning without Forgetting* [10] or *Natural Gradient Descent* [11]. As different scenarios need different approaches, the need for a general framework covering various approaches to minimize catastrophic forgetting arises.

The motivation for the creation of an online learning framework is the easing of the path for the development of models, especially in Industry 4.0 applications. Our approach offers a high variety of configuration possibilities for various online learning scenarios, whereat it is possible to select from different models and solution approaches to identify the most fitting approach. As a base for each online learning cycle, an already pre-trained offline model is used, which is then further trained using online data. The availability of various configurations and models allows users to experiment with similar and unresembling methods for easier comparisons regarding which configuration is more suitable for a scenario. Data can be added in mini-batches or per sample, enabling an industry-like usage as clients often have varying requirements. A main advantage of the framework is the visualization of diverse metrics, e.g., mean-squared error, for each adaptation step over time, that is further referred to as time-step. The amount of time-steps can be dynamically adapted over time and represents the next adaptation of a model with online data. Using such a representation of the model development, it is straightforward to estimate if a model adapts well to new data.

The paper covers the following points: Section 2 addresses related work and highlights missings in current literature. In Section 3, the problem setup and the basic idea of our framework are discussed. A more detailed description of the framework structure is discussed in Section 4, followed by experiments and results in Section 5. Closing, Section 6 covers a conclusion and a prospective future work.

II. Related Work

In this section, we briefly recap existing frameworks and relevant topics for online and offline learning. The frameworks' advantages and drawbacks are highlighted regarding their usage in industrial environments and applicability. The second part of this section focuses on deep learning using state-of-the-art methods for decreasing catastrophic forgetting.

A. Existing Frameworks

Currently, only few frameworks with the support for online learning are mentioned in literature. An unsupervised online learning framework for moving object detection was presented 2004 by Nair et al. [12], focusing on the adaptation of the classifier. They start similar to our approach with an offline trained model and fit it continuously with new data. In comparison to their framework our framework has a broader field of application, as it can deal with various regression problems. An online multi-task learning framework for ensemble learning was developed by Xu et al. [13], focusing on time series data and insensitive loss functions. Both of the mentioned frameworks have a fixed model structure whereat Xu et al. also enable the selection of the loss function. Our framework supports the usage of various models to select from and provides additional configuration possibilities, e.g., loss function and optimizer for neural networks. These features enable the training of a broad variety of different models using the same setup and provide straightforward comparisons as the visualization of results is incorporated into the framework.

B. Relevant Deep Learning Topics

1) Multi-Task Learning: Multi-task learning uses the common knowledge of tasks to improve all tasks, whereat the bottom layers of the model are usually shared and the top layers are task-specific [14]. Such models are often used in the industrial field, as they enable the training of similar tasks in one model to save time and even enhance results. Continuously adding new tasks or data to an already trained multi-task model could lead to unwanted side-effects such as catastrophic forgetting [10].

The proposed framework supports the use for (online) domain adaptation [15], i.e., learning a model based on a source domain that performs sufficiently well on different but related target domains [16]. This scenario frequently arises in industrial settings, e.g., when a previously learned model needs to be applied in a different machine/tool setting or with different materials. This is often achieved by finding a common feature representation where source and target distributions become as similar as possible [17] [18].

2) Catastrophic Forgetting: Catastrophic forgetting is one of the main constraints in online learning and is caused by the stability-plasticity dilemma. This dilemma states that a model requires a certain plasticity for the integration of new knowledge, but also stability to prevent the fading of previously gained knowledge [19]. Researchers engaged themselves in finding a solution for this problem and came up with various approaches. Li et al. [10] propose a method which preserves the original capabilities of a multi-task deep learning model by taking the response of old tasks for the new data into account for the loss calculation. Kirkpatrick et al. [4] developed an algorithm analogous to synaptic consolidation in brains, which decelerates learning on certain weights in deep learning models depending on how important they seem to previously seen tasks. Zhang et al. [20] created a new optimizer where they exploit a connection between natural gradient descents and variational inference to enable further adaptive training. Rusu et al. [21] propose an approach where the network can utilize preliminary knowledge via lateral connections to previously learned features. Castro et al. [9] store the most representative samples of a task in the representative memory of the model and later reuse them for the fitting process.

III. Problem Setup

Consider an unknown target function $f : \mathbb{R}^N \to \mathbb{R}$ with N input features, e.g., a virtual sensor in an industrial manufacturing setting predicting an unobserved process quality measure based on N physical sensor measurements. Typically, the physical sensors are cost efficient, whereas the quality measure can only be recorded with significant effort with respect to hardware cost and setup time.

