Evaluation of Idempotency & Block size of Data on the Performance of Normalized Compression Distance Algorithm

Kamran Ali Bajwa & Mandhapati Venkata Srikanth
This thesis is submitted to the School of Computing at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Authors:
Kamran Ali Bajwa
Personal number: 870826-3598
E-mail: kamranalibajwa@gmail.com

Venkata Srikanth Mandhapati
Personal Number: 880324-1259
E-mail: mandhapatims@gmail.com

University advisor:
Dr. Stefan Axelsson
Stefan.axelsson@bth.se
School of Computing,
Blekinge Institute of Technology

School of Computing
Blekinge Institute of Technology
SE – 371 79 Karlskrona
Sweden

Internet : www.bth.se/com
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57
ABSTRACT

Context: Normalized compression distance (NCD) is a similarity distance metric algorithm which is used for the purpose of analyzing the type of file fragments. The performance of NCD depends upon underlying compression algorithm to be used. We have studied three compressors bzip2, gzip and ppmd, the compression ratio of ppmd is better than bzip2 and the compression ratio of bzip2 is better than gzip, but which one out of these three is better than one another in the viewpoint of idempotency is evaluated by us. Then we have applied NCD along with k nearest neighbour as a classification algorithm to a randomly selected public corpus data with different block sizes (512 byte, 1024 bytes, 1536 bytes, 2048 bytes). The performance of two compressors bzip2 and gzip is also compared for the NCD algorithm in the perspective of idempotency.

Objectives: In this study we have investigated the combine effect of both of the parameters namely compression ratio versus idempotency and varying block size of data on the performance of NCD. The objective is to figure out that in order to have a better performance of NCD either a compressor for NCD should be selected on the basis of better compression ratio of compressors or better idempotency of compressors. The whole purpose of using different block sizes was to evaluate either the performance of NCD will improve or not by varying the block size of data to be used for making the datasets.

Methods: Experiments are performed to test the hypotheses and evaluate the effect of compression ratio versus idempotency and block size of data on the performance of NCD.

Results: The results obtained after the analysis of null hypotheses of main experiment are retained, which showed that there is no statistically significant difference on the performance of NCD when varying block size of data is used and also there is no statistically significant difference on the NCD’s performance when a compressor is selected for NCD on the basis of better compression ratio or better idempotency.

Conclusions: As the results obtained from the experiments are unable to reject the null hypotheses of main experiment so no conclusion could be drawn of the effect of the independent variables on the dependent variable i.e. there is no statistically significant effect of compression ratio versus idempotency and varying block size of data on performance of the NCD.

Keywords: Normalized Compression Distance, NCD, idempotency, compression ratio, block size of data.
# Table of contents

*Evaluation of Idempotency & Block size of Data on the performance of Normalized Compression Distance Algorithm*

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abstract</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>1 Introduction</strong></td>
<td>8</td>
</tr>
<tr>
<td>1.1 Overview</td>
<td>8</td>
</tr>
<tr>
<td>1.2 Aims and Objectives</td>
<td>9</td>
</tr>
<tr>
<td>1.3 Research Questions/Hypotheses</td>
<td>9</td>
</tr>
<tr>
<td>1.4 Research Methodology</td>
<td>11</td>
</tr>
<tr>
<td><strong>2 Technical Information</strong></td>
<td>12</td>
</tr>
<tr>
<td>2.1 File Fragment Analysis</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Classification</td>
<td>13</td>
</tr>
<tr>
<td>2.2.1 Hit Rate</td>
<td>13</td>
</tr>
<tr>
<td>2.3 File Fragment Corpus</td>
<td>14</td>
</tr>
<tr>
<td>2.4 Normalized Compression Distance (NCD)</td>
<td>14</td>
</tr>
<tr>
<td>2.5 Compression Ratio</td>
<td>15</td>
</tr>
<tr>
<td>2.6 Idempotency</td>
<td>15</td>
</tr>
<tr>
<td>2.7 Compression Ratio versus Idempotency</td>
<td>16</td>
</tr>
<tr>
<td>2.8 Compressors</td>
<td>16</td>
</tr>
<tr>
<td><strong>3 Experimentation</strong></td>
<td>18</td>
</tr>
<tr>
<td>3.1 Supporting Experiment</td>
<td>18</td>
</tr>
<tr>
<td>3.1.1 Experiment Definition</td>
<td>18</td>
</tr>
<tr>
<td>3.1.2 Experiment Planning</td>
<td>18</td>
</tr>
<tr>
<td>3.1.2.1 Context selection</td>
<td>18</td>
</tr>
<tr>
<td>3.1.2.2 Hypothesis Formulation</td>
<td>18</td>
</tr>
<tr>
<td>3.1.2.3 Experiment Variables</td>
<td>19</td>
</tr>
<tr>
<td>3.1.2.4 Experimental Dataset</td>
<td>19</td>
</tr>
<tr>
<td>3.1.2.5 Experiment Design</td>
<td>21</td>
</tr>
<tr>
<td>3.1.2.6 Experiment Instrumentation</td>
<td>22</td>
</tr>
<tr>
<td>3.1.2.7 Experiment Validity Evaluation</td>
<td>22</td>
</tr>
<tr>
<td>3.1.3 Experiment Operation</td>
<td>23</td>
</tr>
</tbody>
</table>
3.1.4 Analysis and Interpretation

