AN APPROXIMATE FUNCTIONAL DEPENDENCIES (AFD) BASED APPROACH TO IMPROVE SKYLINE QUERIES COMPUTATION AND MISSING VALUES ESTIMATION OF SKYLINES ON CROWDSOURCED-ENABLED INCOMPLETE DATABASE

BY

MARWA BEHJAT SWIDAN

A thesis submitted in fulfilment of the requirement for the degree of Doctor of Philosophy in Computer Science

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

APRIL 2021

ABSTRACT

Data incompleteness becomes a frequent phenomenon in contemporary non-trivial database applications such as web autonomous databases, incomplete databases, big data and crowd-sourced mobile databases. Processing queries over these incomplete databases impose several challenges that negatively influence processing the queries. Most importantly, the query results derived from incomplete databases are also incomplete as certain values of the query result are not present. Result incompleteness may lead to misguiding the user in multi-criteria decision-making and decision support systems. Skyline queries are one of the most prominent queries applied over these recommendation and decision-making systems. Most recently, several studies have suggested exploiting the crowd-sourced databases in order to estimate the missing values by generating plausible substitute values using the crowd resources. Crowd-sourced databases have proved to be a powerful solution to perform user-given tasks by integrating human intelligence and experience to process the tasks. However, task processing using crowd-sourced platform incurs additional monetary cost and increases the time latency. Also, it is not always possible to produce a satisfactory result according to the user's preferences. Thus, an efficient and cost-effective approach for estimating the missing values of the skylines on crowd-sourced enabled incomplete databases is necessary which is achieved by exploiting the available data and the implicit relationships in the database before referring to the crowd is needed. This thesis proposes a new approach for estimating the missing values of the skylines over incomplete databases. The approach attempts to eliminate the unwanted tuples from the initial incomplete database using data filtration to simplify the value estimation process. Furthermore, the approach utilizes the remaining data and exploits the implicit relationships between the attributes to impute the missing values of the skylines. The approach employs the principle of mining attribute correlations to generate a set of approximate functional dependencies (AFDs) that assist in generating the estimated values. Also, the proposed approach aims at reducing the number of values to be estimated using the crowd when local estimation is inappropriate. Certain factors that influence the data processing such as monetary cost, time latency and accuracy are considered when working on the crowd-sourced platform to estimate the missing values of the skylines. Intensive experiments on both synthetic and real datasets have been accomplished. The experimental results have proven that the proposed approach for estimating the missing values of the skylines over crowd-sourced enabled incomplete databases is scalable and outperforms the other existing approaches. The proposed approach simplifies the process of missing value estimation for the skylines with a total reduction of up to 80% in the number of the values to be considered for the estimation in the initial incomplete database. Furthermore, the experimental results have also shown that the proposed solution has achieved the lowest relative error rate between the real missing and the estimated values in comparison with the other recent approach. Most importantly, our proposed strategy is capable of estimating up to 40% of the total missing values with accuracy up to 90% by exploiting the available data in the initial incomplete database. Lastly, the results of the experiments have also demonstrated that our approach has significantly decreased the monetary cost and the time latency involved when estimating the missing values of the skylines using crowd-sourced databases.