In a first step (offline step), a regression model $h : \mathbb{R}^N \to \mathbb{R}$ is trained using a homogeneous set $(X_S, Y_S = f(X_S))$ of all (source) data available at this point in time. The trained model is then used in production to cyclically predict the quality measure in order to control the production process.

At some point (online step), the model is required to adapt to additional operating environments, e.g., different machine/tool settings or additional materials. At this point, new (target) data $(X_T, Y_T = f(X_T))$ is gathered under the new operating environment that is different but related to the original data set (X_S, Y_S). The framework then has to learn an updated model $h' : \mathbb{R}^N \to \mathbb{R}$, which is able to predict the new data (X_T, Y_T) while still retaining the performance of the original source data (X_S, Y_S).

As during the production process data can be generated continuously or in bulk, the framework needs to be able to train the updated model incrementally, i.e., using only the next available sample or alternatively all the available data from the new operating environment at once.

IV. Framework

In this section, we describe the main parts of our framework, focusing on the central configuration file, as well as the learning algorithm. The framework was implemented in Python 3, using the PyTorch deep learning framework.

ISBN: 978-1-61208-857-0

```
CONFIG = {
  'App_Scen1': {
    'TASK_DICT': {
      'task_1': [(1, 50, False),
        (2, 20, True, True),
        (3, 10, False, True)]
      'task_2': [(2, 40, False),
        (3, 30, True, True)]
    }
     # additional dictionary entries
  }
  'App_Scen2': {
     # other scenario entries
  }
}
```

Fig. 1. General Configuration Dictionary.

A. *Configuration*

The configuration file is the core of our framework and is represented as a dictionary.

1) Application Scenario Configuration: The dictionary, see Fig. 1, is able to store configurations for more than one application scenario in sub-dictionaries, making it possible to access a specific scenario setup by using its key. The scenario dictionary contains, next to information about loading and saving paths for models and (training) data, information about the training process represented in a dictionary containing the different tasks of a scenario.

2) Task Training Configuration: The task dictionary consists of n entries, each containing a task name and an array of tuples, which represent the processing of the available data. One tuple contains four pieces of information. First of all, the time-step is specifying at which time-slot the task data is used in the model. Time-slots represent the the course of adding new online data in our mock-up scenario and are defined by the data assigned to the single time-steps. As the results and intermediate models are stored, it is straightforward to add new time-steps and start the model at specific steps, which enables a dynamic training. The second tuple value defines the used percentage of the available task data. It has a range between 1-100 whereat the sum of percentages for one task should not exceed 100. If the overall percentage of a task is below 100, the remaining data is used for testing. The third element of the tuple is a boolean flag indicating if the data should be trained elementwise or batchwise.

3) Optimizer and Loss Configuration: The framework can be used with arbitrary regression models as long as they are supported by the frameworks wrapper class (see Section IV-B). In case neural networks are used as training models, the framework enables the selection of an optimzer and a loss function. Currently, the framework supports the following optimizers and loss functions:

- ○ Stochastic Gradient Descent (Optimizer)
- ○ Noisy Natural Gradient Descent (Optimizer), as described by Zhang et al. [20]
- ○ Mean Squared Error (Loss)
- ○ Learning without Forgetting (Loss), as described by Li et al. [10]
- ○ Elastic Weight Consolidation (Loss), as described by Kirkpatric et al. [4]

We adapted the Learning without Forgetting approach [10] by using mean squared error instead of multinomial logistic loss used for the loss calculation of the new task in order to support regression tasks. We also had to adapt the elastic weight consolidation approach [4] by using mean squared error instead of cross entropy loss.

4) Other Configurations: Furthermore, the framework enables the selection of source and target column(s) as maybe not all features are used during training in order to allow different experiments with the available data for new models. Additionally, it is possible to define at which steps to start and end a learning run. This is especially useful for stopping training at a specific step and continuing later, as the model can be reloaded and trained further at a later time.

B. *Wrapper*

A wrapper stores the model and enables equal treatment of models. It is used as intersection point between the model and the learning algorithm of the framework and stores additional information, e.g., prediction results. The wrapper class also contains the calculation methods for the metrics, of which Root Mean Squared Error (RMSE), Maximum Absolute Error (MaxAE), sigma, sigma^2 and R^2 are available. Currently, the wrapper supports neural networks, linear regression models, elastic net models and random forest models.

