جامعة عين شمس
كليه الحاسبات والمعلومات
قسم الحسابات علمية

نمذجة السلوك الزمني في الشبكات الديناميكية

لقسم الحسابات العلمية هذه الرساله مقدمه
كلية الحاسبات والمعلومات
جامعة عين شمس

كإنجاز جزئى من مطالب درجه الماجستير
فى الحاسبات والمعلومات

إعداد

أية محمد زكى إسماعيل

بكالوريوس الحاسبات والمعلومات (2012)
معيدة بقسم االحسابات العلمية
كلية الحاسبات والمعلومات
جامعة عين شمس

تحت إشراف

د.صفاء السيد أمين

أستاذ مساعد بقسم الحسابات العلمية
كلية الحاسبات والمعلومات
جامعة عين شمس

د.دعاء عبد الكريم حجازى

مدرس بقسم الحسابات العلمية
كلية الحاسبات والمعلومات
جامعة عين شمس

2016 - القاهرة

ملخص الرسالة

أهم ومعظم الشبكات المهمة في العالم الحقيقي تتغير باستمرار وتتطور مع مرور الوقت. وقد حازت هذه الطبيعة الديناميكية للشبكات على الكثير من الاهتمام بدافع أهميتها المتزايدة و الانتشار الواسع لهذا النوع من الشبكات. وبسبب خصائصها الجوهرية والخاصة، هذه الشبكات تمثل بنماذج هياكل البيانات الديناميكية. ومن أجل التأقلم مع الطبيعة المتطورة، يجب على النموذج الممثل الاحتفاظ بالمعلومات التاريخية للشبكة جنبا إلى جنب مع وقتها الزمني. فتخزين هذا الكم من البيانات، يطرح العديد من المشاكل من وجهة نظر إدارة هياكل البيانات الديناميكية.

نحن نقدم في هذه الرسالة نظرة شاملة بعمق للمشاكل المتعلقة هياكل البيانات الديناميكية. كذلك نقدم تصنيف لنماذج هياكل البيانات الديناميكية بطريقة منهجية وشاملة. علاوة على ذلك تناقش الرسالة العمليات على هياكل البيانات الديناميكية بما في ذلك الخوارزميات وتمثيل المخرجات، بالإضافه إلى إعطاء فكرة عن كيفية إدارة والتعامل مع إضافة عنصرالوقت لهياكل البيانات الديناميكية.

ومع الأزدهار الملحوظ للشبكات في العالم الحقيقي القائم على هياكل البيانات، أصبح من المهم وجود نموذج بياني ديناميكي قادر على إدارة شبكة التطوربكفاءة. لذلك، تقدم هذه الدراسة النظام المعدل (MG*) القادرا على إدارة الشبكات بأداء كفء. (MG*) تستهلك الحد الأدنى من وقت التحديث، وقت الاسترجاع ، والحد الأدنى من ذاكرة التخزين بطريقة فعالة بالمقارنة مع نماذج هياكل البيانات الديناميكية الحالية. وعلاوة على ذلك، فإنه يعطى نتائج بجودة أفضل.

معظم النماذج الحالية تستخدام هياكل البيانات لتخزين سلسلة من اللقطات، وهذه اللقطات إما تاريخية أو مسترجعه للخضوع للعمليات. على الرغم من استهلاك هذه الهياكل الحد الأدنى لوقت التحديث، فهى تحتوى على تكرار فى البيانات المخزنه، وذلك لأن القطات المتتالية تشترك فى معظم العقد والحواف. العديد من الهياكل المضغوطة تقلل هذا التكرار، ولكن على حساب زيادة وقت التحديث المطلوب لإدراج لقطة جديدة في الهيكل. ونتيجه لذلك، نقدم فى هذه الدراسه الهيكل البيانى Fast-CGI لتحقيق التوازن فى معالجة هذه السلبيات وذلك عن طريق فصل الأجزاء القابله للتغير فى أى تحديث مستقبلى عنباقى الأجزاء التى لا تتغير أبدا.

