Faculty of Information and Communication Technology

DUPLICATES DETECTION APPROACH WITHIN INCOMPLETE DATA SETS USING DYNAMIC SORTING KEY AND HOT DECK COMPENSATION METHOD

Abdulrazzak Ali Mohamed Abdulrahim

Doctor of Philosophy

2022

DUPLICATES DETECTION APPROACH WITHIN INCOMPLETE DATA SETS USING DYNAMIC SORTING KEY AND HOT DECK COMPENSATION METHOD

ABDULRAZZAK ALI MOHAMED ABDULRAHIM

A thesis submitted
in fulfilment of the requirements for the degree of Doctor of Philosophy

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2022

DECLARATION

I declare that this thesis entitle "Duplicates Detection Within Incomplete Data Sets Using Dynamic Sorting Key and Hot Deck compensation Method" 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 : Abdulrazzak Ali Mohamed Abdulrahim

Date : 07/11/2022

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 Doctor of Philosophy.

Signature : .....

Supervisor Name : Associate Professor Dr. Nurul Akmar Emran

Date : 07/11/2022

DEDICATION

To my beloved mother and family

ABSTRACT

Duplicate record is a common problem within data sets, especially in huge-volume databases. The accuracy of duplicate detection determines the efficiency of the duplicate removal process. However, duplicate detection has become more challenging due to the presence of missing values within the records where during the clustering and matching process, missing values can cause records deemed similar to be inserted into the wrong group, hence, leading to undetected duplicates. Keeping a database free of duplicates is crucial for most use-cases, as their existence causes false negatives and false positives when matching queries against it. These two data quality issues have negative implications for tasks, such as in the medical field, where the patient may get drugs overdosage, which could, unfortunately, cause loss of life, or parcel delivery, where a parcel can get delivered to the wrong address. While research in duplicate detection is well-established and covers different aspects of both efficiency and effectiveness, our work in this thesis focuses on both. We propose novel method to improve preprocessing task to overcome the challenge posed by the presence of missing values on the efficiency of duplicates detection before duplicate detection takes place and apply the latter in datasets even when prior labeling is not available. In this thesis, duplicate detection improvement is proposed to deal with the presence of missing values within a data set through Duplicate Detection within the Incomplete Data set (DDID) method. DDID is based on a set of procedures to address the problem of missing data, which is to adopt a generic approach based on high-rank attributes (high uniqueness, low missing values), followed by compensating the missing values in high-rank attributes using the Hot Deck compensation method. Dynamic sort keys and matching strings are created from the high-rank attributes in certain lengths. These procedures that were adopted in DDID aimed to validate the expected results in successive stages of detection and achieve a high matching rate of duplicate records despite the presence of missing values through a specific detecting mechanism. The experiments included the use of four benchmark data sets (restaurant, CDDB, MusicBrainz (A), MusicBrainz (B)) to detect duplicates. The missing values were hypothetically added to the key attributes with 4% for the Restaurant data set and 1.5% for the CDDB data set, using an arbitrary pattern to simulate both complete and incomplete data sets. DuDe toolkit was used to detect duplicates as a benchmark to make a relative comparison. Duplicates detection measures have been used to evaluate DDID in terms of accuracy and use performance improvement (PI) and statistical analysis to evaluate DDID in terms of elapsed time. The results of the experiments showed that the procedures adopted in the proposed method DDID achieved a significant improvement in the accuracy of detecting duplicates compared to DuDe as it reached in the first implementation stage, 18% with the Restaurant data set while 16% with the CDDB data set; and its reached 19% and 4% for both MusicBrainz(A) and MusicBrainz(B) respectively, as compared to DuDe.

Similarly, DDID achieved significant improvement in the accuracy of detecting duplicates as compared to DuDe in the second implementation stage, reaching 24%, 18%, 30%, and 3% for Restaurant, CDDB, MusicBrainz(A), and MusicBrainz(B), data sets respectively. The analysis proved that even though the data sets were incomplete, DDID was able to offer better accuracy and faster duplicate detection as compared to DuDe. The adopted procedures also had a positive effect on limiting the defect of window size in the sorted neighbourhood method, as it maintained the stability of the accuracy of detection of duplicates, in addition to improving the performance of the tested blocking methods within this study. The results of this thesis not only contribute to expanding the body of knowledge in data management specifically in the area of data quality, where the focus is given to the problem of how to detect the presence of duplicates within data sets that are incomplete. But it can also contribute to the problem of industry-scale duplicate detection.