خلاصة البحث

أصبح عدم اكتمال البيانات ظاهرة متكررة في عدد لا يستهان به من تطبيقات قواعد البيانات المعاصرة ؛ كقواعد البيانات الغير كاملة، والبيانات الضخمة وقواعد البيانات الحشود المتنقلة. ان نتائج الاستعلام المستمدة من قاعدة البيانات الغير كاملة هي أيضا غير كاملة، اي انه لا يمكن الحصول على قيم مؤكدة لنتائج الاستعلام؛ وقد يؤدي هذا إلى تضليل المستخدم خاصة فيما يتعلق بنظم اتخاذ القرار و نظم دعم القرار. وفي الآونة الأخيرة اقترحت مجموعة من الدراسات استخدام قواعد بيانات الحشد الجماعي (crowd-sourcing databases) لتقدير القيم المفقودة في قاعدة البيانات عن طريق توليد قيم مقبولة باستخدام موارد الحشد. وقد اثبتت الدراسات أن قواعد البيانات ذات الحشد الجماعي تمثل حلاً قوياً من خلال دمج ذكاء ،وقدرات ،وخبرات البشر في معالجة المهام. ومع ذلك فإن معالجة المهام اعتماداً على الحشود الجماعية تضع على عاتق المستخدم تكلفة نقدية وزمن انتظار ولا تقدم دائمًا نتيجة دقيقة ترضي المستخدم؛ وبالتالي نحتاج طريقة فعّالة لتقدير القيم المفقودة لـ Skylines على قواعد بيانات مصادر الحشد الجماعي (crowd-sourcing) الغير كاملة. تقترح هذه الأطروحة طريقة لتقدير القيم المفقودة لـ Skylines على قواعد بيانات مصادر الحشد الجماعي (crowd-sourcing) الغير كاملة وذلك بالتخلص من عناصر البيانات (tuples) الغير مرغوب فيها من قاعدة البيانات الأولية الغير كاملة باستخدام فلترة البيانات لأجل تبسيط عملية تقدير القيم المفقودة. علاوة على ذلك فانها تحاول ايضا الاستفادة من البيانات المتبقية واستغلال العلاقات الضمنية بين الخواص (attributes) لتخمين القيم المفقودة في Skylines.حيث انه تم توظيف فكرة استنباط علاقات الخواص التي تؤدي إلى توليد مجموعة من التبعيات الوظيفية التقريبية (AFDs) للمساعدة في تقدير القيم المفقودة. بالاضافة الى ذلك تركز هذه الاطروحة على تقليل عدد القيم المراد تقديرها باستخدام الحشد (crowd) وذلك بان تتم عملية التقدير من خلال استخدام قاعدة البيانات الموجودة في الـ crowd عندما يكون التقدير باستخدام العلاقات الضمنية غير مناسب. بالاضافة الى ذلك عند العمل على منصة الحشد الجماعي لتقدير القيم المفقودة لـ Skyline سيؤخذ في الاعتبار العوامل التي تؤثر على معالجة البيانات على منصة الحشد الجماعي (crowd-sourcing platform) مثل التكلفة النقدية ، وقت الانتظار ودقة النتائج. تم إجراء تجارب مكثفة على مجموعات البيانات الاصطناعية والحقيقية و قد أثبتت نتائج التجارب أن الطريقة االمقترحة لتقدير القيم المفقودة لـ Skylines على قواعد بيانات مصادر الحشد الجماعي الغير كاملة قابلة للتطوير وتتفوق على الطريقة الحالية. كما انها تعمل على تبسيط عملية تقدير القيمة للقيم المفقودة في Skylines بشكل كبير عن طريق تقليل عدد القيم المفقودة المراد تقديرها في قاعدة البيانات الأولية التي تصل الى اكثر من 80%. وقد أظهرت النتائج أيضًا أن الطريقة المقترحة قد أنتجت قيمًا تقديرية ذات معدل خطأ نسبي أقل مقارنةً بأحدث الأساليب. حيث يمكن ان تقدير حوالي 40٪ من القيم المفقودة محليًا بدقة عالية تصل إلى 90٪ ، بينما يرتفع معدل التقدير المحلي للقيم المفقودة إلى 95٪ في قاعدة بيانات الارتباط .أخيرًا أظهرت نتائج التجارب أن الطريقة المقترحة أدت إلى انخفاض كبير في

التكلفة النقدية و زمن الانتظار عند تقدير القيم المفقودة لـ Skylines باستخدام قواعد بيانات مصادر الحشد الجماعي.

APPROVAL PAGE

The thesis of Marwa Behjat Swidan has been approved by the following:

_____
Ali A. Alwan Al-Juboori
Supervisor

_____
Sherzod Turaev
1st Co-Supervisor

_____
Yonis Gulzar
2nd Co-Supervisor

_____
Norsaremah Salleh
Internal Examiner

_____
Feras Hanandeh
External Examiner

_____
Fatimah Sidi
External Examiner

_____
Radwan Jamal Elatrash
Chairman

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Marwa Behjat Swidan

Signature ..... Date .....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

SKYLINE QUERIES COMPUTATION AND MISSING VALUES ESTIMATION OF SKYLINES ON CROWDSOURCED-ENABLED INCOMPLETE DATABASE

I declare that the copyright holders of this thesis are jointly owned by the Student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may only be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Marwa Behjat Swidan

……..…….. ……..
Signature Date

DEDICATION

To my Dearest and First Teachers: My Father and Mother

To my lovely husband

To my beloved children

To my beloved sisters and brother

ACKNOWLEDGEMENTS

In the Name of Allah "And if you would count the blessings of Allah you would not be able to count them…" (Surah Ibrahim 14: 34). Foremost, I would like to thank Allah who guides me in my education and life path. He, Almighty, has been blessing me through every phase of my life and supporting me with good people to whom I express my deepest appreciation.