AINSHAMSUNIVERSITY
Faculty of Computer
&Information Sciences
Scientific Computing Department

Modeling the Temporal Behavior in Dynamic Networks

Thesis submitted to the Department of Scientific Computing
Faculty of Computer and Information Sciences
Ain Shams University

In partial fulfillment of the requirements for the degree
of Master in Computer and Information Sciences

By

Aya Mohamed Zaki Ismail
B.Sc. in Computer and Information Sciences (2012)
Faculty of Computer and Information Sciences
Ain Shams University – Cairo

Under the supervision of

Dr. Safaa Amin
Associate Professor of Scientific Computing Department
Faculty of Computer and Information Sciences
Ain Shams University – Cairo

Dr. Doaa A. El-Kereem Hegazy
Assistant Professor of Scientific Computing Department
Faculty of Computer and Information Sciences
Ain Shams University – Cairo

Cairo-2016

Acknowledgements

First of all I thank Allah, the most merciful and gracious, who gave me the knowledge, patience and strength to complete this thesis ,and blessed me with his inspired gifts to overcome the obstacles I encountered. Allah, no words can describe your help or be enough appreciation worthy of you, the almighty GOD, thank you.

I would like to thank both ***Dr. Safaa Amin*** and ***Dr. Doaa Hegazy*** for their supervision on my work. I am grateful for their invaluable advices, generous support and visionary guidance.

I would like also to thank ***Dr. Mahmoud Attia Sakr*** the one who proposed this master's idea for his scientific contributions, and technical tips that greatly enhanced the work quality. He taught me a lot and was very patient and considerate. I can never repay you no matter what I do.

To my fiance ***TA. Mohamed Osama***, the one of a kind man I am lucky to have by my side, I extend my utmost gratitude and appreciation for your technical and scientific help, continuous supportive guidance, and unique skillset. This thesis is dedicated to you, to make you proud. Without you, everything is nothing.

No words can be enough to express my gratitude and thanks to all my family, my great parents who have devoted themselves to support me in my whole life, not just this work. Also, I want to thank my friends ***Eman Reda, Ghada Hamed and Eman Hamdy*** for their motivational support and encouragement.

Finally, I thank my ***great mother*** for her endless passionate support and encouragement and the sleepless nights she spent to make it easier for me.

Abstract

Most of the critical real-world networks are continuously changing and evolving with time. The dynamic nature of these networks have gained a lot of attention motivated by the growing importance and wide spread impact of this type of networks. Because of their intrinsic and special characteristics, these networks are best represented by dynamic graph models. In order to cope with their evolving nature, the representation model must keep the historical information of the network along with its temporal time. Storing such amount of data, poses many problems from the perspective of dynamic graph data management.

In this thesis, we provide an in depth overview on dynamic graph related problems. A novel categorization and classification of the state of the art dynamic graph models is also presented in a systematic and comprehensive way. Moreover, we discuss processing on dynamic graphs including both its algorithms and output representation, and give an insight on how to manage and handle the added time parameter to dynamic graph models.

With the notable flourish of real-world networks based on graphs, it becomes crucial to find a dynamic graph model that is able to manage efficiently network evolution. So, this study proposes Modified G* (MG*) system that is able to manage the network consuming efficient performance. MG* consumes minimum update time, retrieve time, and minimum memory storage in an efficient manner compared to the existing dynamic graph models. Moreover, it provides results with a better quality.

Most of the existing models use data structures to store sequence of snapshots, which are either historical or retrieved for processing purposes. Despite consuming minimal update time, these data structures induce storage redundancy, since consecutive snapshots share most of their nodes and edges in common. Compressed variants reduce this redundancy, but at the cost of increasing the update time, required to insert a new snapshot into the structure. Therefore, we propose Fast-CGI data structure to balance handling these downsides.