PENDEKATAN PENGESANAN PENDUA DALAM SET DATA TIDAK LENGKAP MENGGUNAKAN KUNCI PENGISIHAN DINAMIK DAN KAEDAH PAMPASAN DEK PANAS

ABSTRAK

Kelewahan data adalah masalah biasa dalam set data terutamanya dalam pangkalan data bersaiz besar. Ketepatan pengesanan pendua menentukan kecekapan proses penyingkiran pendua. Walau bagaimanapun, pengesanan pendua telah menjadi lebih mencabar kerana kehadiran nilai yang hilang dalam rekod di mana semasa proses pengelompokan dan pemadanan, nilai yang hilang boleh menyebabkan rekod yang dianggap serupa dimasukkan ke dalam kumpulan yang salah, justeru, membawa kepada pendua tidak dapat dikesan. Mengekalkan pangkalan data bebas daripada pendua adalah penting untuk kebanyakan kes penggunaan, kerana kewujudannya menyebabkan negatif palsu dan positif palsu apabila memadankan pertanyaan dengannya. Kedua-dua isu kualiti data ini mempunyai implikasi negatif untuk tugas, seperti dalam bidang perubatan, di mana pesakit mungkin mendapat dos berlebihan ubat, yang malangnya boleh menyebabkan kehilangan nyawa atau penghantaran bungkusan, di mana bungkusan boleh dihantar ke alamat yang salah . Walaupun penyelidikan dalam pengesanan pendua sudah mantap dan merangkumi aspek berbeza bagi kedua-dua kecekapan dan keberkesanan, kerja kami dalam tesis ini memfokuskan pada kedua-duanya. Kami mencadangkan kaedah baru untuk meningkatkan tugas prapemprosesan untuk mengatasi cabaran yang ditimbulkan oleh kehadiran nilai yang hilang pada kecekapan pengesanan pendua sebelum pengesanan pendua berlaku dan menggunakan yang terakhir dalam set data walaupun pelabelan sebelumnya tidak tersedia. Dalam tesis ini, penambahbaikan pengesanan pendua dicadangkan untuk menangani kehadiran nilai yang hilang dalam set data melalui Pengesanan Pendua dalam kaedah Set Data Tidak Lengkap (DDID). DDID adalah berdasarkan satu set prosedur untuk menangani masalah kehilangan data, iaitu menggunakan pendekatan generik berdasarkan atribut peringkat tinggi (keunikan tinggi, nilai hilang rendah), diikuti dengan mengimbangi nilai yang hilang dalam peringkat tinggi atribut menggunakan kaedah pampasan Hot Deck. Kekunci isihan dinamik dan rentetan padanan dicipta daripada atribut peringkat tinggi dalam panjang tertentu. Prosedur ini yang diterima pakai dalam DDID bertujuan untuk mengesahkan keputusan yang dijangkakan dalam peringkat pengesanan berturut-turut dan mencapai kadar pemadanan rekod pendua yang tinggi walaupun terdapat nilai yang hilang melalui mekanisme pengesanan khusus. Eksperimen tersebut termasuk penggunaan empat set data penanda aras (restoran, CDDB, MusicBrainz (A), MusicBrainz (B)) untuk mengesan pendua. Nilai yang hilang telah ditambahkan secara hipotesis pada atribut utama dengan 4% untuk set data Restoran dan 1.5% untuk set data CDDB, menggunakan corak arbitrari untuk mensimulasikan kedua-dua set data yang lengkap dan tidak lengkap. Kit alat DuDe digunakan untuk mengesan pendua sebagai penanda aras untuk membuat perbandingan relatif. Langkah pengesanan pendua telah digunakan untuk menilai DDID dari segi ketepatan dan menggunakan peningkatan prestasi (PI) dan analisis statistik untuk menilai DDID dari segi masa berlalu. Keputusan eksperimen menunjukkan bahawa prosedur yang diterima pakai dalam kaedah yang dicadangkan DDID mencapai peningkatan yang ketara dalam ketepatan pengesanan pendua berbanding DuDe kerana ia mencapai pada