C. *Algorithm*

For an outline of our main learning loop see Fig. 2. At first, the data is loaded and stored in memory according to the definition in the configuration, see Fig. 1. Afterwards, an offline model is either loaded or trained and placed into a wrapper, which is described in section IV-B. After this step, the main part of the algorithm starts. As long as the last time-step is not reached, the data belonging to the time-step is fetched from the dictionary and used to fit the model. Depending on the configuration of the task at the time-step, the model is either fitted batchwise or elementwise with the available task data. Adding new tasks is performed in the wrapper and, therefore, not incorporated in the algorithm, see Fig. 2. After fitting of the model, it is evaluated with the test data. After the last time-step is completed, the model can be saved depending on the configuration. Finally, the selected metrics are calculated and visualized.

D. *Visualization*

The framework enables visualizations of the training and testing results as Portable Document Format (PDF) plots, e.g., see Fig. 3 and Fig. 4. Numerical results are additionally stored in Excel sheets. In the configuration files, a user can specify

```
Step 1 : Load data and partion data according to
         configuration file into train/test,
         batchwise/elementwise and online/offline
Step 2 : Load or train offline model
while time-steps available do
    Step 3 : Get data for time-step
    Step 4 : Fit model with according data
    Step 5 : Evaluate model with test data
    Step 6 : Get next time-step
end
Step 7 : If required, save model
Step 8 : Calculate and visualize metrics
```

Fig. 2. Learning Algorithm. The algorithm can be started multiple times with varying time-steps and, therefore, enables a dynamic type of online learning.

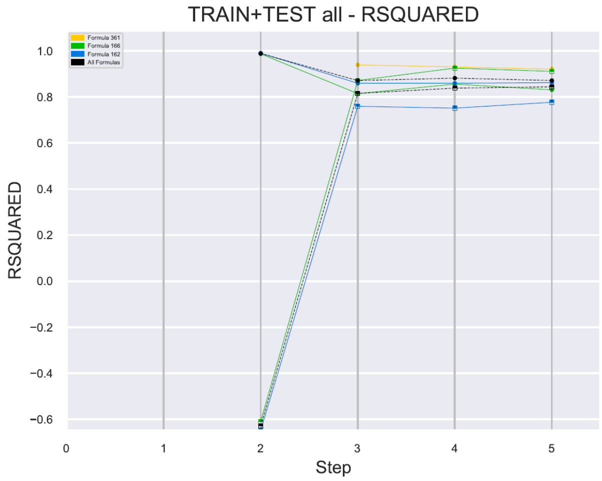


Fig. 3. Visualization example with black mean-line. Dots represent training results, half-filled squares testing results.

which metrics are visualized (multiple metrics are possible) and whether training and testing results should be visualized separately. Additionally, it is possible to anonymize the results in case of sensitive information has to be visualized.

V. Experiment

A. Dataset

We show the usability of our proposed framework using a resin production dataset provided by the Austrian company Metadynea. The dataset consists of three different recipes, each containing 5639 samples. Each sample of the dataset consists of 2692 features, which are composed of the following values: sample id, sample time, date, batch, spectrum light intensity, process pressure, process temperature, condensation time and various values representing the spectrum trend. The target is represented by a reference value measured in C.

B. Setting

1) Dataset Partition: For the experiment, we use five time-steps to simulate online learning whereat the first one is used to train an offline model. To simulate a real world application, with regards to adding new tasks and enhancing existing tasks, we partitioned the dataset the following way:

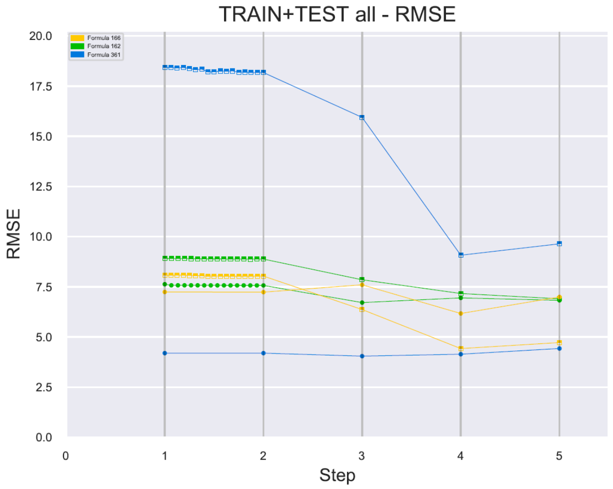


Fig. 4. Visualization example including elementwise adding of data.