3.1.4.1 Results

3.1.4.2 Hypothesis Testing

3.2 Main Experimentation

3.2.1 Experiment Definition

3.2.2 Experiment Planning

3.2.2.1 Context selection

3.2.2.2 Hypothesis Formulation

3.2.2.3 Experiment Variables

3.2.2.4 Experiment Dataset

3.2.2.5 Experiment Design

3.2.2.6 Experiment Instrumentation

3.2.2.7 Experiment Validity Evaluation

3.2.3 Experiment Operation

3.2.4 Analysis and Interpretation

3.2.4.1 Results

3.2.4.2 Hypothesis Testing

4 Discussion

5 Conclusion And Future work

6 Related Work

7 References
List of Tables

Table 1: Experimental Dataset ........................................................................................................ 20
Table 2: Results from ANOVA Test .................................................................................................. 33
Table 3: Results from ANOVA Two Factor Test .............................................................................. 34
Table 4: Experimental Dataset ........................................................................................................ 38
Table 5: Hit rate for blocksize-512 ................................................................................................. 43
Table 6: Standard deviation for blocksize-512 ................................................................................ 44
Table 7: Hit rate for blocksize-1024 ............................................................................................... 46
Table 8: Standard deviation for blocksize-1024 ............................................................................. 47
Table 9: Hit rate for blocksize-1536 ............................................................................................... 49
Table 10: Standard deviation for blocksize-1536 .......................................................................... 50
Table 11: Hit rate for blocksize-2048 .............................................................................................. 52
Table 12: Standard deviation for blocksize-2048 .......................................................................... 53
Table 13: Results from ANOVA Test .............................................................................................. 56
Table 14: Results from ANOVA two factor Test ............................................................................ 57

List of Figures

Figure 1: Experiment Design of Supporting Experiment ................................................................. 21
Figure 2: Experiment Design of Supporting Experiment ................................................................. 21
Figure 3: Experiment design of Main Experiment ......................................................................... 38
Figure 4: Experiment design of Main Experiment ......................................................................... 39
List of Graphs

Graph 1: A Sample block size 512 Java File ................................................................. 25
Graph 2: A Sample block size 1024 Java File .............................................................. 25
Graph 3: A Sample block size 1536 Java File .............................................................. 26
Graph 4: A Sample block size 2048 Java File .............................................................. 26
Graph 5: Mean of all block size 512 Java Files .............................................................. 27
Graph 6: Standard Deviation of all block size 512 Java Files ....................................... 28
Graph 7: Mean of all block size 1024 Java Files ............................................................ 28
Graph 8: Standard Deviation of all blocks size 1024 Java Files ..................................... 28
Graph 9: Mean of all block size 1536 Java Files ............................................................ 29
Graph 10: Standard Deviation of all blocks size 1536 Java Files .................................... 29
Graph 11: Mean of all block size 2048 Java Files ............................................................ 29
Graph 12: Standard Deviation of all blocks size 2048 Java Files .................................... 30
Graph 13: Summary of all Java Files for all Block sizes ................................................ 31
Graph 14: Summary of all Files of all Block sizes ......................................................... 32
Graph 15: Summary of all Files of all Block sizes ......................................................... 32
Graph 16: Hit rate for Blocksize-512 & averaged k ....................................................... 42
Graph 17: Hit rate for Blocksize-1024 & averaged k ..................................................... 45
Graph 18: Hit rate for Blocksize-1536 & averaged k ..................................................... 48
Graph 19: Hit rate for Blocksize-2048 & averaged k ..................................................... 51
Graph 20: Hit rate for All Block size with averaged k by Bzip2 ...................................... 54
Graph 21: Hit rate for All Block size with averaged k by Gzip ......................................... 55
1 INTRODUCTION

1.1 Overview

Computer forensics undergoes the process of collecting, classifying, analyzing, recovering the data in digital storage devices using the scientific knowledge for the purpose of investigation, which is very important in computer security as computers are getting involved in criminal activities more and more. A computer forensic investigator is often involved in identifying and reconstructing the files (partial complete files) from the fragments remained left after the deletion, corruption, overwriting, or damaging of the storage devices. Sometimes there is meta data available about the fragments which the investigator can use for the identification and reconstruction of the files but this is not the case all the time. So having some information or clue about the fragment types (like if fragment is of type exe, doc or txt) can assist the investigator in the reconstruction process of the files and also such file fragmentation analysis is very crucial in computer forensics for a fair investigation. [6]

There are many techniques available out there for the purpose of file fragmentation analysis; classification is one of those techniques. Mostly the file fragments with which someone have to deal with are complex n valued type (n valued means depending on more than one file types), so to that end normalized compression distance (NCD) algorithm along with a classification algorithm has been used for the purpose of classification and clustering, as NCD is parameter free too [10] so one do not have to care about the inside content of file fragments being classified which means it is also very suitable for complex n valued problems. A lot of work has been done where NCD is used for the classification purposes like work done by Axelssoon [1] where the accuracies reached 70 percent for few of the file types and very low for some of the file types. Similarly the work done by Cilibrasi [5] where he have used NCD to classify the music files, they have used three file fragment types like classical, jazz and rock and they reported that the result of the classification is reasonable but not perfect. Similarly Yoshizawa [17] assesses the degree of similarity between student essay drafts using NCD, they said that the NCD measure is fast and effective way to identify the changes in the content of essays but the quality cannot be determined by the NCD results alone.

In order to use the NCD with its maximum potential there are some factors which were left to be investigated like firstly the effect of changing block size of data (data means file fragments) and secondly the idempotency of compressors being used. The performance of NCD mainly depends upon the usage of a normalize compression algorithm [4] and idempotency is one of the factors which is used to evaluate the normality of a compression algorithm and compression ratio is one of the factor on which performance of compression algorithm depends. So it is important to investigate that in order to have better classification performance of NCD either a compression algorithm for NCD should be selected on the basis of better compression ratio or better
idempotency. In addition to that the performance of compression algorithms also have effect of block size of data being used [3]. So the purpose of this study is to evaluate the combine effect of both of the parameters namely compression ratio versus idempotency and block size of data on the performance of NCD.

The detail about all the technical terms used like File fragmentation analysis, NCD algorithm, Classification, Idempotency, Block size of Data is given in the section 2 under Technical Information.