peringkat pelaksanaan pertama, 18% dengan set data Restoran manakala 16% dengan set data CDDB; dan mencapai 19% dan 4% masing-masing untuk kedua-dua MusicBrainz(A) dan MusicBrainz(B), berbanding DuDe. Begitu juga, DDID mencapai peningkatan ketara dalam ketepatan pengesanan pendua berbanding DuDe pada peringkat pelaksanaan kedua, mencapai 24%, 18%, 30% dan 3% untuk Restoran, CDDB, MusicBrainz(A), dan MusicBrainz(B), set data masing-masing. Analisis membuktikan bahawa walaupun set data tidak lengkap, DDID mampu menawarkan ketepatan yang lebih baik dan pengesanan pendua yang lebih pantas berbanding DuDe. Prosedur yang diterima pakai juga mempunyai kesan positif dalam mengehadkan kecacatan saiz tetingkap dalam kaedah kejiranan yang disusun, kerana ia mengekalkan kestabilan ketepatan pengesanan pendua, di samping meningkatkan prestasi kaedah menyekat yang diuji dalam kajian ini. Hasil tesis ini bukan sahaja menyumbang kepada pengembangan badan pengetahuan dalam pengurusan data khususnya dalam bidang kualiti data, di mana tumpuan diberikan kepada masalah bagaimana untuk mengesan kehadiran pendua dalam set data yang tidak lengkap. Tetapi ia juga boleh menyumbang kepada masalah pengesanan pendua skala industri.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the almighty ALLAH swt, for giving me the strength, courage, health, creativity and ability to complete this thesis successfully, I would like to gratefully acknowledge my deepest gratitude to my supervisor, Professor Madya Dr. Nurul Akmar Binti Emran for believing in me, and her outstanding guidance and valuable advise during my study. My sincere thanks also for my second supervisor, Dr. Siti Azirah Binti Asmai.

I would like to express my thank to Hasso Plattner Institute (HPI) and Humburg University for making it possible to use data sets and the Duplicate Detection Tool (DuDe) in this research.

Last but not least, I would like to thank my family for their love and support, especially my mother and my wife for their sacrifice and patience of parting and their continuous prayers for me to complete this thesis, and also my beloved son Mohammed Abdulrazzak Ali in whom I see the future. I particularly want to thank my brothers (Salah, Shaafal, Mukhtar, Abdo, Kamal and Fahd) and my sisters (Samira, Malukah, Amira and Suhair) for their unconditional moral supports.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DDID	-	Duplicates Detection Within Incomplete Data Sets
DuDe	-	The Duplicate Detection Toolkit
SNM	-	Sorted Neighborhood Method
UF	-	Uniqueness Function
CF	-	Completeness Function
LD	-	Levenshtein Distance
TP	-	True Positives
FP	-	False Positives
TN	-	True Negatives
FN	-	False Negatives
PPS	-	Pusat Pengajian Siswazah
SPSS	-	Statistical Package for the Social Sciences
FTMK	-	Fakulti Teknologi Maklumat dan Komunikasi

LIST OF PUBLICATIONS

1. Ali, A., Emran, N.A. and Asmai, S.A., 2021. Missing Values Compensation in Duplicates Detection Using Hot Deck Method. *Journal of Big Data*, 8(1), pp. 1–19.

2. Ali, A., Emran, N.A., Asmai, S.A. and Ismail, A.R., 2019. An Assessment of Open Data Sets Completeness. *International Journal of Advanced Computer Science and Applications*, 10(6), pp.557-562.

3. Ali, A., Emran, N.A., Asmai, S.A. and Thabet, A., 2018. Duplicates Detection Within Incomplete Data Sets Using Blocking and Dynamic Sorting Key Methods. *International Journal of Advanced Computer Science and Applications*, 9(9), pp.629–636.

4. Thabet, A., Emran, N.A, Muda, A. K. and Ali, A., 2022. Missing Values Imputation in Arabic Datasets Using Enhanced Robust Association Rules (ERAR). *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, 28(2), pp.1067–1075.

