Research Statement
Past and Current Research

George R. Obaido
PhD (Computer Science)
Wits University,
South Africa.
March 8, 2021

Introduction

I am a Visiting Researcher at the Formal Structures and Algorithms Lab at the University of Johannesburg and Computational Intelligence Lab at the University of the Witwatersrand in South Africa. My research interests lie in using Computational methods in solving many societal problems. These approaches range from Natural Language Processing (NLP), Formal Language and Automata (FLA), Machine Learning and Deep Learning, Computer Science Education (CSEd) and Mechanism Design for Social Good (MD4SG).

This research statement is organised as follows: the first section discusses my previous works, especially in the area of source-code plagiarism detection and using practice aids to students and non-technical experts to understand SQL queries. The second section discusses engagements with multiple interdisciplinary projects in Education, Transportation, Big Data, Open Data Sharing, and Responsible AI initiatives.

1 Previous Research

1.1 Computer Science Education: Source-code plagiarism

The problem of source-code plagiarism in programming assignments is a common problem in universities. This problem motivated me to taking a Master's study at Wits University. Further, I investigated this problem using the MOSS[1] plagiarism detection system [1], built and maintained by Prof Alex Aiken. In addition to this resource, I employed and extended graph-based approaches to MOSS to investigate the similarities in students' programs [2].

[1]Measure Of Software Similarity

These combined approaches initiated my curiosity in investigating many other practical CSEd problems. This Master's study further produced two conference publications [3, 4]. After completing my Master's studies, I indicated interests in undertaking a PhD research within the CSEd domain. This interest led to joining the Formal Structures, Algorithms and Industrial Applications Lab through Prof Abejide Ade-Ibijola, and the Computational Intelligence Lab under Prof Hima Vadapalli's supervision.

1.2 Computer Science Education: SQL comprehension

Structured Query Language (SQL) is popular with relational databases. Despite the simple and highly structured nature of SQL, end-users often find it difficult to comprehend written or new queries. During my PhD study, I investigated this problem, focusing on non-technical end-users in business sectors and students in academic institutions. This research presented several approaches to this problem. It covered a range of applying principles from Formal Language and Automata Theory (FLAT) techniques to SQL comprehension and synthesis.

This study yielded five publications with several software prototypes. These publications are as follows: first, a tool that uses regular expressions was designed to generate a narration from a query to aid the comprehension of SQL queries [5]. Second, a tool that describes the automatic generation of narrations from nested SQL queries using a context-free grammar (CFG) was built [6]. Third, another system that made use of visual specifications which represented SQL commands was used to build queries [7]. Fourth, a tool that uses a Jumping Finite Automaton, an abstract model for recognising natural language specifications of queries, was designed and helped translate into a SQL query [8]. Finally, a speech-based conversational system that takes speech inputs from a user was intended to interact with a hypothetical database [9]. Also, a journal article [10], which used a CFG to automatically synthesise large datasets, was published.

2 Current Research

My curiosity for knowledge has motivated me to join multiple interdisciplinary works. These areas of research are presented in the following sections.

2.1 Computer Science Education

Under-resourced languages are a significant challenge for statistical approaches to machine translation. Novice programmers of native origins struggle to understand programming since these languages are mostly written in English. These challenges have made it imperative for researchers and developers to create *culturally-agnostic* aids that would assist learners in understanding programming. Recently, I have worked on synthesising SQL queries from South African local language narratives [11]. This study aims to promote

South African local languages' inclusivity, mainly under-resourced within the Computer Sciences. I am interested in exploring computational techniques, such as Machine Learning and Deep Learning approaches to this problem.

Other notable areas of interests in CSEd are: how facial expression influences student understanding of a course concept and how to recommend which programming language is desirable for industry employers. Other areas of interests are using data sciences techniques, and gaming approaches to solve many CS problems – I have one contribution to this area [12].

2.2 Synthetic Data Generation

The ability to generate high-fidelity synthetic data is crucial when available (real) data is limited or where privacy and data protection standards allow only for limited use of the given data, e.g., in medical, agriculture, education and financial datasets. Since synthetic data can support AI / deep learning model development and software testing, I am particularly interested in contributing immensely in this area. Currently, I have applied the CFG approach to automatically synthesise hypothetical datasets that are similar to the Northwind database [10] – a sample database created by Microsoft to demonstrate features of some of their products, including SQL Server. I am particularly interested in exploring the neural approach, such as GANs[2] to synthesise many datasets for testing software applications.

2.3 Biomedical Imaging

Biomedical imaging techniques have significantly improved the health care of patients. Image-guided therapy has reduced the high risk of human errors with improved accuracy in disease detection and surgical procedures. Several computational models have been developed to assist medical practitioners to distinguish between benign and malignant cases. Together with a team of researchers in South Africa, we have reviewed some of these computational approaches and how they could be applied to solve medical-related problems, especially in breast cancer diagnosis [13]. I am interested in understanding in applying computational methods to save many lives in this area.

2.4 Transportation and Logistics

The outbreak of the coronavirus pandemic has had a significant impact on airlines worldwide – more than at any other time. With a team of researchers, we explored several airline passengers' experiences and their frustration about airlines providing vouchers instead of refunds.

