UNIVERSITI PUTRA MALAYSIA

PROCESSING SKYLINE QUERIES IN CENTRALISED AND DISTRIBUTED INCOMPLETE DATABASES

ALI AMER ALWAN

PROCESSING SKYLINE QUERIES IN CENTRALISED AND DISTRIBUTED INCOMPLETE DATABASES

ALI AMER ALWAN

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA

2013

PROCESSING SKYLINE QUERIES IN CENTRALISED AND DISTRIBUTED INCOMPLETE DATABASES

By

ALI AMER ALWAN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

June 2013

DEDICATION

To my Dearest and First Teachers: My Father and Mother
To my lovely wife "Zubaidah"
I will always be grateful for your endless love, unlimited support
and deep faith in me

To my beloved son Amer

To my lovely brother and sisters

Ali

Abstract of thesis presented to Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

PROCESSING SKYLINE QUERIES IN CENTRALISED AND DISTRIBUTED INCOMPLETE DATABASES

By

ALI AMER ALWAN

June 2013

Chairman : Professor Hamidah Ibrahim, PhD

Faculty : Computer Science and Information Technology

Skyline queries incorporate and provide a flexible query operator that returns data items (skylines) which are not being dominated by other data items in all dimensions (attributes) of the database. Most of the existing skyline techniques determine the skylines by assuming that the values of dimensions for every data item are available (complete). However, this assumption is not always true particularly for multidimensional database as some values may be missing. The incompleteness of data leads to the loss of the *transitivity property* of skyline technique and results into failure in *test dominance* as some data items are incomparable to each other. Furthermore, missing values will influence negatively on the process of finding skylines, leading to high overhead, due to exhaustive pairwise comparisons between the data items. This problem becomes more complicated when multiple tables with incomplete data need to be accessed in determining the skylines. Since in distributed database tables are spread over various locations, therefore, join operation is needed in identifying the skylines. Joining these dimensions without any filteration will result in a huge amount of data to be joined. Furthermore, most of the previous works in the area of skyline queries in incomplete database emphasized only on retrieving skylines without estimating the missing values. In other words, the derived skylines have some missing values in one or more dimensions. However, in many cases users are concerned about the values in these dimensions.

This thesis aims at proposing an efficient approach which is able to identify skylines in incomplete database. The approach employs the concepts of clustering data to partition the initial database into a set of distinct clusters. Then, the derived clusters are further divided into smaller groups and local skylines of each cluster are then identified. Next, a set of virtual skylines called *k*-dom that is derived from the local skylines are merged to derive a global *k*-dom skyline which is inserted at the top of each cluster to identify the candidate skylines. The final skylines are retrieved after conducting pairwise comparisons among the candidate skylines. The approach is extended to process skyline queries in incomplete distributed databases by pruning the input relations before conducting the join and skyline operations.

The thesis also proposes an approach to estimate the missing values in the skylines. The approach utilizes the concept of mining attribute correlations to generate approximate functional dependencies (AFDs) that capture the relationships between dimensions. Besides, the strength of probability correlations between dimensions is computed in order to estimate the values. Then, the skylines are ranked according to the confidence of the generated AFD and the strength of probability correlations of the dimensions.

Several experiments on synthetic and real datasets have been conducted. The results showed that our proposed approach for processing skyline queries in incomplete database has reduced the number of pairwise comparisons in the range of 75%-93% and the processing time in the range of 50%-89% compared to the previous approach. While the approach for processing skyline queries in incomplete distributed databases achieved between 56% to 88% reduction in the processing time and 84% to 90% for network cost compared to the previous approach. Lastly, the results for imputing the missing values of the skylines have shown that our approach achieved 25% error rate between the real missing values and the estimated values of the skylines.

Abstrak tesis yang dikemukakan kepada senat Uinversiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

MEMPROSES PERTANYAAN *SKYLINE* DI DALAM PANGKALAN DATA TIDAK LENGKAP BERPUSAT DAN TERAGIH

Oleh

ALI AMER ALWAN

Jun 2013

Pengerusi : Profesor Hamidah Ibrahim, PhD

Fakulti : Sains Komputer dan Teknologi Maklumat

Pertanyaan Skyline menggabungkan dan menyediakan operator pertanyaan fleksibel yang memulangkan item data (*skyline*) yang tidak didominasi oleh item data lain dalam semua dimensi (atribut) pangkalan data. Kebanyakan teknik *skyline* yang sedia ada menentukan *skyline* dengan mengandaikan bahawa nilai dimensi bagi setiap item data adalah tersedia (lengkap). Walau bagaimanapun, andaian ini tidak selalunya benar khususnya bagi pangkalan data pelbagai dimensi memandangkan beberapa nilai mungkin tiada. Ketidaklengkapan data menyebabkan kehilangan sifat transitiviti teknik *skyline* dan mengakibatkan kegagalan dalam ujian penguasaan kerana beberapa item data tidak sepadan antara satu sama lain. Tambahan pula, nilai yang hilang akan memberi pengaruh negatif terhadap proses mendapatkan item data *skyline*, menyebabkan atas yang tinggi, akibat perbandingan berpasangan yang tuntas. Masalah ini menjadi lebih rumit bila pelbagai jadual yang mempunyai data tak lengkap perlu dinilai untuk menentukan skyline. Memandangkan dalam pangkalan data teragih jadual disebarkan kepada berbagai lokasi, maka operasi sambungan diperlukan untuk mendapatkan *skyline*. Penggabungan dimensi-dimensi ini tanpa sebarang tapisan akan mengakibatkan penggabungan sejumlah data yang besar. Lagipun, kebanyakan kajian terdahulu dalam bidang pertanyaan *skyline* dalam pangakalan data tak lengkap hanya menekankan kepada mendapatkan semula *skyline* tanpa mengambil kira nilai-nilai yang tiada. Dalam kata lain *Skyline* yang diperolehi tidak mengandungi nilai dalam satu atau lebih dimensi. Walau bagaimanapun, dalam banyak kes, pengguna lebih menitik beratkan nilai-nilai dalam dimensi-dimensi ini.

