THE COMPRESSIBILITY AND THE RANDOMNESS OF COMPRESSED DATA BASED ON FIBONACCI CODE: A NOVEL APPROACH

BY

KAMAL AHMED MULHI AL-KHAYYAT

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

DECEMBER 2021

ABSTRACT

The tremendous growth of data generated daily has made the science of data compression an important and renewable field. It has become the first way to reduce the volume of data to optimize the use of storage units and accelerate the process of transferring data across various types of networks, chiefly the World Wide Web, thus reducing the cost of transport and storage. Compressed data grows with the same frequency as the data itself, which, in turn, created an urgent necessity to understand and analyze the compressed files themselves, and since efforts are focused only on inventing and developing new compression algorithms, few efforts remain trying to understand and analyze compressed files. This research invests in compressed files introducing a new way to analyze and understand compressed data from new angles. This analysis contributes to solutions to practical problems, including the problem of servers in classifying files before actually compressing them with what is known as compressibility. The issue of studying compressibility in systems servers is a sensitive and important issue, given that they provide for the optimum utilization of the physical and programmatic server resources. This research presented a new method by which server systems can distinguish between compressed files from uncompressed files on the one hand, and on the other hand, distinguish between compressed files that need more compression and those that do not need all of this in one frame. Moreover, as the randomness study programs cannot distinguish compressed data from uncompressed data in most cases, this study provided an integrated package of methods for studying the randomness of compressed files called (RTCD). This package can analyze the randomness of compressed files from new practical angles and open the way for the ability to compare compressed files with each other and distinguish between them successfully. This package includes quantitative and graphical measures all set to be standard in practice. The analysis in this study relies on the use of the Fibonacci code as a strong analytical basis capable of knowing the common characteristics of compressed files and can thus distinguish them from uncompressed files successfully. Moreover, the difference in these characteristics within the compressed files circle enables one to know the files that still need more compression. Compared to the well-known techniques that study compressibility and those that study randomness of data, this analysis shows its distinction and its ability to overcome the deficiencies of these methods.

ملخص البحث

جعل ان النمو الهائل للبيانات المتولدة على اساس يومي علم ضغط البيانات مجالا هاما ومتجددا حيث اصبح الوسيلة الأولى في تقليل حجم البيانات من أجل تحسين استخدام وحدات التخزين وتسريع عملية نقل البيانات عبر مختلف انواع الشبكات وعلى راسها الشبكة العنكبوتية مما خفض كلفة النقل والتخزين.وهكذا نمت الملفات المضغوطة بنفس وتيرة البيانات نفسها ,وهذا بدورة ولد ضرورة ملحة لفهم وتحليل الملفات المضغوطة نفسها .وحيث ان الجهود متركزة فقط على اختراع وتطوير خوارزميات ضغط جديدة ,تبقى قلة من الجهود من تحاول فهم وتحليل الملفات المضغوطة والسبب في ذلك أن البيانات المضغوطة عبارة عن بيانات معقدة التركيب ويصعب التنبأ بسلوكها ولذلك تعتبر عشوائية بشكل عام .ان هذا البحث يستثمر في الملفات المضغوطة ويقدم طريقة جديدة لتحليل وفهم البيانات المضغوطة من زوايا جديدة. يساهم هذا التحليل في حل مشاكل عملية من ذلك مشكلة السرفرات في تصنيف الملفات قبل ضغطها فعليا بما يعرف بقابلية الضغط (compressibility) .ان مسالة دراسة قابلية الضغط في انظمة السرفرات هي مسالة حساسة وهامة لما توفرة من استغلال امثل لمصادر السرفرات المادية والبرمجية .قدم هذا البحث طريقة جديدة تستطيع بها أنظمة السرفرات على سبيل المثال من التمييز بين الملفات المضغوطة من الملفات غير المضغوطة من جهة ومن جهة اخرى ميز بين الملفات المضغوطة المحتاجة لمزيد من الضغط وتلك التي لا تحتاج كل ذلك في اطار واحد. علاوة على ذلك , وحيث أن البيانات المضغوطة لا تستطيع برامج دراسة العشوائية من التمييز بينها وبين البيانات الغير مضغوطة في معظم الاحوال , فان هذه الدراسة قدمت حزمة متكاملة من طرق دراسة العشوائية للملفات المضغوطة سميت (RTCD). هذه الحزمة تسطيع أن تحلل عشوائية الملفات المضغوطة من زوايا عملية جديدة وتفتح المجال للقدرة على مقارنة الملفات المضغوطة مع بعضها البعض والتميز بينها وبين الملفات الغير مضغوطة بنجاح . ان هذه الحزمة تشتمل