This work analyses the experiences shared by airline passengers during the COVID-19 pandemic, especially concerning refund. The study (currently in press) contributes to the growing body of knowledge about the impact of the COVID-19 pandemic on various sectors of the world's economies. To our knowledge, this work is the first that examines how flight passengers experienced the refund process of airlines during the COVID-19 crisis.

[2]Generative Adversarial Networks

2.5 Open Data Sharing

I am actively involved in the open data sharing project with outstanding researchers. This project aims to critically examine and shed light on the complex data ecosystem in the Global South and do so in a manner that centres the stories, research, and practices of those from the region. We do so focusing on the African continent and the Middle East and North African (MENA) region. We examine questions such as "Who benefits from data sharing?" "Who are the omitted stakeholders in the data ecology?", and "How can data practices be carried out in a manner that pays back to communities where data is extracted from?". Our work is conducted by researchers and practitioners across the globe and done in collaboration with researchers, data scientists, non-profit organisations, and government organisations focused on improving our understanding of the data landscape and removing data inequities in their many forms.

This project has led to contributions in multiple venues, such as ACM FAccT[3] '21 [14], and preliminary versions in ACM CI[4]'20, PLSC[5] '20, Contested Data Academic Workshop at D&S [6] '20, ML4D[7]'19. The website for this project is available here. There are several ongoing projects in this area.

Bibliography

1. Kevin W Bowyer and Lawrence O Hall. Experience using "MOSS" to detect cheating on programming assignments. In *FIE'99 Frontiers in Education. 29th Annual Frontiers in Education Conference. Designing the Future of Science and Engineering Education. Conference Proceedings (IEEE Cat. No. 99CH37011*, volume 3, pages 13B3–18. IEEE, 1999.

2. George Rabeshi Obaido. Structural analysis of source code plagiarism using graphs, 2017. MSc thesis.

3. George Obaido, Pravesh Ranchod, and Richard Klein. Catching plagiarists: Detecting plagiarism in student source code assignment in a virtual learning environment. In *10th International Technology, Education and Development Conference*, pages 7369–7376. IATED, 2016.

4. George Obaido, Pravesh Ranchod, and Richard Klein. Constructing and analysing plagiarism in student programs using graphs. In *Proceedings of the Second International Conference on the Internet, Cyber Security and Information Systems*, pages 1–8. ICICIS, 2017.

5. Abejide Ade-Ibijola and George Obaido. S-NAR: generating narrations of SQL queries using regular expressions. In *Proceedings of the South African Institute of Computer Scientists and Information Technologists*, pages 1–8. ACM, 2017.

[3]Fairness, Accountability, and Transparency
[4]Conference on Collective Intelligence
[5]Privacy Law Scholars Conference
[6]Data and Society
[7]Machine Learning for Development Workshop

6. George Obaido, Abejide Ade-Ibijola, and Hima Vadapalli. Generating narrations of nested SQL queries using context-free grammars. In *2019 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6. IEEE, 2019.

7. George Obaido, Abejide Ade-Ibijola, and Hima Vadapalli. Generating SQL queries from visual specifications. In *Annual Conference of the Southern African Computer Lecturers' Association*, pages 315–330. Springer, 2018.

8. George Obaido, Abejide Ade-Ibijola, and Hima Vadapalli. Synthesis of SQL queries from narrations. In *2019 6th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 195–201. IEEE, 2019.

9. George Obaido, Abejide Ade-Ibijola, and Hima Vadapalli. TalkSQL: A tool for the synthesis of SQL queries from verbal specifications. In *International Multidisciplinary Information Technology and Engineering Conference (IMITEC) Vanderbijlpark, South Africa*, pages 1–9. IEEE, 2019.

10. Abejide Ade-Ibijola and George Obaido. XNorthwind: Grammar-driven synthesis of large datasets for DB applications. *International Journal of Computer Science, IAENG*, 46(4), 2019. IAENG.

11. George Obaido and Abejide Ade-Ibijola. Sythesis of SQL queries from south african local language narrations. *Advances in Science, Technology and Engineering Systems Journal (ASTESJ)*, 2020.

12. Solomon Sunday Oyelere, Nacir Bouali, Rogers Kaliisa, George Obaido, Abdullahi Abubakar Yunusa, and Ebunayo R Jimoh. Exploring the trends of educational virtual reality games: a systematic review of empirical studies. *Smart Learning Environments*, 7(1):1–22, 2020. Springer.

13. Kehinde Aruleba, George Obaido, Blessing Ogbuokiri, Adewale Oluwaseun Fadaka, Ashwil Klein, Tayo Alex Adekiya, and Raphael Taiwo Aruleba. Applications of computational methods in biomedical breast cancer imaging diagnostics: A review. *Journal of Imaging*, 6(10):105, 2020. Multidisciplinary Digital Publishing Institute (MDPI).

14. Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L Remy, and Swathi Sadagopan. Narratives and counternarratives on data sharing in africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 329–341. ACM, 2021.