1.2 Aims and Objectives

The whole aim of this project is as follows:

Aim:

- To evaluate the combine effect of both of the parameters namely compression ratio versus idempotency and block size of data on the performance of NCD

Objectives:

- To evaluate the compression algorithms on the basis of idempotency along with the evaluation of the effect of changing block size of data on the measured idempotencies by the compression algorithms
- To analyze and compare the compression algorithms on the basis of evaluated idempotency
- To evaluate the combine effect of both of the parameters namely compression ratio versus idempotency and block size of data on the performance of NCD
- To analyze and compare the combine effect of both the parameters compression ratio versus idempotency and block size of data on the performance of NCD

The compression algorithms, the datasets and the block size of data used are described in the section 2 under technical information and also in the section 3.1 under supporting experiment & in section 3.2 under main experiment.

1.3 Research Questions/Hypotheses

As it is stated above in aims and objectives that we are doing evaluation of the two parameters so the methodology we used here for the evaluation purposes is experimentation (methodology is described in section 1.4). We needed to have performed two experiments namely:

- Supporting Experiment for evaluating Idempotency of compressors
- Main Experiment for evaluating performance of NCD
So the hypotheses for both the experiments are given below:

**Hypotheses for Supporting Experiment**

We have two null hypotheses for the supporting experiment. The purpose of first null hypothesis is to test the compression algorithms (compressors) on the basis of idempotency and the second null hypothesis is used to test the effect of changing block size of data on the measured idempotencies by the compression algorithms. The hypotheses are:

1. **Null Hypothesis, H0**: There is no difference between the idempotencies measured by the compressors.

   *Alternate Hypothesis, H1*: There will be difference between the idempotencies measured by the compressors.

2. **Null Hypothesis, H2**: There is no effect of changing block size of data on the measured idempotencies by the compressors.

   *Alternate Hypothesis, H3*: There will be effect of changing block size of data on the measured idempotencies by the compressors.

**Hypotheses for Main Experiment**

Here again we have two null hypothesis for the main experiment. The purpose of first null hypothesis is to test either a compressor for the NCD should be used on the basis of idempotency or compression ratio for having better classification performance. The results of supporting experiment will tell us which compressor is better in idempotency and as we already know which one is better in compression ratio from the work done by Manuel [3]. The second null hypothesis is used to test the effect of changing block size of data on classification performance of the NCD by the compression algorithms. The hypotheses are:

1. **Null Hypothesis, H0**: There is no difference on NCD’s classification performance when a compressor is used for NCD algorithm on the basis of compression ratio or idempotency.

   *Alternate Hypothesis, H1*: There will be difference on NCD’s classification performance when a compressor is used for NCD algorithm on the basis of compression ratio or idempotency.

2. **Null Hypothesis, H2**: There is no difference on NCD’s classification performance by compressors when a block size of data is changed.

   *Alternate Hypothesis, H3*: There will be difference on NCD’s classification performance by compressors when a block size of data is changed.
1.4 Research Methodology

The research methodology used to fulfil the aims & objectives and to test the hypotheses is of course experimentation. The experimentation is conducted by following the guidelines given by Wohlin [16]. As stated before, we have performed two experiments for testing the hypotheses namely the supporting experiment and the main experiment. The reason of performing supporting experiment is that we want to evaluate the compressors on the basis of idempotencies which we intended to use in the main experiment. As we knew which compressor is better than other in compression ratio from the work done by Manuel [3] but for idempotency which one is better than other was needed to be evaluated.

The main experiment was then performed to test the hypotheses of main experimentation. The detail of independent and dependent variables selected, experiment design, experiment context and instrumentation used in both of the supporting and main experiments are presented in the section 3 under experimentation.
2 TECHNICAL INFORMATION

The technical terms used, the formulae and the information about Normalized Compression Distance are explained in this section:

2.1 File Fragment Analysis

According to Poisel [13] today 80 to 90% of crime cases have involved some kind of digital evidence. Digital forensics is used to investigate such crime cases and computer forensics as a part of digital forensics is used to investigate cases that involve computer devices and media. In computer forensics the investigator often come across collection of file fragments from hard drives, disk images, logs that needed to put together to reconstruct the partial complete files. Sometimes enough meta data (file headers) of files is sufficient to reconstruct the files but often a random collection of file fragments is left out. So knowing or having a clue of the types of file fragments (like fragment is exe, doc, or java etc.) really aids in the reconstruction process. The clue or information about the type of files can be provided by the file fragmentation analysis.

The whole purpose of file fragmentation analysis is to analyze the fragments in order to predict their actual class type (file type) to which they belong. There are many techniques for the purpose of file fragmentation analysis, classification is one of those available techniques (described in section 2.2). By the classification technique the fragments will be classified according to their file type (like if there are 3 file types exe, doc, and java and there are 100 fragments to be classified into these 3 file types).

For example:

File types = exe, doc, java

Fragments (to be classified) =100 (there could be n number of fragments)

Where file types is also called class type and fragments to be classified is also called number of instances. After the classification process, assume 100 fragments are classified as:

20 fragments (are classified as) = exe
30 fragments (are classified as) = doc
50 fragments (are classified as) = java

So after the classification process the 100 fragments are analyzed as: 20 fragments belong to file type exe, 30 fragments belong to file type doc and 50 fragments belong to file type java.

In our work (in section 3.2 under main experiment) the NCD along with k-nearest neighbour is applied for analyzing and classifying the file fragments.
2.2 Classification

Classification is a supervised machine learning technique which is used to classify the unknown test instances to the known class (the instances could be file fragments). When NCD is used then we have to use a classification algorithm for the purpose of classification. The NCD algorithm gives us a distant matrix (this distant matrix is based on similarity values in range of 0 to 1 described in section 2.4 under NCD). As NCD is a similarity distance metric so we have to used $k$-nearest neighbour [12] to handle the distant matrix for classification purposes.

In our work we have used $k$-nearest neighbour in main experiment section 3.2. The NCD as described in section 2.4 will give a distant matrix file for a group of file fragments, and these file fragments could be classified using the $k$-nearest neighbour on the basis of the values contained in the distant matrix file.