على مقاييس كمية واخرى بيانية كلها موضوعة لتكون قياسية من الناحية العملية . يعتمد التحليل في هذه الدراسة على استخدام كود الفيبوناتشي كاساس تحليلي قوي قادر على معرفة الخصائص المشتركة للملفات المضغوطة ويستطيع بذلك تمييزها من الملفات الغير مضغوطة بنجاح وعلاوة على ذلك فان الاختلاف في هذه الخصائص داخل دائرة الملفات المضغوطة يمكن من معرفة الملفات التي لا تزال بحاجة الى مزيد من الضغط . بالمقارنة مع التقنيات المعروفة التي تدرس قابلية الضغط وتلك التي تدرس عشوائية البيانات يتبن مقدار تميز هذا التحليل وقدرته على تجاوز النقص في هذه الطرق.

APPROVAL PAGE

The thesis of Kamal Ahmed Mulhi Al-Khayyat has been approved by the following:

_____
Imad Fakhri Al-Shaikhli
Supervisor

_____
Raini Binti Hassan
Co-supervisor

_____
Abdul Wahab Abdul Rahman
Internal Examiner

_____
Shamala K Subramaniam
External Examiner

_____
Hassanuddeen Abdul Aziz
Chairman

DECLARATION

I hereby declare that this thesis is the result of my 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.

Kamal Ahmed Mulhi Al-Khayyat

Signature…… Date…21-9-2020……

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

THE COMPRESSIBILITY AND THE RANDOMNESS OF COMPRESSED DATA BASED ON FIBONACCI CODE: A NOVEL APPROACH

I declare that the copyright holder of this thesis/dissertation 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 purpose.

3. The IIUM library will have the right to make, store in a retrieval 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 Kamal Ahmed Mulhi Al-Khayyat

…… ……21-9-2020…..

Signature Date

DEDICATION

I dedicate this thesis to my dear parents, without whom I would not have reached this stage, to my wife, who was an example of patience and sacrifice, and to my children Mustafa, Fatima, Abdel-Kareem, and Abdel-Hameed, who were indeed a treasured gift from Allah.

ACKNOWLEDGEMENTS

All glory is due to Allah, the Almighty, whose Grace and Mercies have been with me throughout my program. Although it has been tasking, His Mercies and Blessings on me ease the herculean task of completing this thesis.

I am most indebted to my supervisor, Prof. Dr. Imad Fakhri Al-Shaikhli, whose enduring disposition, kindness, promptitude, thoroughness, and friendship have facilitated the successful completion of my work. I put on record and appreciate his detailed comments, useful suggestions, and inspiring queries which have considerably improved this thesis. His brilliant grasp of the aim and content of this work led to his insightful comments, suggestions, and queries which helped me a great deal. Despite his commitments, he took the time to listen and attend to me whenever requested. The moral support he extended to me is in no doubt a boost that helped in building and writing the draft of this research work. I am also grateful to my co-supervisor, Dr. Raini Binti Hassan whose support and cooperation contributed to the outcome of this work.

Lastly, my gratitude goes to my beloved wife and lovely children; for their prayers, understanding, and endurance while away.

Once again, we glorify Allah for His endless mercy on us one of which is enabling us to successfully round off the efforts of writing this thesis. Alhamdulillah

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Despite technological advancements, information grows exponentially, far outpacing growth in hardware. The expected growth in data in 2020 is 44 Zettabytes (ZB), and for 2025, 463 Exabytes (EB) (Desjardins, 2019). Millions of data are created and transmitted worldwide daily. For example, ~147,000 photos are uploaded to Facebook every 60 s, over ~95 million photos and videos are shared on Instagram every day (Aslam, 2020), and ~1.2 billion photos are uploaded to Google Photos per day (Porter, 2019).