- **Recipe 166**: 30% used in offline training at step 1; 40% added at step 4; 25% added at step 5; 5% used for testing
- **Recipe 162**: 25% used in offline training at step 1; 25% added at step 2; 25% added at step 3; 25% used for testing
- **Recipe 361**: 80% added at step 3; 19% added at step 4; 1% used for testing

2) Deep Learning Models: We use the following network architectures for our learning scenarios:

- **Feed-Forward Network** with 5 hidden layers, containing 35/30/25/15/7 hidden nodes
- **Feed-Forward Multi-Task Network** with 4 hidden layers, containing 300/50/30/15 hidden nodes

The models use learning rate schedulers and early stopping, as well as warm starts in the online learning part.

3) Model Configurations: To demonstrate the framework's possibilities, we chose the following model configurations for our experiments:

- Feed-Forward (Single and Multi-Task) with Stochastic Gradient Descent and MSE
- Feed-Forward Multi-Task with Stochastic Gradient Descent and Learning without Forgetting (LwF)
- Feed-Forward (Single and Multi-Task) with Natural Gradient Descent and MSE
- Feed-Forward Multi-Task with Natural Gradient Descent and LwF
- Feed-Forward Multi-Task with Stochastic Gradient Descent and Elastic Weight Consolidation (EWC)
- Linear Regression
- Random Forest
- Elastic Net

C. Results

The results of the experiments are presented in Table I, which contains the RMSE results of the training and testing environments. It is possible to see an improvement in nearly every model configuration regarding the first and last step of

TABLE I
RMSE TRAIN AND TEST VALIDATION RESULTS.

	Train/Test	Step 1 (Offline)	Step 2 (Online)	Step 3 (Online)	Step 4 (Online)	Step 5 (Online)
FF S with SGD/MSE	Train	10.73	13.49	10.15	7.80	9.11
FF S with SGD/MSE	Test	8.85	9.93	9.97	6.65	7.26
FF S with NGD/MSE	Train	12.87	12.7	9.81	8.32	8.28
FF S with NGD/MSE	Test	8.35	11.34	7.36	7.15	6.34
FF M with SGD/MSE	Train	13.24	11.76	10.30	10.38	9.19
FF M with SGD/MSE	Test	12.59	13.32	10.18	10.31	8.99
FF M with SGD/LwF	Train	10.72	13.71	10.15	7.08	9.11
FF M with SGD/LwF	Test	8.85	13.24	9.97	6.65	6.27
FF M with SGD/EWC	Train	19.13	18.47	20.01	18.78	21.46
FF M with SGD/EWC	Test	17.19	17.46	18.67	18.13	20.33
FF M with NGD/MSE	Train	10.91	12.42	9.80	10.02	10.50
FF M with NGD/MSE	Test	10.64	11.43	9.48	9.71	9.26
FF M with NGD/LwF	Train	12.87	12.70	9.81	8.32	**8.20**
FF M with NGD/LwF	Test	9.33	11.34	7.36	7.16	**6.29**
Linear Regression	Train	2.06E-11	1.89	5.98	5.45	**5.8**
Linear Regression	Test	41.92	21.34	9.54	7.32	**7.47**
Random Forest	Train	6.44	2.30	4.67	5.43	9.78
Random Forest	Test	12.14	10.28	10.12	10.10	10.13
Elastic Net	Train	14.12	12.69	11.50	10.32	10.75
Elastic Net	Test	10.32	10.37	10.10	10.05	10.08

both training and testing results. The first time-step, Step 1 (Offline), represents the base model, which is trained offline from scratch. Step 2 to Step 5 represent the adaptation of models with online data over time. The models should not have a decreasing performance due to the integration of online data, as a worse RMSE would indicate a less fitting model.

In general, the models performed quite differently at the same steps as for some models the RMSE is increasing whereat the RMSE is decreasing for others. The best performing machine learning model in our experiment is the linear regression model, although it is stagnating during the training process. The best performing neural network model is the Feed-Forward Multi-Task model with Natural Gradient Descent using Learning without Forgetting (FF M with NGD/LwF). It is one of the few models continually improving during the online training process.