For example in our case there are thousands of fragments of different file type and in order to classify a test fragment the $k$-nearest neighbour will look the nearest neighbours around the test fragment and it will then classify on the basis of majority vote. The number of nearest neighbours to look for depends upon the value of the $k$, like in order to classify a fragment for the value of $k$ equal to 10, the $k$-nearest neighbour will look the 10 nearest neighbour around that test fragment and let’s say (e.g. out of 10 fragments, 6 fragments nearest around are exe, 2 fragments are java and 2 fragments are doc) then the $k$-nearest neighbour will classify a fragment as exe because it is in majority.

2.2.1 Hit Rate

In classification the number of correctly classified instances is called hit rate. E.g. as described above if there are 3 file types exe, doc, and java and there are 100 fragments to be classified into these 3 file types. (Fragments to be classified mean: to find out their actual class type or file type to which they belong)

Like assume we have:

File types = exe, doc, java

Fragments (to be classified) = 100 (there could be n number of fragments)

Where file types is also called class type and fragments to be classified is also called number of instances. After the classification process, assume 100 fragments are classified as:

20 fragments (are classified as) = exe
30 fragments (are classified as) = doc
50 fragments (are classified as) = java
After the classification process, out of 20 fragments (which are classified as) = exe, actually 15 were classified correctly as exe and 5 were belonging to doc file type, then the hit rate for the file type exe will be calculated as:

\[
\text{Hit rate (exe file type)} = \frac{\text{Number of Instances correctly classified (exe file type)}}{\text{Total number of instances (exe file type)}} = \frac{15}{20}
\]

Similarly the hit rates of the other file types can be calculated. We have used hit rate as a performance indicator in section 3.2 under main experiment.

2.3 File Fragment Corpus

For making the research reproducible and comparable we used publicly available corpora [7, 20]. The datasets made for both supporting and main experiment depend upon this corpus data. This corpus contains millions of different files which is very helpful to run the experiments randomly and repeatedly. The 28 file types which we used are: pdf, java, jpg, html, text, xls, doc, ppt, gif, ps, xml, csv, eps, gz, swf, pps, sql, png, pptx, docx, ttf, js, pub, bmp, xbm, zip, xlsx and jar.

We used these file types because in our daily life one often come across a lot these file types and also as we want to compare our results with the work done by Axelsson [1] where he used the same public corpus, and which will be helpful while comparing the results. The datasets used out of this public corpus are described in the experimentation section 3.

2.4 Normalized Compression Distance (NCD)

Normalized compression distance is based on the normalized information distance [11]; it measures similarity between two elements in the form of a distance matrix [4]. The NCD formula is given below:

\[
\text{NCD}(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}
\]

Where, \(C(x)\) is compression of \(x\),

\(C(y)\) is compression of \(y\),

\(C(xy)\) is compression of concatenation of \(x\) and \(y\),

and \(x, y\) are two file fragments.

In our work we have used NCD in main experiment in section 3.2 to calculate the distances between the file fragments, when the NCD will be applied on these fragments it will give a value
in range of 0 to 1, where 0 means the fragments are completely similar to each other and 1 means they are entirely different.

### 2.5 Compression Ratio

Compression ratio is defined as the ratio of compressed size of a file to uncompressed size of a file. It is given as:

\[
\text{Compression ratio} = \frac{\text{compressed size (x)}}{\text{uncompressed size (x)}}
\]

where x could be a file (or file fragment)

### 2.6 Idempotency

Idempotency is a metric which is used to evaluate the normality of a lossless compressor [4]. The idempotency value of a compressor tells us that compression of an empty string should be 0, and when the compression of the concatenation of a file fragment is taken it should be equal to the compression of the file fragment itself. The idempotency formula is given below:

\[
C(xx) = C(x), \quad \text{and} \quad C(\lambda) = 0
\]

Where, \( \lambda \) is the empty string,

\( C \) is compressor,

\( C(x) \) is compression of x,

\( C(xx) \) is compression of concatenation of x

and x is a file fragment.

For idempotency measurement we used the NCD’s formula described in section 2.4 as:

\[
\text{NCD}(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}
\]

The idempotency should be calculated among the same file fragments. So when both the fragments will be same then the above NCD’s formula could be turned into idempotency formula by replacing the y with x as:

\[
\text{NCD}(x, x) = \frac{C(xx) - C(x)}{C(x)}
\]

This can be used for the idempotency measurement.
In our work we have used this derived idempotency measurement formula in the supporting experiment section 3.1 to evaluate the idempotency of the three compressors bzip2, gzip and ppmd.

2.7 Compression Ratio versus Idempotency

- The compression ratio of a compressor is the ability to compress a file in order to have a smaller size.

For example: Consider a file x having a size 100MB.

(Then using the above formula from section 2.5)

\[ \text{Compression ratio} = \frac{\text{compressed size (x)}}{\text{uncompressed size (x)}} \]

Assume after compressing, the size of file x is 50MB.

\[ \text{Compression ratio} = \frac{50}{100} \]

But assume a second instance that after compressing, the size of file x is also 100MB.

\[ \text{i.e. Compression ratio} = \frac{100}{100} \]

➢ Which means compressor is unable to compress or reduce the size of file x.

- The Idempotency of a compressor is the ability to compress a file in order to find the redundant information in that file.

For example: Consider a file x having a size 100MB.

(Then using the above formula from section 2.6)

\[ C(xx) = C(x) \]

Assume after compressing, the size of C(x) of file x is 50MB & then file is concatenated with itself and we got xx, then the size of C(xx) is also 50MB.

\[ \text{i.e. } 50 = 50 \]

➢ Which means the compressor is able to find out all the redundant information.