The huge data produced on a daily based known as “big data” which become an active research area. (Kolajo et al., 2019), (Ghani et al., 2019)

Not only the social media creates huge of data daily, but the advancement in various sciences also resulted in enormous data, for example in the field of genomics science each personal genome produces a data file of around 100 GB, which will be very costly when thinking about recording the genome of the population for the whole country (Greenfield et al., 2019).

The tremendous growth in data necessitates the use of data compression to minimize the use of storage spaces, increasing the throughput of the network and saving other resources (Jin et al., 2019). Data compression is the science of reducing the size of data which is essential to decrease the financial costs of accommodating and transmitting data. The importance of data compression is even more evident when we are aware of the available storage capacity only being able to store less than 15% of produced data in 2020 (EMC, 2014).

As the data grows, the compressed format of that data grows as well, either stored in storage spaces or transmitted from a point to another on the web. This compressed form is considered the final form of data and it is smaller than the original.

In the end, the compressed data needs to be restored to its original form. If the restoration (called decompression), resulted in the exact original data, the compression is considered lossless compression, but if the restoration resulted in a similar version of the original, it is considered lossy compression. Both lossy and lossless are used to reduce the size of data, and each has its applications.

Data compression has become an integral part of any storage system, including cloud computing which is widely used nowadays. (Barik et al., 2020),(Azar et al., 2020),and(Lu et al., 2021). Since the server should store the data in a compact form so the cost of storage will be minimum, there is a problem if the data itself is incompressible. The server in this case will expend resources on an ultimately futile task. For larger-sized incompressible data, for example, the process of such compressing may take several hours, whereas for the same size of compressible data, it may take several minutes (Harnik et al., 2013) and (Kim et al., 2020).

This longer processing time will also cause other problems such as data delivery latency due to the share of the resources being committed to compressing incompressible data.

The amount and growth of data now and in the future begs the question of how do we identify compressible and non-compressible data?

This question leads to the proposal of a compressibility test for data. The genesis of the test is to minimize the allocation of resources for an ultimately futile undertaking. Larger data are also harder to compress, and the rate of data growth and its size would make this undertaking ever more challenging.

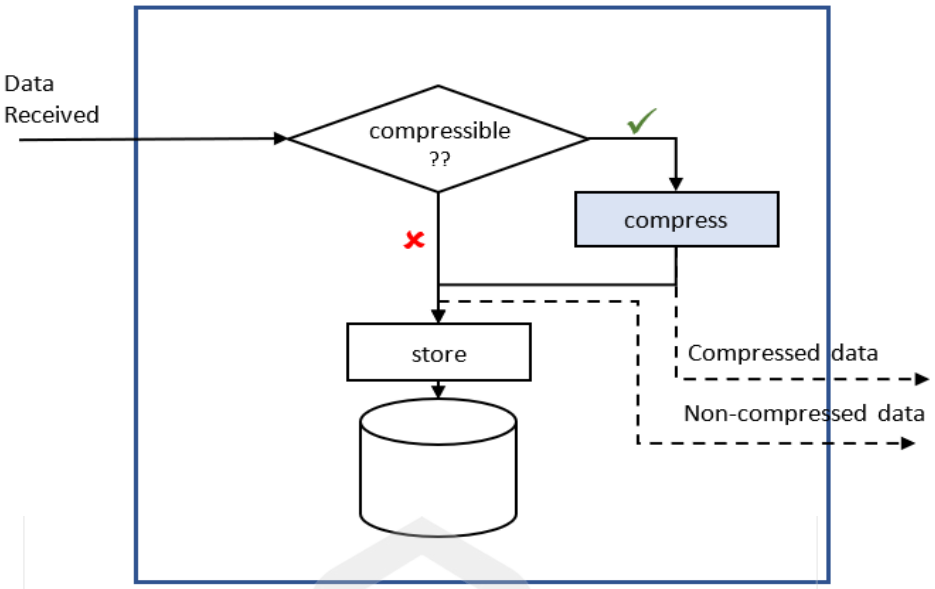

Figure 1.1 Shows the position of compressibility test in server systems.

Upon receiving the data, the compressibility is taken place, so the decision is made to compress or not, whether this process is resulted of online compression (as shown in dotted lines) or for allocating storage spaces, this process is important to save the resources.

The importance of the compressibility test in increasing the throughput as in (Kim et al., 2020) is about 34.15% with sacrificing only 0.09% of the compression ratio for incompressible data, and this percentage is very high in saving resources.

