BIG DATA TECHNOLOGY INFORMATION EXTRACTION AND FUSION FROM NON-HOMOGENOUS WEB DATA SOURCES

LEE QI ZIAN

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION TECHNOLOGY

2025

Faculty of Information and Communication Technology

BIG DATA TECHNOLOGY INFORMATION EXTRACTION AND FUSION FROM NON-HOMOGENOUS WEB DATA SOURCES

Lee Qi Zian

Master of Science in Information and Communication Technology

2025

BIG DATA TECHNOLOGY INFORMATION EXTRACTION AND FUSION FROM NON-HOMOGENOUS WEB DATA SOURCES

LEE QI ZIAN

A thesis submitted
in fulfillment of the requirements for the degree of
Master of Science in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2025

DECLARATION

I declare that this thesis entitled “Big Data Technology Information Extraction And Fusion From Non-homogenous Web Data Sources” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :.....

Name : LEE QI ZIAN

Date : 11/6/2025

APPROVAL

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of Master of Science in Information Technology and Communication Technology.

Signature :.....

Supervisor Name : DR NUR ZAREEN ZULKARNAIN

Date : 11/6/2025

DEDICATION

To my beloved parent Lee Choon Hoe and Yeoh Seok Peng who provide love and support are my greatest inspiration on completing this study. To my dearest supervisor, Dr Nur Zareen Zulkarnain and Ts. Dr Yogan Jaya Kumar for being helpful, responsible and always spending precious time to guide and motivate me. Finally, to all my friends who provided me support and motivation during the duration of this research.

ABSTRACT

Big data has played an ever-increasing role in various sectors of the economy. Despite the availability of big data technologies, many companies and organizations in Malaysia remain reluctant to adopt them. This study was conducted to develop a web extraction framework to extract data from the internet to assist adoption of big data technology. Web scrapping has been a popular method for collecting data from websites. This is because data on the internet is updated frequently thus making it a good source for getting accurate information. Analyzing data requires a large quantity of information to yield a good analysis result. However, the non-homogeneous nature of each website may cause the data from the different internet web sources to have different data making the quality of the data inconsistent. Previous study has propose the use of record linkage method to merge data from multiple website. The record linkage method proposed by previous study used deterministic technique to match data which match the string of matching variable to merge data. However, deterministic technique requires the matching variable to be an exact match to be able to match. Therefore, deterministic matching cannot take into account the dissimilarity such as spacing and different letter cases which can be common in web data due to it non-homogenous nature. This study will explore the use of fuzzy matching technique in matching web data. Fuzzy matching uses Levenshtein distance to calculate the similarity of string and a threshold will be used to decide how similar to trigger a match. This enables fuzzy matching to match string that are only partially match instead of exact match. This study will begin by conducting a systematic review to determine the challenge of big data adoption and what data to extract. This study will implement the Technology-Organization-Environment (TOE) framework to examine the challenges faced by Malaysian organizations with regards to big data adoption. After the systematic review, a web data extraction framework will be developed to extract data that can assist big data adoption. The extracted data will then be merged to enhance the quality of the data. A comparison is made between deterministic matching and fuzzy matching on the performance of merging web data. The finding from this comparison shows that fuzzy matching has a slightly better performance in merging web data. This is due to fuzzy matching can match the string of matching variable that has different spacing and letter cases. A survey case study carried out in this study also shows that the extracted data is very helpful in helping user while purchasing the required big data software on the software market.