2.8 Compressors

The computation of Normalized compression distance entirely depends upon the compressor to be used. The three compressors which we used in our work for NCD are:

1. PPMZ (we have used PPMD which is an extension to PPMZ)
2. Bzip2
3. Gzip

The reason of choosing these three compressors is that they are open source data compressor and as we know the compression ratio of PPMZ is better than Bzip2 and the compression ratio of Bzip2 is better than Gzip [3], but we do not know which compressor among these three is better than other in idempotency, so we have performed supporting experiment in order to find that. The purpose of evaluating compressors on the basis of idempotency is that if a compressor is used for the NCD which is better than other compressors in idempotency it can produce good results rather than using a compressor for NCD which is better than other compressors in compression ratio.
3 EXPERIMENTATION

3.1 Supporting Experiment

3.1.1 Experiment Definition

Analyze the bzip2, gzip and ppmd compressors for the purpose of evaluation of their idempotencies with respect to the effectiveness of compression ratio and block size of data from point of view of the researcher in the context of M.Sc. students on laptop computers.

Object of Study: The objects studied are the bzip2, gzip and ppmd compressors.

Purpose: The purpose is to evaluate the idempotencies of bzip2, gzip and ppmd compressors.

Quality focus: The quality focus is effectiveness of compression ratio of bzip2, gzip and ppmd compressors and effectiveness of block size of data in producing idempotency values.

Perspective: The perspective is from researcher’s point of view.

Context: The experiment is run using M.Sc. students as subjects on laptop computers.

3.1.2 Experiment Planning

3.1.2.1 Context selection

The experiment is run on laptop computers to measure the idempotencies of the compressors using the sh shell scripts.

- Laptop’s Configuration
  
  Processors : Intel(R) Core(TM) i3 CPU M 350 @ 2.27GHz & Intel(R) Core(TM) i3 CPU M 330 @ 2.13GHz
  
  Memory (RAM): 4 GB & 2GB
  
  Disk Space: 320 GB & 320 GB
  
  Operating System: Bunt 11.04

3.1.2.2 Hypothesis Formulation

The hypotheses are presented here in more precise form.

1. Null Hypothesis, H0: There is no difference between the idempotencies measured by the bzip2, gzip and ppmd.
Alternate Hypothesis, H1: There will be difference between the idempotencies measured by the bzip2, gzip and ppmd.

2. Null Hypothesis, H2: There is no effect of changing block size of data on the measured idempotencies by the bzip2, gzip and ppmd.

Alternate Hypothesis, H3: There will be effect of changing block size of data on the measured idempotencies by the bzip2, gzip and ppmd.

The hypotheses mean that we have to collect the following data:

- Idempotency values measured by the bzip2, gzip and ppmd
- Idempotency values measured by the bzip2, gzip and ppmd when the block size of data will be changed

3.1.2.3 Experiment Variables

The experiment variables used in the supporting experiment are:

- Independent Variable
  Independent variable1: Compressor
  Independent variable2: block size

- Dependent Variable
  Dependent variable: Idempotency (measured by compressors)

3.1.2.4 Experimental Dataset

The dataset used in this experiment out of the corpus described before is given in the table below. The experiment is performed for all of these four block sizes 512, 1024, 1536 and 2048. The number of files used per each file type is 50.

<table>
<thead>
<tr>
<th>File Ending</th>
<th>File Type</th>
<th>Number of Files (for block size in kilobytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>.pdf</td>
<td>Portable document format</td>
<td>50 50 50 50</td>
</tr>
<tr>
<td>.html</td>
<td>Hypertext markup language</td>
<td>50 50 50 50</td>
</tr>
<tr>
<td>.jpg</td>
<td>Joint Photographic Experts</td>
<td>50 50 50 50</td>
</tr>
<tr>
<td>.txt</td>
<td>Text Files (Dos/Windows)</td>
<td>50 50 50 50</td>
</tr>
<tr>
<td>File Extension</td>
<td>Desktop Application</td>
<td>Column 1</td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------</td>
<td>----------</td>
</tr>
<tr>
<td>.doc</td>
<td>Microsoft Word (Various vs)</td>
<td>50</td>
</tr>
<tr>
<td>.xls</td>
<td>Microsoft Excel (various vs)</td>
<td>50</td>
</tr>
<tr>
<td>.ppt</td>
<td>Microsoft PowerPoint</td>
<td>50</td>
</tr>
<tr>
<td>.xml</td>
<td>Extensible Markup Language</td>
<td>50</td>
</tr>
<tr>
<td>.gif</td>
<td>Graphics Interchange Format</td>
<td>50</td>
</tr>
<tr>
<td>.ps</td>
<td>PostScript</td>
<td>50</td>
</tr>
<tr>
<td>.csv</td>
<td>Comma Separated Values</td>
<td>50</td>
</tr>
<tr>
<td>.gz</td>
<td>GNU zip compression</td>
<td>50</td>
</tr>
<tr>
<td>.eps</td>
<td>Encapsulated PostScript</td>
<td>50</td>
</tr>
<tr>
<td>.png</td>
<td>Portable Network Graphics</td>
<td>50</td>
</tr>
<tr>
<td>.swf</td>
<td>Shockwave Flash</td>
<td>50</td>
</tr>
<tr>
<td>.pps</td>
<td>Microsoft PowerPoint Show</td>
<td>50</td>
</tr>
<tr>
<td>.sql</td>
<td>Sql database Scripts and Dumps</td>
<td>50</td>
</tr>
<tr>
<td>.java</td>
<td>Java Source Code</td>
<td>50</td>
</tr>
<tr>
<td>.pptx</td>
<td>Microsoft PowerPoint (xml format)</td>
<td>50</td>
</tr>
<tr>
<td>.docx</td>
<td>Microsoft Word (Xml format)</td>
<td>50</td>
</tr>
<tr>
<td>.ttf</td>
<td>True type font</td>
<td>50</td>
</tr>
<tr>
<td>.js</td>
<td>JavaScript source code</td>
<td>50</td>
</tr>
<tr>
<td>.pub</td>
<td>SSH public key files</td>
<td>50</td>
</tr>
<tr>
<td>.bmp</td>
<td>Bitmap pictures</td>
<td>50</td>
</tr>
<tr>
<td>.xbm</td>
<td>X Bitmap</td>
<td>50</td>
</tr>
<tr>
<td>.xlsx</td>
<td>Microsoft Excel (Xml format)</td>
<td>50</td>
</tr>
<tr>
<td>.jar</td>
<td>Java Archive</td>
<td>50</td>
</tr>
<tr>
<td>.zip</td>
<td>Data Compression and archiving</td>
<td>50</td>
</tr>
</tbody>
</table>