5. Ali, A., Emran, N.A., Asmai, S.A. and Thabet, A.,. Improving the Efficiency of Clustering Algorithm for Duplicates Detection. *Indonesian Journal of Electrical Engineering and Computer Science* (Accepted).

CHAPTER 1

INTRODUCTION

This thesis proposes an improved approach to duplicate detection within incomplete data set. The proposed method aims to improve the factors of accuracy and elapsed time in the duplicates detection within incomplete data sets. This chapter is organized as follows: The problem of duplicates within the incomplete data set and its impact on data quality is presented in Section 1.1. In Section 1.2 research motivation is covered. Research objectives and its mapping with the research questions is in Section 1.3. Section 1.4 presents the main challenges in duplicates detection within incomplete data sets. Section 1.5 presents the contributions of the research in expanding data quality body of knowledge. In Section 1.6 the chapter concludes with a summary of the expected results in the chapters of the thesis.

1.1 Overview

Applications of business and projects rely mostly on databases for storage purposes and information processing for determining results. However, the database usage over the past years faced specific constraints and most databases were configured only for specific applications and purposes usage. In short, the configuration of applications and databases are tailored for attaining specific results for end-users. However, at present, many of these applications produce outputs that are not inevitable but may be incomplete, inaccurate or somewhat ambiguous (Panse, 2015). The databases suffer from quality problems which are lacking inconsistency in which the kind of data consistency often relates to the real-world entities unique representation such as individuals' data and merchandises stored in databases. An inconsistent database refers to a scenario where an entity is represented several times in the database or when the data itself is represented differently in different databases because of the schematic heterogeneity (Wang and Zhang, 1996; Chen, Zobel and Verspoor, 2017b). Thus, the database becomes 'dirty' because of the presence of duplicates (Elmagarmid, Ipeirotis and Verykios, 2007; Naumann and Herschel, 2010; Christen, 2012). Definitions

of duplicates or redundancy rely on the context. In standard databases, duplicate occurs if a unique entity is represented several times (Songchun and Hideto, 1993; Tamilselvi and Gifta, 2011; Chen, Zobel and Verspoor, 2017a; Panse et al., 2021) or the objects mirror the identical real-world object but have a variety of representations in the database (Elmagarmid, Ipeirotis and Verykios, 2007; Baumgartner et al., 2009; Barcelos, Mendoza and Moreira, 2021).

Data duplication can lead to a disaster. In the medical field, for example, non-detection of duplicates can lead to an increase in the quantity of drugs (overdose) (Di Rico et al., 2018). Consequently, the prescription of wrong medications may endanger the patient's health due to the unintended interdependencies between the administered medications. Therefore, data sets within databases must be cleaned from duplicate representations of entities to ensure data consistency. (Elmagarmid, Ipeirotis and Verykios, 2007; Naumann and Herschel, 2010; Christen, 2012; Ehsani-Moghaddam, Martin and Queenan, 2021). One of contributing causes of duplicates is the integration of data generated from multiple sources (Lenzerini, 2002; Brazhnik and Jones, 2007; Doan, Halevy and Ives, 2012; Picado et al., 2020). The absence of a universal identifier or the damage of a value of this identifier may lead to data duplication. Duplicate detection requires linking of several data sources that help to capture aspects of the same real-world entities (van Gennip et al., 2018). The integrated data contain a ratio of duplicate data which is between $(1\% - 5\%)$ (Kelkar, Manwade and Prof, 2012). Addressing duplicated data is more important to avoid false inflation of the database which makes data retrieval becomes costly and difficult. Besides, duplicates detection is used to improve the quality of schemas between different databases (Panse, 2015). Duplicate detection is a complex task especially in the presence of noise such as misspellings, abbreviations, and missing values. This situation causes similar records that represent the same physical entity, for example, having the same staff represented multiple times in the company's database, with several different personnel numbers. This situation is also known as semantic duplicates (Nguena, Ophélie and Richeline, 2017; Ansari and Sharma, 2020).