PENGEKSTRAKAN MAKLUMAT TEKNOLOGI DATA RAYA DAN GABUNGAN DARIPADA SUMBER DATA WEB YANG TIDAK HOMOGEN

ABSTRAK

Data raya telah memainkan peranan yang semakin meningkat dalam pelbagai sektor ekonomi. Walaupun terdapat teknologi data raya, banyak syarikat dan organisasi di Malaysia masih enggan menerimanya. Kajian ini dijalankan untuk membangunkan rangka kerja pengekstrakan web untuk mengekstrak data daripada internet untuk membantu penggunaan teknologi data raya. Pengikisan web telah menjadi kaedah popular untuk mengumpul data daripada tapak web. Ini kerana data di internet kerap dikemas kini sekali gus menjadikannya sumber yang baik untuk mendapatkan maklumat yang tepat. Proses menganalisis data memerlukan kuantiti maklumat yang banyak untuk menghasilkan keputusan analisis yang baik. Walau bagaimanapun, sifat tidak homogen setiap laman web mungkin menyebabkan data daripada sumber web internet yang berbeza mempunyai data yang berbeza menjadikan kualiti data tidak konsisten. Kajian terdahulu telah mencadangkan penggunaan kaedah pautan rekod untuk menggabungkan data daripada beberapa laman web. Kaedah kaitan rekod yang dicadangkan oleh kajian lepas menggunakan teknik deterministik untuk memadankan data yang sepadan dengan rentetan pembolehubah padanan untuk menggabungkan data. Teknik deterministik memerlukan pembolehubah padanan menjadi padanan tepat untuk dapat dipadankan. Oleh itu, teknik padanan deterministik tidak boleh mengambil kira ketidaksamaan seperti jarak dan kes huruf yang berbeza yang boleh menjadi perkara biasa dalam data web kerana sifatnya tidak homogen. Kajian ini akan meneroka penggunaan teknik logic kabur dalam pemadanan data web. Logik kabur menggunakan jarak Levenshtein untuk mengira persamaan rentetan dan ambang akan digunakan untuk menentukan kesamaan untuk mencetuskan padanan. Ini membolehkan padanan kabur untuk memadankan rentetan yang hanya separa padanan dan bukannya padanan tepat. Kajian ini akan dimulakan dengan menjalankan semakan sistematik untuk menentukan cabaran penggunaan data raya dan data yang perlu diekstrak. Kajian ini akan melaksanakan rangka kerja Technology-Organization-Environment (TOE) untuk mengkaji cabaran yang dihadapi oleh organisasi Malaysia berkaitan penggunaan data raya. Selepas semakan sistematik, rangka kerja pengekstrakan data web akan dibangunkan untuk mengekstrak data yang boleh membantu penerimaan data besar. Data yang diekstrak kemudiannya akan digabungkan untuk meningkatkan kualiti data. Perbandingan dibuat antara padanan deterministik dan logik kabur pada prestasi penggabungan data web. Dapatan daripada perbandingan ini menunjukkan bahawa padanan kabur mempunyai prestasi yang lebih baik sedikit dalam penggabungan data web. Ini disebabkan padanan kabur boleh memadankan rentetan pembolehubah padanan yang mempunyai jarak dan huruf huruf yang berbeza. Walau bagaimanapun, kajian ini juga menunjukkan bahawa data yang diekstrak sangat membantu dalam membantu pengguna semasa membeli perisian data raya yang diperlukan di pasaran perisian.

ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to express my sincere gratitude to my supervisors Dr Nur Zareen Zulkarnain and Ts. Dr Yogan Jaya Kumar from the Faculty of Information and Communication Technology Universiti Teknikal Malaysia Melaka (UTeM) for their essential supervision, support and encouragement towards the completion of this study and thesis.

I would also like to extend my sincere appreciation to the staff of both Universiti Teknikal Malaysia Melaka and Faculty of Information and Communication Technology for assisting me with the administration during my study. I also thank Prof. Dr. Azah Kamilah Binti Draman @ Muda and Centre for Research and Innovation Management (CRIM) for providing financial assistance as part of research assistance allowance during my study.