I would like to express my gratitude to the many people who supported me during my scientific career and, without whom this thesis would not have been possible.

My profound gratitude and heartfelt thanks to my family: my dearest father, mother, husband and my children, who I am always grateful for them. They encouraged me to pursue my PhD and without their support, patience, concern, love, and prayer, it was impossible for me to complete my PhD studies.

I am also grateful to my supportive supervisor Asst. Prof. Dr. Ali A. Alwan and co-supervisor Assoc. Prof. Dr. Amelia Ritahani Ismail who have continuously encouraged me throughout my research. I would like to thank, my main supervisor Dr. Ali very much, whose patience, encourages, generosity was endless and whose critical advice contribute to the quality of my work. Our conversations were an invaluable asset in developing my research, and without his encouragement, I would not finish this final work in my PhD study.

Of course, the work presented in this thesis was not accomplished alone. In addition to my supervisor, I am very fortunate to have worked with my friends and colleagues in the Lab, Dr. Yonis, Dr. Arjumand Dr. Imama and Dr. Abdolrahem thanks a lot for your support during my study.

I highly indebted to all my family and friends in the IIUM and outside it, their presence made me strong and help me to overcome all the difficulties I faced during my studies.

Thank you so much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 OVERVIEW

In most of contemporary database applications, the issue of missing data has become a frequent phenomenon, especially when databases are generated automatically using various information extraction or information integration approaches. There are many factors responsible for rendering the databases imprecise, for example, data integration from different huge databases or users providing incomplete input by ignoring some data, whether intentionally or unintentionally. These activities negatively influence the database contents and deteriorate their quality. These factors impact on the completeness and the correctness of the query result (El Maarry et al., 2015; Lofi et al., 2013a; Lofi et al., 2013b; Wolf et al., 2009). Some queries cannot be optimally answered through traditional database management techniques as the process of answering certain queries relies on information that is incomplete, imprecise, or uncertain. Examples of such problems include translation, handwriting recognition, image understanding, and web databases.

Crowd-sourcing has become an effective solution for such types of queries by exploiting knowledge, ideas, experiences, and skills of crowd workers to process information and obtain accurate answers to difficult or very cost-intensive queries on the web. The process of integrating individuals who carry out computations using software systems is known as hyped human/machine systems (Franklin et al, 2011; Li et al., 2016; Li et al., (2017b); Parameswaran & Polyzotis, 2011; Parameswaran et al., 2012; Schmidt & Jettinghoff, 2016; Xintong et al., 2014). Several crowd-sourcing database systems have already been developed to extend the traditional databases

system into crowd-sourced databases systems. This extension supports more types of queries by means of the power of people such as CrowdDB (Franklin et al., 2011), Qurk (Marcus et al., 2011), Deco (Park et al., 2012). These crowd-sourced database systems are associated with certain crowd-sourcing marketplaces such as AMT (Amazon Mechanical Turk) (Franklin et al., 2011) and CrowdFlower (Li et al., 2016) to attract crowds to work for them. The crowd-sourced database leverages many aspects of traditional database systems.

In recent years, skyline queries have gained considerable attention for their usefulness in multi-criteria decision making and decision support applications. The main aims of skyline queries is to generate the best result (skylines) for the user based on his or her preferences by reducing the search space as much as possible by focusing exclusively on those sets of tuples that may potentially be the skylines. Data incompleteness has a negative impact on the skyline queries processing and may lead to loss of the *transitivity property* of the skyline technique. This might also result in the issue of *cyclic dominance* between the tuples as some tuples are incomparable with each other and thus no tuple is considered as skyline (Khalefa et al., 2008). Most importantly, the skylines produced from the incomplete database are also incomplete as some missing values in one or more attributes are introduced. Retrieving skylines with incomplete data is undesirable as it results in misleading the users and ends with inappropriate selection. Therefore, manipulating these incomplete skylines by replacing the missing values with some plausible values is needed as it provides the users with complete skylines that help them to make the best decision. This research work attempts to investigate the impact of processing skyline queries in crowd-sourced enabled incomplete databases. The focus is directed towards how to generate precise estimated values for the missing values of the skylines by exploiting the available data and the

embedded relationship between the attributes in the incomplete database and the crowd-sourced databases.