**Total:** 1400 1400 1400 1400

Table 1: Experimental Dataset
3.1.2.5 Experiment Design

The experimental design can be understood from the figure 1 and figure 2 below:

![Figure 1: Experiment Design of Supporting Experiment](image)

In this experiment we are studying the effects of compression ratio of Compressors and Block size of data on the idempotency. There are two factors:

**Factor A:** Compressors

**Factor B:** Block size of Data

The compressors have three levels – Bzip2, gzip, ppmd

The block size of data has four levels – 512, 1024, 1536, 2048

<table>
<thead>
<tr>
<th>Block size of Data</th>
<th>Compressors</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bzip2</td>
<td>512 bytes</td>
<td>Treatment 1</td>
<td>Treatment 3</td>
</tr>
<tr>
<td></td>
<td>1024 bytes</td>
<td>Treatment 4</td>
<td>Treatment 6</td>
</tr>
<tr>
<td></td>
<td>1536 bytes</td>
<td>Treatment 7</td>
<td>Treatment 9</td>
</tr>
<tr>
<td></td>
<td>2048 bytes</td>
<td>Treatment 10</td>
<td>Treatment 12</td>
</tr>
</tbody>
</table>

**Figure 2: Experiment Design of Supporting Experiment**

- Treatment1(bzip2, 512), Treatment4 (bzip2, 1024), Treatment7 (bzip2, 1536), Treatment10 (bzip2, 2048)
- Treatment2 (gzip, 512), Treatment5 (gzip, 1024), Treatment8 (gzip, 1536), Treatment11 (gzip, 2048)
- Treatment3 (ppmd, 512), Treatment6 (ppmd, 1024), Treatment9 (ppmd, 1536), Treatment12 (ppmd, 2048)
• All these treatments are given to the derived NCD formula (discussed in section 2.5) for calculating the idempotency values by the three compressors.
• ANOVA (ANalysis Of VAriance) is used for testing

3.1.2.6 Experiment Instrumentation

The following required instruments for the experiment are selected and made ready:

• Compressors (bzip2,gzip,ppmd)
• Implementation of NCD for Idempotency
• Dataset of Block sizes (512,1024,1536,2048)
• Scripts are developed to perform experiment

3.1.2.7 Experiment Validity Evaluation

In our particular case, we have several levels of validity to consider.

Internal Validity: (treatment causes the outcome)

The experiment is run one time, but a large number of dataset files are used, their results can be compared with one another other to ensure that the treatment causes the outcome.

Conclusion Validity: (relationship b/w treatment and outcome)

As our experimental design is two factors with more than two treatments so we have used ANOVA for statistical testing to draw the correct conclusion relationship between the treatment and outcome. And we have used proper instrumentation to get the reliable measures.

Construct Validity:

The theory is made clear and proper constructs are defined so that we they are translated to measure and treatments we can draw the treatment reflects the construct of the cause.

External Validity:

This threat is critical because as we are selecting the samples from the population, so we have to take care of it while generalizing the results.
3.1.3 Experiment Operation

The experiment is executed on laptop computers for over a week, during which idempotency data measured by the three compressors for each of the file types is collected. The data is collected in text files and is validated properly.
3.1.4 Analysis and Interpretation

We have used descriptive statistics to present the results, and we have used ANOVA test for testing the hypotheses.

3.1.4.1 Results

3.1.4.1.1 Descriptive Statistics

As a first step in analyzing the data, we used descriptive statistics to visualize the data collected. As there are space limits so we have presented just a few samples of the data and rest of the data is available at [18]. We have start presenting the results for individual files, then for single file type and then for all the file types.

3.1.4.1.1.1 By Individual Files

Firstly the results of individual files by the three compressors bzip2, gzip and ppmd are presented. The purpose of showing the graphs for individual file is to give an overview of how the three compressors performed for the individual files. The Graph1 shown below is a sample representation of a java file for block size 512. In this graph, x-axis is representing number of fragments, y-axis is representing idempotency values and colours are representing the three compressors.

The number of fragments on x-axis depends upon the size of a file (i.e. File1 split by block size 512 bytes shown in Graph 1, the same File1 when split by block size 1024 bytes in Graph 2, the same File1 split by block size 1536 bytes in Graph 3, the same File1 split by block size 2048 bytes in Graph 4).

The Idempotency values on y-axis are based on the calculations performed by using the Idempotency formula (described in section 2.6). According to that formula idempotency values will lie in range of 0 to 1, where 0 means the compressor is Idempotent and 1 means it is non idempotent. But here in our case none of the compressors is idempotent as we can see in the graphs 1, 2, 3 & 4 (because none of the compressors is giving idempotency value equal to 0).

In Graph 1, the idempotency values of gzip compressor for a sample Java file (with block size 512) is in range of 0.04 to 0.05, the idempotency values of bzip2 compressor for a sample Java file (with block size 512) is in range of 0.06 to 0.24 and the idempotency values of ppmd compressor for a sample Java file (with block size 512) is in range of 0.12 to 0.25.
The performance of the three compressors can be visualized by the coloured lines, the more closer a line to zero the better the performance of the compressor will be. If a line lie at zero then it means the compressor idempotency is zero and it is idempotent compressor.