Finally I would like to extend my gratitude to my family for their unwavering support, love and encouragement, prayers and motivation throughout the duration of this study. Their constant encouragement and support helped me to stay focus and determined to achieve my goals. Without their support, this thesis would not be possible. I am forever grateful to all the person who support me high and lows during this research journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BD	-	Big Data
BDA	-	Big Data Adoption
TOE	-	Technology-Organization-Environment

LIST OF APPENDICES

LIST OF PUBLICATIONS

The followings are the list of publications related to the work on this thesis:

Qi Zian, Lee and Zulkarnain, Nur Zareen and Jaya Kumar, Yogan, 2024. Challenges in big data adoption for Malaysian organizations: a review. *Indonesian Journal of Electrical Engineering and Computer Science*. 33. 507-517. 10.11591/ijeecs.v33.i1.pp. 507-517. 2024. (SCOPUS indexed, Q2 (2024)

Qi Zian, Lee and Jaya Kumar, Yogan and Zulkarnain, Nur Zareen, 2021. Tools for Big Data and Analytics. *3rd International Conference on Intelligent and Interactive Computing*. pp. 76-79.

CHAPTER 1

INTRODUCTION

1.1 Background

Big data is a technology that had been rapid gaining popularity from organizations looking to seize advantage opportunities. There are three main key characteristic of big data evolution, Volume, Variety, and Velocity (Kaur et al., 2020). Big data can be qualified as any data that are difficult be processed by a traditional data pipeline. Big data has huge application opportunities across all sectors of industries. By adopting big data, organization can leverage their power of its unique characteristic and function to improve the performance of sales and marketing department.

Big data has great potential in research and development due to its unique ability to process data from multiple data sources such as documents, records, protocol, video and images. However, like every technology an organization will always need to invest a significant amount of resources to use it to its fullest potential. The main problem faced by organizations while trying to adopt big data technologies are technology readiness. Despite the simple two words of technology readiness, it has a huge impact on the organization decision on how to move forward. Technologies are an ever-evolving thing and choosing a suitable technology is crucial for the long-term planning of an organization. Decision making has always been in the core of human interaction ranging a simple decision like choosing what to eat during dinner to crucial decision making that can possibly reshape history (Bossaerts and Murawski, 2017). Information has always been a crucial asset in human life since the invention of writing which is also the cornerstone of modern world. By recording

information with words preserves the idea for a long time. They display the vibrations of the human soul in written form, thus the knowledge can be pass down to endless generations. With the invention of computers, human found out that writing can be digitalize and can be store in virtually endless quantities.

Big data has certainly been a tremendous help in the decision making with has largely affect the all the economy sectors (Berardino and Vona, 2023). However, there is no doubt that the financial sector will have the most obvious benefits from big data tools as large volume of data can be process in a much cheaper and efficient ways (Hasan et al., 2020). This can also enable a company to have competitive advantages in the marketing sector such as new product, services, new business model, customer satisfaction and customer loyalties (Ramadan et al., 2020). Moreover, big data can ensure all the financial information of a company to be integrated and explored unlike the traditional model that cause all the information to be kept in silos of each department in the company that can cause crucial information to be missed out (Almeida, 2017).

1.2 Problem Statement

Nowadays there are a lot of tools and software with various capabilities such as Python, Matlab, Tableau, Hadoop, and MongoDB. These tools can provide service in helping organizations and companies tackling big data problems. Although there are a lot of big data tools some companies are still reluctant to use it. In a study conduct in 2018 (Saleh et al., 2018), one of the reason that make Malaysian company especially SME turning away from big data tools is that it can take up a lot of time and resources to use it.

The technologies chosen to tackle big data can usually affect the recruitment of expertise and the training of employee in new technical skills. Choosing the wrong

technologies to use can also affect the quality of the gathered data. According to a study conducted by Gartner in 2017, researchers reveal that poor data quality can cause $15 million annually and will also undermine their digital initiative which may lead to poor customer trust (Moore, 2018). Despite all the issues, many organizations still underestimate the importance of choosing which tools to be deployed to tackle big data. Business may usually opt for a tool that can be deploy faster, but the tools may be less effective. This issue can also sow distrust between the data execution team and management.

Moreover, the operational cost of big data tools is also a key factor when an organization is deciding on which tools to invest in. A cost of a big data tools ranged from open source (zero cost) to paid license software with good after sales services. Different tools will require different knowledge to operate it such as the programming language and the architecture of the tools. When an organization decide to adopt a new big data tool, they might need to retrain their employee which will result in more cost on the training. Other than operational cost, data sensitivity which is how willingly a company will share their data within department. Data sensitivity has significant contribute to a factor that makes company reluctant to adopt big data as one company might has different data sensitivity to other company (Kaisler et al., 2019).