Tesis ini bertujuan mencadangkan satu pendekatan cekap yang boleh mengenal pasti skyline dalam pangkalan data tak lengkap. Pendekatan ini menggunakan konsep pengelompokan data untuk membahagi pangkalan data asal kepada satu set kelompok yang berbeza. Kemudian, kelompok yang diterbitkan dibahagikan lagi kepada kumpulan lebih kecil dan *skyline* setempat bagi setiap kelompok kemudiannya dikenal pasti. Selepas itu, satu set skyline maya yang dipanggil *k*-dom yang diterbitkan daripada skyline setempat digabungkan untuk menerbitkan satu *skyline* *k*-dom sejagat yang diselitkan di atas setiap kelompok untuk mengecam calon-calon *skyline*. *Skyline* terakhir didapatkan semula selepas dilaksanakan perbandingan berpasangan dalam kalangan calon *skyline*. Pendekatan ini dilanjutkan untuk memproses pertanyaan *skyline* dalam pangkalan data teragih tak lengkap dengan memangkas hubungan input sebelum dilakukan operasi cantuman dan *skyline*.

Tesis ini juga mencadangkan satu pendekatan untuk menyenggara nilai-nilai yang tiada dalam *skyline* melalui anggaran nilai hampir. Pendekatan ini menggunakan konsep perlombongan korelasi atribut untuk menjana satu Kebergantungan Fungsian Hampiran (AFD) yang menangkap hubungan antara dimensi. Di samping itu, kekuatan korelasi kebarangkalian antara dimensi dikira untuk menganggarkan nilai tersebut. Kemudian, *skyline* tersebut disusun mengikut kekuatan AFD yang dijana dan kekuatan korelasi kebarangkalian antara dimensi.

Beberapa eksperimen ke atas set data sintetik dan sebenar telah dijalankan. Keputusan menunjukkan pendekatan yang dicadangkan untuk memproses pertanyaan *skyline* dalam pangkalaan data tak lengkap telah menurunkan bilangan perbandingan berpasangan dalam julat 75%-93% dan masa pemprosesan dalam julat 50%-89% berbanding dengan pendekatan-pendekatan terdahulu. Manakala pendekatan untuk memproses pertanyaan skyline dalam pangkalaan data teragih tak lengkap telah mencapai penurunan 56% hingga 88% dalam masa pemprosesan dan 84% hingga 90% bagi kos rangkaian. Akhir sekali, keputusan bagi menggantikan nilai *skyline* yang tiada telah menunjukkan bahawa pendekatan ini menghasilkan kadar ralat 25% di antara nilai sebenar *skyline* yang tiada dengan nilai yang dianggarkan.

ACKNOWLEDGEMENTS

In the name of *ALLAH*, the most merciful and most compassionate. Praise to *ALLAH* S. W. T. who granted me strength, courage, patience and inspirations to complete this work.

I do not think that words can express the gratitude that I have for my supervisor Professor Dr. Hamidah Ibrahim. As a top researcher in database system, her smart intuition on computer science, stimulating suggestions and encouragement helped me in all aspects of study.

First and foremost, I heartiest would to sincere thanks my supervisor Prof. Dr. Hamidah Ibrahim, for her incredible guidance, continuous support, and encouragement. Always having time for me and readily providing her technical expertise throughout the period of my study. I owe more than I can ever repay. Only has the successful completion of this work become possible due to her supervision, she is the first person to thank for making my Ph.D. at the Universiti Putra Malaysia a very enjoyable experience. Her high stand of diplomatic power and professionalism set a great model for me to follow.

To my thesis committee members, Assoc. Prof. Dr. Nor Izura Udzir, and Dr. Fatimah Sidi, I would like to express appreciation for their insightful comments, questions, criticisms, and suggestions on the work. Their critical appraisals of my papers and presentations are extremely valuable for the improvement in my thinking also her kindness and willingness to help is unforgettable.

I delighted to gratefully acknowledge the Universiti Putra Malaysia for giving me the opportunity to complete my study and their financial support during my Ph.D journey. Thanks also are due to other members of the academic, and the technical staff in the Faculty of Computer Science and Information Technology for their help and effort to provide facilities, equipments, and an excellent environment to accomplish this research.

I also I would like to thank many people I have met during my stay in Malaysia for their help, enjoyable discussions and some good times.

Finally, I would like to express my love and deepest thanks to my noblest father Amer Alwan Al-Jubouri and my great mother Ebtisam Mohammad were the reason of my success, I'm indebted to them. To my brother, and sisters, who have loved and support me throughout my life, thank you.

Ali Amer Alwan

June 2013

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

Hamidah Ibrahim, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Nur Izura Udzir, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Fatimah Sidi, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Putra Malaysia or other institutions.

ALI AMER ALWAN

Date: 19 June 2013

TABLE OF CONTENTS