As no one of compressor is idempotent but which compressor among the three is better in idempotency can be visualized by comparing the idempotency values graphs. According to the graph gzip's idempotency values are closest to zero, then bzip2 whose idempotency values are overlapping with ppmd but still better than ppmd and then the idempotency values of ppmd lie at last.

The graphs of block sizes 1024, 1536 and 2048 of the same java file are shown in the Graph 2, Graph 3 and Graph 4 accordingly. The effect of increasing the block size (from 512 to 1024, 1536 & 2048) on the idempotency can also be visualized, which is not very significant here.
Graph 3: A Sample block size 1536 Java File (x-axis: Number of Fragments of File, y-axis: Idempotency)

Graph 4: A Sample block size 2048 Java File (x-axis: Number of Fragments of File, y-axis: Idempotency)

### 3.1.4.1.1.2 By All Java Files

The idempotency values measured by the three compressors of all the java files for block size 512 can be visualized below in Graph 5. The purpose of this graph is to give an overview of how the three compressors performed for the files of a whole file type (like java). In this graph, x-axis is representing number of files (the 50 number of the java files which were used in the supporting experiment), y-axis is representing idempotency values and colours are representing the three compressors.

On x-axis, the performance of each Java file by the three compressors can be visualized. The Idempotency values on y-axis are based on the calculations performed by using the Idempotency formula (described in section 2.6). According to that formula idempotency values will lie in range of 0 to 1, where 0 means the compressor is Idempotent and 1 means it is non idempotent. But here in our case none of the compressors (for all the files of java file type) is idempotent as
we can see in the graphs 5, 7, 9 & 11 (because none of the compressors is giving idempotency value equal to 0).

The performance of the three compressors can be visualized by the coloured vertical columns, the smaller and closer a column to zero is the better the performance of the compressor will be. According to the graph gzip’s idempotency values (for all java files) are best and closest to zero, then bzip2 whose idempotency values are better than ppmd and then the idempotency values of ppmd comes at last.

Graph 5: Mean of all block size 512 Java Files (x-axis: Number of Files, y-axis: averaged Idempotency)

The standard deviation is also shown below in Graph 6, which is showing the deviation in the measured idempotency values by the three compressors of all java files for block size 512. According to the graph the deviation by gzip’s idempotency is comparatively smallest and closest to zero, then ppmd and bzip2.

The graphs of mean and standard deviation of all java files for block sizes 1024, 1536 and 2048 can be visualized below (in the similar way as described above) in the Graph 7, Graph 8, Graph 9, Graph 10, Graph 11 and Graph 12 accordingly. One thing is common among all these graphs that the performance of gzip compressor is better than the rest of compressors.
Graph 6: Standard Deviation of all block size 512 Java Files (x-axis: Number of Files, y-axis: Average Idempotency)

Graph 7: Mean of all block size 1024 Java Files (x-axis: Number of Files, y-axis: Average Idempotency)

Graph 8: Standard Deviation of all blocks size 1024 Java Files (x-axis: Number of Files, y-axis: Average Idempotency)
Graph 9: Mean of all block size 1536 Java Files (x-axis: Number of Files, y-axis: Average Idempotency)

Graph 10: Standard Deviation of all blocks size 1536 Java Files (x-axis: Number of Files, y-axis: Average Idempotency)

Graph 11: Mean of all block size 2048 Java Files (x-axis: Number of Files, y-axis: Average Idempotency)
Graph 12: Standard Deviation of all blocks size 2048 Java Files (x-axis: Number of Files, y-axis: Average Idempotency)

3.1.4.1.3 Overall Summarized Statistics of all Java Files for all Block sizes

A summary of mean of average idempotency values and mean of standard deviation of idempotency values of all the java files for all block sizes is shown below in the Graph 13. The purpose of this graph is to give an overview of the effect of increasing the block size of data on the performance of compressors (in order to do that mean of averages is used so that in a less space we can provide a visualization).

The x-axis here is representing the mean of average Idempotency values (on left side) and mean of standard deviation of idempotency values (on right side) of all the 50 Java files used. The different colours are representing the four block sizes.

The Idempotency values on y-axis are based on the calculations performed by using the Idempotency formula (described in section 2.6). According to that formula idempotency values will lie in range of 0 to 1, where 0 means the compressor is Idempotent and 1 means it is non idempotent. But here in our case none of the compressors is idempotent as we can see in the graph 13 (because none of the compressors is giving idempotency value equal to 0).

From this Graph, we can analyse the overall results of the three compressors when the block size is changed from 512 bytes to 1024 bytes, 1536 bytes & 2048 bytes. As a whole the performance of gzip is better than other two compressors. The gzip compressor is more consistent in performance as compared to the other compressors when the block size is increased. The bzip2 is giving a bit fluctuation as compared to gzip and the performance of ppmd is going down (mean its idempotency is becoming worse) when the block size is increased.
Graph 13: Summary of all Java Files for all Block sizes (x-axis: Compressors with block sizes, y-axis: mean Idempotency of all Java Files)

3.1.4.1.1.4 Overall Summarized Statistics of all File Types

The overall summary of mean of average idempotency values and the mean of standard deviation of the idempotency values of all the file types used are shown in the Graph 14 and Graph 15 below. The purpose of these graphs is to give an overview of how the three compressors performed for all the 28 file types (in order to do that mean of averages is used so that in a less space we can provide a visualization of the results of all 28 file types).

The x-axis here is representing the 28 file types (each individual file type out of 28 file types here is representing the mean of the average idempotency values of all the 50 files used). The colours are representing the three compressors bzip2, gzip and ppmd.

The Idempotency values on y-axis are based on the calculations performed by using the Idempotency formula (described in section 2.6). According to that formula idempotency values will lie in range of 0 to 1, where 0 means the compressor is Idempotent and 1 means it is non idempotent. But here in our case none of the compressors is idempotent (for any of the 28 file types) as we can see in the graph 14 (because none of the compressors is giving idempotency value equal to 0).