Even if a company had already decided what type of big data tools to adopt, they will need to find the suitable big data tools on the e-market. The e-market of big data tools tends to be a software review website where software developer or publisher will publish their software on the website for user to review it. The typical review page of a software will usually have the software features, approximate prices, user rating and the URL link to the software website. These type of website makes finding software easier for end user. However, there a lot of software tools review websites and all the information on the review

page are varied. For example, a software from review website A may have less listed features compared to another review website B. This can be due to the information in review website A not being updated. One method which can be used to address this issue is by using deterministic matching technique. This technique select a feature in the data as matching variable to match the data thus merging the data sources.

Although previous studies has explored and proven that these implication can be solved by using deterministic matching technique in data merging to match data by comparing the string of the matching variable for match however, deterministic matching can only match data that has exactly same matching variable (Namazi-Rad et al., 2017; Tuoto et al., 2018; Suman et al., 2020; d'Ovidio et al., 2021; Shook-Sa et al., 2022; Jamali-Phiri et al., 2023). The nature of deterministic matching can cause implication in matching data extracted from internet as the data from different sources do not have the exact same matching variables (Shook-Sa et al., 2022).

As a conclusion, there are two major concerns regarding assisting the adoption of big data technology in Malaysia. First is to identify the key challenges concerning Malaysian organizations to use big data technologies. Next is to investigate how relevent information from software tools review websites can be merged to provide useful insights in aiding the adoption of big data technologies. For the latter, the main challenges in merging data from web data source is due to the nature of non-homogenous website that does not have the exact same data. Here we put forward the research questions that will be further investigated in this study.

- What are the key challenges concerning Malaysian organizations to use big data technologies?

- How can we match data regarding big data technologies extracted from different sources for merging without requiring the data to be exactly same for matching?

1.3 Research Objective

To solve and mitigate the problem stated in the problem statement, a study is conducted to develop a framework to assist organizations and companies to adopt big data technologies. The first aim is to find out what are the challenges on big data technologies adoption in Malaysia. This can help us identify what are the helpful information to be extracted from the internet. The second aim is to explore data matching techniques to investigate the optimal techniques in merging data extracted from different data sources. Therefore the main hypothesis of this study is that an alternative data matching techniques such as fuzzy matching are more can match web extracted data better than deterministic matching as it can match data with partially same matching variable. Thus, to achieve the above-mentioned aim, the following objectives have been set:

a) To identify the challenges concerning big data technology adoption among organization and industries in Malaysia using systematic review.

b) To propose a web information extraction framework with an optimal approach to extract and combine big data tools information from non-homogenous data sources to assist big data adoption.

To investigate the concerns regarding the use of big data technologies within Malaysian organizations, a systematic review is conducted to identify the criteria and challenges for big data technologies adoption. The information gained from this study is crucial as it will give a clear picture on what type of information are commonly seeked by the organization in deciding to use big data technologies.

Next, a web extraction framework will be developed to extract the features concerning of big data tools from the internet based on the information identified from the systematic review. The data extracted from the data sources will be merged. This study will explore an alternative matching technique known as fuzzy matching. The performance of data merging will be compared between deterministic matching and fuzzy matching techniques.

1.4 Scope of Research

The scope of this research are as follows:

- Conducting a systematic review of existing literature to identify the criteria and challenges for the adoption of big data technologies within Malaysian organizations.
- Analyzing the identified criteria and challenges to understand common trends and issues in big data technology adoption.
- Developing a web extraction framework to collect information on features of big data tools from the internet, based on the findings of the systematic review.
- Merging the extracted data from different sources and exploring the use of fuzzy matching as an alternative technique for data merging.
- Comparing the performance of data merging between deterministic matching and fuzzy matching techniques to assess the effectiveness of each method.