Regarding the storage spaces, the compressibility test can save from 40% to up to 60% of storage spaces, see (Nicolae, 2011). The other resources, such as CPU time, I/O, and memory, will be dramatically affected too.

1.2 PROBLEM STATEMENT

Compressed data is considered as the final minimum form of any data because compressed data, most of the time, is incompressible data (Jon Tate, Christian Burns, Bosmat Tuv-El, 2018). But when the compressed data can be compressed for a second time, do the compressibility tests have the ability to detect their compressibility?

Based on the compressibility tests commonly used, it is possible to figure out the answer.

Starting with the most used compressibility test, named "Prefix estimation". This compressibility test examines the compressibility by compressing a small chunk taken from the head of the data. The main drawback of this method is, no guarantee that the first portion of data will represent the whole data, and for compressed data, this is not an exception.

It is important to note that the search for alternatives or supplements to the test reflects the existence of shortcomings of this test as discussed in (Harnik et al., 2013) which we will refer to next.

"Entropy" is an alternative compressibility test (Balakrishnan & Touba, 2007; Harnik et al., 2013; Oltean et al., 2014b). This compressibility test will not work with compressed data because compressed data have high entropy, whether this data is compressible or not. Compressed data tends to occupy the full range of symbols. Another problem with entropy appeared in (Harnik et al., 2013) which stated that entropy does not imply the compressibility of data in all cases because entropy does not consider the repetition within the data.

Some compressibility tests are not different than entropy, such are "corset" (Harnik et al., 2013), and "byte counting" (Peterson & Reiher, 2016).

To overcome the drawback of entropy, some compressibility tests count on the repetition of patterns (W. Huang et al., 2018; Kipnis & Dror, 2016), others invent a new randomness measurement such as "Pairs distance " (Harnik et al., 2013).

The compressibility tests of this kind are also, not working with compressed data for the same reason of entropy.

Compressible compressed data is not as random as most compressed data, but at the same time, the obvious inner correlations and redundancies are removed by the compression algorithm (D Salomon & Motta, 2010).

Since the compressibility tests will not work on compressed data, we can think of compressibility tests as randomness tests of data.

Despite the word "random" is a vague term and has no exact meaning (Shen et al., 2017), the researchers will follow the Kolmogorov algorithmic definition of randomness (Allender et al., 2006), which implies that data is not random if it is compressible and vice versa. In this case, we can consider the compressibility test as a randomness test too. But can other randomness tests discover the randomness of compressed data and work as a compressibility test?

Like the problem of compressibility tests for compressed data, randomness tests also cannot deal with compressed data predictably. Unfortunately, the result of applying randomness tests on compressed data varies so much from a file to another without any certain rules (Klein & Shapira, 2020).

The previous discussion resulted in two main problems, the first is regarding the compressibility tests which cannot deal with the compressibility of the compressed data, and the second is the inability of randomness tests to recognize the randomness of compressed data (hence the compressibility) in a standard manner.

Both problems reflect the absence of a sophisticated analysis method dedicated to the compressed data, so the compressibility and the randomness of such data be identified.

1.3 RESEARCH QUESTIONS

The research questions are listed as follows:

1. What are the current compressibility tests used to determine the compressibility of data? And can these tests determine the compressibility of compressed data?

2. What are the randomness tests used to measure the randomness of data? and can they work to identify the randomness of compressed data?
3. How do we develop an alternative compressibility test and randomness test sophisticated to deal with compressed data?

1.4 RESEARCH OBJECTIVES

The objectives of this study are:

1. To explore the various methods of compressibility tests used to examine data for compressibility, and whether or not it works for already compressed data.
2. To study the randomness of compressed data and assess whether the available methods are reliable analytical tools that can provide a better understanding of compressed data and serve for detecting its randomness in terms of compressibility.
3. To identify what differentiate compressed data from its uncompressed counterpart and look for possible commonality between compressed data.
4. To develop a new approach that can determine the compressibility and the randomness of compressed data in a measurable and practical way.

1.5 THE IMPORTANCE OF THE STUDY

This study suggests a better understanding of the compressed data by introducing a new approach for analyzing compressed data, with a two-fold approach:

1. It will provide an alternative for studying the randomness of compressed data instead of test packages such as that of NIST (Rukhin et al., 2001)