The insertion of a new task at the third step enhances all of the neural network models, except the one model using the Elastic Weight Consolidation. The linear regression model and the random forest model are not able to deal with the new task as well as other models, which decreases their improvement although they still have good results. According to our experimental setting, a feed-forward multi-task model using LwF to minimize catastrophic forgetting fits best to our scenario and should be considered for further usage.

VI. CONCLUSION AND FUTURE WORK

Concluding, we present a framework which is able to improve offline learning models with online data. The need for such a framework increases as continuously more data is produced, especially in Industry 4.0 environments, and the training of a new model is either computationally expensive or not possible, e.g., old data may not be available anymore. The online learning setting in our framework enables the adaption of models using new data, either batchwise or elementwise, and additionally allows the inclusion of new tasks. The learning process of a model can be visualized to enable the evaluation of its development using different metrics. The framework enables various configuration possibilities regarding the training process and the models itself. The main focus of the model configuration is on the avoidance of catastrophic forgetting and, therefore, includes different state-of-the-art approaches for decreasing this problem. The amount of possible configurations make the framework flexible and adaptive regarding new regression problems The framework can easily be extended regarding supported models as currently only neural network and some selected machine learning models are available.

As the framework is still under development and we presented its current state in this paper, we will briefly outline further steps. Currently, we are working on the handling of censored online data, which are highly likely to cause negatively biased models, to reach a broader application area. Additionally, we want to enable more flexible neural network architectures and the parallel training of selected models resulting in the automatic selection of the best to continue working with.

ACKNOWLEDGMENT

The research reported in this paper has been supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry for Digital and Economic Aairs, and the Province of Upper Austria in the frame of the COMET center SCCH.

REFERENCES

[1] K. Zhou, T. Liu, and L. Zhou, “Industry 4.0: Towards future industrial opportunities and challenges,” in *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. IEEE, 2015, pp. 2147–2152.

[2] Y. Lu, “Industry 4.0: A survey on technologies, applications and open research issues,” *Journal of Industrial Information Integration*, vol. 6, pp. 1–10, 2017.

[3] A. Schütze, N. Helwig, and T. Schneider, “Sensors 4.0–smart sensors and measurement technology enable industry 4.0,” *Journal of Sensors and Sensor Systems*, vol. 7, no. 1, pp. 359–371, 2018.

[4] Kirkpatrick *et al.*, “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[5] D. Sahoo, Q. Pham, J. Lu, and S. C. Hoi, “Online deep learning: learning deep neural networks on the fly,” in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 2660–2666.

[6] L. C. Jain, M. Seera, C. P. Lim, and P. Balasubramaniam, “A review of online learning in supervised neural networks,” *Neural Computing and Applications*, vol. 25, no. 3-4, pp. 491–509, 2014.

[7] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” *Neural Networks*, vol. 113, pp. 54–71, 2019.

[8] A. Gepperth and B. Hammer, “Incremental learning algorithms and applications,” in *European symposium on artificial neural networks (esann)*, 2016.

[9] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, “End-to-end incremental learning,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 233–248.

[10] Z. Li and D. Hoiem, “Learning without forgetting,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.

[11] M. Rattray, D. Saad, and S.-i. Amari, “Natural gradient descent for online learning,” *Physical review letters*, vol. 81, no. 24, p. 5461, 1998.

[12] V. Nair and J. J. Clark, “An unsupervised, online learning framework for moving object detection,” in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[13] J. Xu, P.-N. Tan, J. Zhou, and L. Luo, “Online multi-task learning framework for ensemble forecasting,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 6, pp. 1268–1280, 2017.

[14] R. Caruana, “Multitask learning,” *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[15] B.-D. Shai *et al.*, “A theory of learning from different domains,” *Machine Learning*, vol. 79, pp. 151–175, 2010.

[16] K. Crammer, M. Kearns, and J. Wortman, “Learning from multiple sources,” in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 321–328.

[17] Y. Ganin *et al.*, “Domain-adversarial training of neural networks,” *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[18] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz, “Central moment discrepancy (CMD) for domain-invariant representation learning,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[19] M. Mermillod, A. Bugaiska, and P. Bonin, “The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects,” *Frontiers in psychology*, vol. 4, p. 504, 2013.

[20] G. Zhang, S. Sun, D. Duvenaud, and R. Grosse, “Noisy natural gradient as variational inference,” in *International Conference on Machine Learning*, 2018, pp. 5847–5856.

[21] A. Rusu *et al.*, “Progressive neural networks,” *arXiv preprint arXiv:1606.04671*, 2016.