1.2 PROBLEM STATEMENT

Skyline queries have been investigated intensively since their first introduction into the database community by Borzsony in 2001 (Borzsony et al., 2001). Skyline queries emphasize on pruning the search space of large numbers of tuples to a small set of interesting tuples by removing those tuples that are dominated by others. Numerous algorithms have been designed to process skyline queries in databases with complete data assuming that the database is fully complete (no missing values) (Bartolini et al., 2006; Borzsony et al., 2001; Chomicki et al., 2003; Godfrey et al., 2005; Kossmann et al., 2002; Lee et al., 2010; Lin et al., 2013; Mortensen et al., 2015; Mullesgaard et al., 2014; Tan et al., 2001; Wang et al., 2016; Wong et al., 2008). The focus of these approaches is on minimizing the exhaustive searching process and shrinking the scanning space in order to derive the skylines. The searching space is determined by the number of pairwise comparisons that need to be performed between the tuples in identifying skylines. That means a higher number of pairwise comparisons results in a larger searching space and vice versa.

A group of researchers have highlighted the issue of missing data on processing skyline queries and have investigated the impact of missing values of the database attributes on performing the skyline queries and the completeness of the skyline results. A critical issue of data incompleteness in skyline queries is that the *transitivity property* of the skyline technique is lost and that the dominance relationship between tuples becomes *cyclic*. To clarify these two problems, the following incomplete database example is given. Assume a database consists of three tuples with missing values in one

or more attributes, namely *a*(?, 4, 2, ?), *b*(1, 3, ?, 4), and *c*(?, ?, 3, 2). Here the missing values have been replaced with (?). Applying the skyline technique on the three tuples with missing values results in the following: First, tuple *a* dominates *b* based on the common non-missing attributes (A_2), and *b* dominates *c* on the common non-missing attribute (A_4). Thus, it can be observed that tuple *c* dominates based on the common non-missing attribute (A_3). Therefore, the *transitivity property* of the skyline technique no longer holds and the dominance relationship is *cyclic* as none of these three tuples can be considered a skyline as each tuple is dominated by at least one other tuple (Alwan et al., 2018; Khalefa et al., 2008; Swaminathan et al., 2019).

To solve the above problem of skyline computation over incomplete data, many approaches of skyline query processing on the traditional incomplete database have been proposed (Alwan et al., 2016; Arefin & Morimoto, 2012; Bharuka & Kumar, 2013a; Bharuka & Kumar, 2013b; Gao et al., 2014; Gulzar et al., 2018; Khalefa et al., 2008; Wang et al., 2017; Zhang et al., 2016). However, these approaches have introduced solutions to the issue of processing skyline queries on incomplete databases without paying attention to manipulating the missing values of the skylines. Hence, the skylines returned from the incomplete database to the user will also be incomplete. Retrieving skylines with incomplete data is thus considered a dissatisfactory approach as it may lead to user misguidance and wrong decision-making.

To the best of our knowledge, the only work that has so far raised the issue of value estimation for skylines in the traditional database is that of Alwan et al., (2018). Nevertheless, it is not always suitable to use the available data in the incomplete database to estimate the missing values. In certain cases, the relative error produced between the real missing and the estimated values can be very large, which may impact

the quality of the skylines. Thus, outsourcing these missing values using a crowd-sourced database, for instance, might provide a more precise estimation.

Skyline queries are generally quite expensive operations, especially when executed on the crowd-sourcing database due to the crowd-sourced database contains a large amount of data for various types of databases. Moreover, to improve the skyline results in the incomplete database, crowd-sourced databases have been exploited a crowd workers to fill the missing values. However, interfering with humans involves additional monetary cost and results in latency as all missing values need to be elicited from the crowd-sourcing database.

To the best of our knowledge, insufficient attention has been paid to processing skyline queries in the crowd-sourcing enabled incomplete database; so far, very few approaches have been proposed. The (El Maarry et al., 2015; Lofi et al., 2013a; Lofi et al., 2013b) strategies introduced a hybrid approach incorporating the heuristic techniques with crowd-sourcing to process the skyline queries in an incomplete crowd-sourcing database. This proposed approach is based on the KNN and Min-Max values algorithms as a first phase for estimating all the missing values in the database, but the results have shown that these algorithms are not suitable when the missing rate is more than 20%. However, the missing values with high relative error have estimated from the crowd before identify the final skyline; hence, the estimation process became expensive.

Lastly, the most recent work related to skyline queries in the crowd-sourcing database has been introduced by Lee et al., (2016). Lee's research team assumed that the user submits a skyline query into the database after which some results have to be extracted from the crowd called virtual attributes with missing values. These virtual attributes do not constitute part of the initial database and are generated by the crowd worker and retrieved during the run-time. Thus, an efficient approach is needed

for estimating the missing values of the skylines in crowd-sourced enabled incomplete databases. The approach should take into consideration reducing the processing time of the value estimation for the missing values of the skylines, improving the accuracy of the estimated values, and minimizing the monetary cost of value estimation and the time latency through using the crowd-sourced databases.