Contents

List of Figures

List of Tables

Abbrevations

MG* Modified G*

Fast-CGI Fast Compact Graph Index

TRD Temporal Relational Database

DB Database

FVF Find Verify and Fix

SM Storage Model

TPM Temporal Provenance Model

FSDNs Fixed Schedule Dynamic Networks

LABS Locality-Aware Batch Scheduling

CGI Compact Graph Index

Split-CGI Split Compact Graph Index

VL-Map Vertex Location Map

VL-Pair Vertex Location Pair

SP Shortest Path

DFS Depth First Search

BFS Breadth First Search

CH Construction Hierarchy

Chapter 1

Introduction

1.1 Motivation

Most real-world networks like social networks [4, 30, 45, 46, 48, 31], transportation networks [42, 18], and other networks contain a vast amount of information. These networks are mostly represented by graphs. These graphs model the network entities and their relations in the form of vertices and edges respectively. The majority of current network graph representations rely on static graphs . Such type of graphs fails to handle the real time changes of networks. That's why there is a significant interest in providing a dynamic graph model that stores the network historical changes and gives the ability to query these changes [24].

Temporal Relational Database "TRD" shares the power of storing historical changes with dynamic graphs. A lot of literature on TRD focus on temporal data model and temporal query language [3, 9, 10, 33, 39, 40, 41]. The two main basic concepts of TRD are valid time and transaction time. Valid time represents the time period that indicates when the fact is true in the real world. Transaction time represents the time period of storing and removing the fact from the DB. In Dynamic graph we focus on valid time, where the goal is to retrieve the graph entities that are valid at any given time instant.

1.2 Problem Definition

Research has focused on static large graph management [1, 6, 7, 11, 16, 23, 28, 29, 32, 44]. However, most of real-world networks evolve with time. Managing these evolving networks has attracted much attention in recent years. The networks evolved data can be kept in a dynamic graph to improve the expressiveness and the quality of search queries as well as snapshot(s) retrieval. Storing the continuous evolution of the network in a dynamic graph makes its storage size grow. Existing dynamic graph models try to limit their storage by eliminating redundant data. However, their update time increases due to the elimination step. This illustrates that there is a tradeoff between the used storage and the update time.

One of the most important features of dynamic graph models is providing the historical information of evolving networks. They represent the historical states of the network. The historical network state at single time point is known as a snapshot. Most approaches answering queries on dynamic graph models depend on retrieving single or several snapshots at specific time point or specific time interval respectively. The accuracy of the retrieved snapshot highly affects the quality of the queries and the retrieval time affects the performance of the queries. Therefore, retrieving historical information in a short time is a crucial objective in dynamic graphs.

1.3 Research Objectives

The goal of our research is to present a formal graph model augmented with time using less cost and exploit the temporal information hidden in a network, which has been mostly ignored so far, to improve the expressiveness and quality of search queries. The objectives of this research can be summarized in the following points:-

1. Study the properties that describe how a dynamic graph changes.

2. Study the already existing dynamic graph models.

3. Provide dynamic graph model that has a better update time, better storage, and better retrieve time.

4. Introduce a comparitive of our proposed model against similar models of the literature regarding the time, storage, and reliability.

1.4 Main Contributions of this Thesis

1. Providing a comprehensive survey on the existing dynamic graph models and its related problems.

2. Proposing an efficient dynamic graph model (MG*) which is a modification on G*. It manages the network evolution efficiently where, it overcomes the update time overhead problem that exists in G*. MG* has better update time and memory storage than G* where the difference is order of magnitude.

3. Providing an enhancement on our proposed MG* for a better retrieve time based on parallelization.

4. Proposing a compact data structure Fast-CGI for storing sequence of snapshots (i.e.,each snapshot represents the network state at a single time point). Fast-CGI marks a significant enhancement in the dynamic graph area where, this type of data structure considered as a main component of many existing dynamic graph models.

1.5 Thesis Organization

This thesis is organized in five chapters including this one. Their contents are described briefly as follows:

. **Chapter 2:** provides the necessary background and overview needed to understand this thesis. It gives an overview of dynamic graphs including related definitions, current existing models ,and processing.

. **Chapter 3:** provides our proposed system MG* that effeciently stores the continuous evolutions of an evolving network.

. **Chapter 4:** provides an enhancement on our proposed MG* system for a better retrieving performance.

. **Chapter 5:** Proposes Fast-CGI data structure for storing sequence of snapshots ”static graphs” in a complete compact manner while maintaining the update time as stable and minimum as possible independent on the continuous growth of the snapshots count.

. **Chapter 6:** Introduces the conclusions and the future works.