التخصص: هندسة كهربية
القسم: هندسة الحاسبات والنظم
الدرجة: الماجستير

جامعة عين شمس

كلية الهندسة

الاسم: احمد فاروق عبدالعزيز ابراهيم
تاريخ الميلاد: ١٩٧٧/٢/٢٣
البكالوريوس : يوليو ٢٠٠١ **جامعة** : الازهر

هيئة الاشراف
أ.د. هدى قرشى محمد
د. إسلام احمد محمد المداح

لجنة الحكم
أ.د. جمال الدين علي
أ.د. ماجده بهاء الدين فايق

تحسين الأداء في طرق التنقيب عن البيانات بإيجاد الانماط المتكررة في قواعد البيانات الضخمة

ان التنقيب عن البيانات من اهم الطرق المستخدمة في تحليل واستخدام البيانات في العصر الحديث.وتعد المشكلة الاساسية في التنقيب عن البيانات هي عملية ايجاد الانماط المتكررة في قواعد البيانات الضخمة .وذلك لان ايجاد الانماط المتكررة تلعب دورا هاما واساسيا في الوصول الي المعلومات الهامة من قاعدة البيانات . و نظرا للتطور السريع في مجال الكمبيوتر و تكنولوجيا الاتصالات مما كان سببا اساسيا في ظهور مفهوم جديد في التنقيب عن البيانات وهو الطرق الموزعة في التنقيب عن البيانات. والذي يعمل علي الحد من المشاكل التي تواجه طرق التنقيب العادية عن البيانات مثل الاعاقة المتتالية و تقليل الوقت الزمني اللازم عند التعامل مع قواعد البيانات الضخمة . فإن الهدف من هذا العمل هو تقديم تحسين في احدي الطرق الموزعة(parallel buddy prima) في التنقيب عن البيانات والتغلب علي كافة المشاكل التي واجهت هذه الطريقة .ومن هذه المشاكل الحاجة الي تحميل كافة البيانات في الذاكرة ويعد هذا الامر غير ملائم لقواعد البيانات الضخمة. يعمل النظام المقترح علي عدم الحاجة الي البحث في كافة السجلات الموجودة بقاعدة البيانات في كل مرة يتم فيها حساب عدد مرات تكرار الانماط المختلفة. و لقد تم تنفيذ الخوارزم المقترح ومقارنته بالخوارزميات السابقة من حيث تقييم الاداء في دقة النتائج والوقت اللازم لتنفيذ الخوارزم.وتظهر النتائج ان الخوارزم المقترح يعطي نتائج اسرع دون ادني تأثير علي دقة النتائج.

Frequent Itemset Mining for Big Data

Data mining has been a powerful technique in analyzing and utilizing data in today's information rich society. A fundamental problem in data mining is the process of finding frequent patterns in large datasets. Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases. With the rapid development of computers and communication networks, data mining in distributed environment is becoming a heated problem. Distributed data mining (DDM) is expected to relieve current mining methods from the sequential bottleneck, and provide the ability to scale to massive data sets and improve the response time. The aim of this work is to present a modified parallel algorithm that makes an enhancement of parallel buddy prima algorithm. The modified parallel algorithm doesn't need to load all transaction in memory. It may be inapplicable for large data sets with many distinct items. Modified algorithm prevent full scan of all transaction set each time we calculate the support count of a candidate set. The algorithm is implemented, and is compared to its predecessor algorithms. The comparisons were made on processing time and the accuracy of each algorithm. Results of applying the proposed algorithm show faster performance than other algorithms without scarifying the accuracy.

ملخص

ان التنقيب عن البيانات من اهم الطرق المستخدمة في تحليل واستخدام البيانات في العصر الحديث.وتعد المشكلة الاساسية في التنقيب عن البيانات هي عملية ايجاد الانماط المتكررة في قواعد البيانات الضخمة .وذلك لان ايجاد الانماط المتكررة تلعب دورا هاما واساسيا في الوصول الي المعلومات الهامة من قاعدة البيانات . و نظرا للتطور السريع في مجال الكمبيوتر و تكنولوجيا الاتصالات مما كان سببا اساسيا في ظهور مفهوم جديد في التنقيب عن البيانات وهو الطرق الموزعة في التنقيب عن البيانات. والذي يعمل علي الحد من المشاكل التي تواجه طرق التنقيب العادية عن البيانات مثل الاعاقة المتتالية و تقليل الوقت الزمني اللازم عند التعامل مع قواعد البيانات الضخمة . فإن الهدف من هذا العمل هو تقديم تحسين في احدي الطرق الموزعة(parallel buddy prima) في التنقيب عن البيانات والتغلب علي كافة المشاكل التي واجهت هذه الطريقة .ومن هذه المشاكل الحاجة الي تحميل كافة البيانات في الذاكرة ويعد هذا الامر غير ملائم لقواعد البيانات الضخمة. يعمل النظام المقترح علي عدم الحاجة الي البحث في كافة السجلات الموجودة بقاعدة البيانات في كل مرة يتم فيها حساب عدد مرات تكرار الانماط المختلفة.