1.3 RESEARCH QUESTIONS

This section describes the research questions to be answered in this research work. The research questions are as follows:

1. How to estimate the missing values in the incomplete crowd-sourcing enabled database on the skyline queries process?
2. How can reduce the execution time, monetary cost, and time latency of the skyline over the crowd-sourced enabled incomplete database?
3. How can the process of value estimation for the missing values of skylines in crowd-sourced enabled incomplete databases be improved?

1.4 RESEARCH OBJECTIVES

This research work aims to achieve the following objectives:

1. To propose an efficient approach that is able to process skyline queries over crowd-sourced enabled incomplete databases. The approach attempts to estimate the accurate values for the incomplete values in the database before identifying the final skylines.
2. To propose an approach that eliminates the dominated tuples from the initial incomplete database in the early stage before identifying the skylines.

3. To exploiting the available data in the initial incomplete database and the implicit relationships between the attributes, taking into consideration the relative error between the real missing and the estimated values. Furthermore, when local estimation is inappropriate, been relying on the crowd workers imputation with the aim of minimizing the monetary cost and time latency while sustaining high accuracy for the estimated values.

1.5 THE SCOPE OF THE RESEARCH

The scope of this research work is outlined in the following points:

- This research work uses the crowd-sourcing relational database model as it is the most dominant model among contemporary database applications.
- The type of preference queries considered in this research work is limited to skyline queries.
- This thesis employs synthetic and real databases. The synthetic dataset contains correlated and independent data while the real dataset contains NBA and Car data. These two databases are the most frequently used types of databases in the area of skyline query processing on incomplete databases.
- This research work exploits the available data in the initial incomplete database to generate the estimated values of the skylines. Thus, we assume that the data involved in the skyline and the estimation processes are in numeric form since it constitutes the most common data form used in skyline queries.

- Lastly, we also assume that the missing values may be present in one or more attributes of the database and that the missing rate may be as high as 60% of the entire amount of the database.

1.6 CONTRIBUTIONS

The main contributions of this research work can be summarized as follows:

- Efficient data filtration that prunes the unwanted dominated tuples from crowd-sourced enabled incomplete databases before applying the process of value estimation for the missing values of the skylines is proposed. The proposed approach aims at eliminating the dominated tuples before scanning the incomplete database to identify the implicit relationships between the attributes of the database. The data analysis assists in generating the approximate functional dependencies between the attributes, which in turn provides an accurate estimation for the missing skyline values. Thus, prior data filtration is very beneficial as it limits the process of value analysis and determines the attribute correlation to those candidate tuples that can contribute in forming the skylines of the incomplete database. This optimization process helps avoid many unnecessary exhaustive processes when estimating the missing values of the skylines. The proposed solution of data pruning is suitable for processing skyline queries over various types of database systems beside crowd-sourced enabled incomplete database such as autonomous web database, decision support system, decision-making system, and recommendation system. It is very common for such databases to contain missing values in one or more attributes.

- An efficient method for estimating the missing values of skylines in crowd-sourced enabled incomplete databases called *AFD*-based is proposed. The *AFD*-based strategy attempts to generate estimated values for the missing values for the skylines of the incomplete database. The principle of the *AFD*-based approach consists of employing the embedded relationship between the attributes to generate a set of approximate functional dependencies that assist in generating precise estimation for the missing values of the skylines. Exploiting the initial incomplete data and the implicit relationships between the attribute allows us to simplify the process of value estimation and minimize the relative error between the real and the missing values of the skylines. Thus, the number of missing values that need to be estimated is reduced, which in turn results in less monetary cost and lower time latency when using crowd resources.
- A crowd-sourced enabled approach to impute the missing values of the skylines using the crowd-sourcing databases is proposed. This approach is used to generate accurate estimated values for the skylines when the result of the local estimation is undesirable. It might occur that the derived estimated values of the skylines using the *AFD*-based approach result into an unacceptable rate of relative error between the real missing and the estimated values. Thus, exploiting crowd-sourced resources is very beneficial as it helps derive skylines with no missing values in incomplete database systems. The proposed approach results in minimum monetary cost and time latency while maintaining high precision for the estimated values of the skylines using the crowd-sourced database.