In this thesis, inexact duplicates issue is highlighted, in which records refer to the same physical entity while not being syntactically equivalent (Tamilselvi and Gifta, 2011; Anitha et al., 2012; Vasiliev et al., 2020). Inexact duplicates occur due to the presence of missing values within a data set. The element that has a null value does not necessarily represent the element that does not exist at all. For example, if a book contains a subtitle field,

but the value is null, then the book contains the subtitle element but the exact value of the field is unknown. On the other hand, stating that the subtitle is missing may simply indicate that the book has no subtitle. Finally, there must be strategies to develop similar measures to help detect the various possible semantics of missing elements and null values that may cause a duplicate (Naumann and Herschel, 2010). Duplicate detection is a process known to traditional databases (Elmagarmid, Ipeirotis and Verykios, 2007; Naumann and Herschel, 2010; Christen, 2012; Aleshin-Guendel and Sadinle, 2022), but little attention has been given for duplication detection within incomplete data sets. Some aspects of duplication detection within incomplete data set are very similar to other databases, however, the detection of duplicates within incomplete data sets poses some new challenges. In this thesis, the effect of missing values in detection of duplicates was analyzed and a method for the detection of duplicates that focuses on incomplete data sets was proposed. In the next section, the concept of incomplete data sets is presented.

1.1.1 Incomplete Data Sets

Incomplete data sets have become almost ubiquitous in various application domains. It has been reported that, the more data are accumulated, and the more tools for integrating and exchanging data become available, the more instances of incompleteness are obtained (Libkin, 2014). A data set with at least one incomplete datum is referred to as an incomplete data set, otherwise, it is called complete data set (Umathe and Chaudhary, 2015). Incomplete data create uncertainties during data analysis, which must be managed during data analysis. Dealing with incomplete data sets is a challenge in order to record high-quality data (Stiglic et al., 2017). Climate and image, sensors and medical data sets are common examples of incomplete data sets. Issue of incompleteness in these data sets may be caused by several factors such as certain measurements reflection might be absent at the time, or the information might be missing because of failure of partial system, sensor node malfunction, certain areas in systematics policies which intentionally skip some values or it might simply be a result of users' privacy concerns. If all of the attributes have few missing significant fraction of the entries, any kind of reasonable extrapolation on the original data is hard to perform (Aggarwal and Parthasarathy, 2001; Jaseena and David, 2014). Thus, incomplete data management, such as the merging of data from different sources for various reasons

brings a new challenge of data duplication (Chen, Zobel and Verspoor, 2017b). The next section presents the impact of duplicates on data quality.

1.1.2 Duplicates as Data Quality Problem

Today, big data problem can be a result of organizations failures to process or analyse data produced by several sources. These organizations have access to a massive information, but they are unable to get the value out of it (Zezula, 2015). The main problem concerning data quality is the data are often 'dirty' at data sources (Kim et al., 2003). Dirty data include inaccurate data, incomplete data, the presence of duplicates, and non-standard representation of data. Dirty data leads to unreliable results for analysis.

Data quality is defined as "fitness for use" (Wang and Strong, 1996; Carlo and Scannapieca, 2006; Goodchild, Wenzhong and Fisher, 2002) and is also defined as "the distance between the data views presented by an information system and the same data in the real world" (Orr, 1998). Such a definition is viewed as an "operational definition", although defining data quality based on comparisons with the real world is an extremely difficult task (Bertolazzi and Scannapieco, 2001) Accordingly, any false decision should be ignored. Data sets must be preprocessed to produce a complete and clean dataset before starting any integration process. Even though a standard set of dimensions for data quality is not available, researchers commonly agree with data quality attributes or dimensions namely accuracy, completeness, consistency, and currency. So for each data quality problem, there is a specific data quality rule that it targets like redundancy (as a duplicate instance), illegal values, functional dependency (Taleb, Dssouli and Serhani, 2015; Sadiq et al., 2018). The systems which rely on these dimensions and rules are of high-quality (Scannapieco, Missier and Batini, 2005). Data quality has become a common challenge for organizations where they struggle with inconsistency, loss and data duplication (Huang et al., 2017).

Moreover, many statistical surveys have shown that data conflicts arise because of duplicate records (Elmagarmid, Ipeirotis and Verykios, 2007). A practical solution for this problem namely duplicate detection (or record linkage or data matching) is proposed to produce a unique and consistent view of the data record. However, it is later found that the techniques are facing a larger set of data problems such as incompleteness, inaccuracy and inconsistency. Therefore, several researchers are dealing with data quality problems.