و لقد تم تنفيذ الخوارزم المقترح ومقارنته بالخوارزميات السابقة من حيث تقييم الاداء في دقة النتائج والوقت اللازم لتنفيذ الخوارزم.وتظهر النتائج ان الخوارزم المقترح يعطي نتائج اسرع دون ادني تأثير علي دقة النتائج.

Abstract

Data mining has been a powerful technique in analyzing and utilizing data in today's information rich society. A fundamental problem in data mining is the process of finding frequent patterns in large datasets. Frequent itemsets play an essential role in many data mining tasks that try to find interesting patterns from databases. With the rapid development of computers and communication networks, data mining in distributed environment is becoming a heated problem. Distributed data mining (DDM) is expected to relieve current mining methods from the sequential bottleneck, and provide the ability to scale to massive data sets and improve the response time.

The aim of this work is to present a modified parallel algorithm that makes an enhancement of parallel buddy prima algorithm. The modified parallel algorithm doesn't need to load all transaction in memory. It may be inapplicable for large data sets with many distinct items. Modified algorithm prevent full scan of all transaction set each time we calculate the support count of a candidate set. The algorithm is implemented, and is compared to its predecessor algorithms. The comparisons were made on processing time and the accuracy of each algorithm. Results of applying the proposed algorithm show faster performance than other algorithms without scarifying the accuracy.

Ain Shams University
Faculty of Engineering
Computer & Systems Engineering Department

Frequent Itemset Mining for Big Data

Thesis

Submitted in partial fulfillment of the Requirements for the M.Sc. Degree in Electrical Engineering
(Computer and Systems)

By

Ahmed Farouk Abd El_Aziz
B.Sc in Computer and Systems Engineering

Supervised By

Dr. Hoda Korashy Mohamed
Computer & Systems Engineering Dept.
Faculty of Engineering, Ain Shams University

Dr. Islam El_Maddah
Computer & Systems Engineering Dept.
Faculty of Engineering, Ain Shams University

Cairo 2015

ABSTRACT

Name: Ahmed Farouk Abd El_Aziz.

Thesis title: Frequent Itemset Mining for Big Data

A thesis for M.Sc. degree. Ain Shams University, Faculty of Engineering, Computer and Systems Engineering Department, 2015.

Discovering the frequent item sets requires a lot of computation power, memory and input/output values, which is better to be provided by parallel processing. The implementation of data mining ideas in high performance parallel and distributed computing environments is becoming crucial for ensuring system scalability and interactivity as data continues to grow in size and complexity.

The aim of this work is to present a modified parallel algorithm that makes an enhancement of parallel buddy prima algorithm. The modified parallel algorithm doesn’t need to load all transaction in memory. It may be inapplicable for large data sets with many distinct items. Modified algorithm prevent full scan of all transaction set each time we calculate the support count of a candidate set. The algorithm is implemented, and is compared to its predecessor algorithms. The comparisons were made on processing time and the accuracy of each algorithm. Results of applying the proposed algorithm show faster performance than other algorithms without scarifying the accuracy.

Keywords:

Frequent itemset, Big data, Parallel data mining

Acknowledgments

First, all my thanks and gratitude to Allah for accepting my prayer successfully complete this work .This thesis arose in part out of years of research. I have worked with a great number of people whose contribution in assorted ways to the research and the making of the thesis deserved special mention. It is a pleasure to convey my gratitude to them all in my humble acknowledgment.

In the first place i would like to thank Prof. Hoda Korashy Mohamed for her qualified support and her availability. Her advices were precious and fundamental for the development of my work. I will be grateful to her.

I want to thank Dr. Islam El_Maddah. I appreciated his help and his continuous support through my work time.

Finally, a special acknowledgement to my father, my mother, and my wife. Their love and support through these years was very important. Words alone cannot adequately express my gratitude.

Contents

Table of Contents

List of Tables

List of Figures

Chapter 1

Chapter 1

INTRODUCTION

1.1 Background

Database has been used in business management, government administration, scientific and engineering data management and many other important applications.

Data mining is an important part of computer science; it aims on finding information from database that is not already known by users of various application areas. The ever-growing amounts of data makes mining in most of them indispensable since the data sets often contain valuable information that cannot be obtained manually.

Frequent itemset mining is a popular part in the area of data mining with the goal of finding values or items that co-occur frequently together in a data set.

In many of applications, the response time of the mining process is important because the process is often executed and the results are required as soon as possible.

In general, there exist a large number of well-designed itemset mining algorithms that are very efficient on small data sets. However, most of them cannot be applied on large data sets because they do not take advantage of the full spectrum of available parallelism provided by modern processors.