The performance of gzip is better than bzip2 and ppmd for most of the file types but for some file types like png, gz, pps, pptx, gif and swf the performance of ppmd is better than the gzip and bzip2.

Similarly (in Graph 15) the gzip is giving less deviation as compared to bzip2 and ppmd for most of the file types but for some file types like png, gz and gif the ppmd is giving less deviation as compared to gzip and bzip2.
Graph 14: Summary of all Files of all Block sizes (x-axis: File Types, y-axis: mean Idempotency of all Files)

Graph 15: Summary of all Files of all Block sizes (x-axis: File Types, y-axis: Std. Deviation in Idempotency of all Files)
3.1.4.2 Hypothesis Testing

The hypothesis testing is performed on the collected data for supporting experiment. The first hypothesis H0 regarding there is no difference between the idempotencies measured by the bzip2, gzip and ppmd is evaluated using ANOVA test. The second hypothesis H2 regarding there is no effect of changing block size of data on the measured idempotencies by the bzip2, gzip and ppmd is evaluated using ANOVA two factor tests. The reason of using ANOVA is that as we have to independent variables so it is more appropriate to test the hypotheses using ANOVA test.

1. Null Hypothesis, H0: There is no difference between the idempotencies measured by the bzip2, gzip and ppmd.

The results from ANOVA test for this hypothesis are shown below in table 2:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>0.317083</td>
<td>2</td>
<td>0.158542</td>
<td>63.45845</td>
<td>2.62E-17</td>
<td>3.109311</td>
</tr>
<tr>
<td>Within Groups</td>
<td>0.202367</td>
<td>81</td>
<td>0.002498</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>0.51945</td>
<td>83</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results from ANOVA Test

- From the results in table 2 above, as F is greater than F crit and as P-value is less than 0.05, so we will reject the null hypothesis and it is significant. I.e. we will reject significantly that there is no difference between the idempotencies measured by the bzip2, gzip and ppmd.

2. Null Hypothesis, H2: There is no effect of changing block size of data on the measured idempotencies by the bzip2, gzip and ppmd.

As here we have two factors with different levels of treatments so we will use factorial ANOVA and the hypothesis will be break down as:

Case-1) $H0$, Block size original = Block size Changed

$H1$, Block size original ≠ Block size Changed

Case-2) $H2$, Bzip2 = Gzip = Ppmd

$H3$, Bzip2 ≠ Gzip ≠ Ppmd
Case-3) $H4$, an interaction is absent

$H5$, an interaction is present

The results from ANOVA test for this hypothesis are shown below in table 3:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>df</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>$F$</th>
<th>$P$-value</th>
<th>$F$ crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>3</td>
<td>0.002409</td>
<td></td>
<td>0.000803</td>
<td>0.232264</td>
<td>0.873882</td>
<td>2.63248</td>
</tr>
<tr>
<td>Columns</td>
<td>2</td>
<td>0.936022</td>
<td>0.468011</td>
<td>135.3883</td>
<td>1.83E-43</td>
<td>3.023603</td>
<td>3.023603</td>
</tr>
<tr>
<td>Interaction</td>
<td>6</td>
<td>0.066441</td>
<td>0.011073</td>
<td>3.20338</td>
<td>0.00454</td>
<td>2.126597</td>
<td>2.126597</td>
</tr>
<tr>
<td>Within</td>
<td>324</td>
<td>1.120005</td>
<td>0.003457</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>335</td>
<td>2.124877</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Results from ANOVA Two Factor Test

From the table 3 the results are as:

Case-1) $H0$ is retained.

As in above table 3 in sample row, $F$ is less than $F$ crit and $p$-value is greater than 0.05, $H0$ will retain and it is not significant. I.e. we cannot accept that there is significant effect of changing block size of data on the measured idempotencies by the bzip2, gzip and ppm.

Case-2) $H2$ will be rejected.

As in above table in Columns row, $F$ is greater than $F$ crit and $p$-value is less than 0.05, $H0$ will be rejected. So there is difference between the idempotencies measured by the bzip2, gzip and ppm.

Case-3) $H4$ will be rejected.

As in above table in Interaction row, $F$ is greater than $F$ crit but $p$-value is greater than 0.05, $H0$ will be rejected but it is not significant. I.e. Interaction is present but it is not significant.
3.2 Main Experimentation

3.2.1 Experiment Definition

Analyze the bzip2 and gzip compressors for the purpose of evaluation of NCD’s classification performance with respect to the effectiveness of idempotency of compressors, and block size of data from point of view of the researcher in the context of M.Sc. students on university’s server.

Object of Study: The objects studied are the bzip2 and gzip compressors.

Purpose: The purpose is to evaluate the classification performance of NCD.

Quality focus: The quality focus is effectiveness of idempotency of bzip2 and gzip when used as compressors in NCD algorithm and effectiveness of block size of data when fed to compressors on the results of NCD classification performance.

Perspective: The perspective is from researcher’s point of view.

Context: The experiment is run using M.Sc. students as subjects on university’s server.

3.2.2 Experiment Planning

3.2.2.1 Context selection

The experiment is run on university’s server to calculate the distance matrices using NCD. The python & sh shell scripts were used as implementation languages.

- Server’s Configuration
- CPU: Dual Core Amd Opteron(tm) Processor 2214 HE
- Memory (RAM): 12 GB
- Server: Linux Harold 2.6.32-41-generic
- Operating System: Ubuntu 10.04

3.2.2.2 Hypothesis Formulation

The hypotheses are presented here in more precise form:

1. Null Hypothesis, H0: There is no difference on NCD’s classification performance when a compressor bzip2 or gzip is used for NCD algorithm on the basis of compression ratio or idempotency.
**Alternate Hypothesis, H1:** There will be difference on NCD’s classification performance when a compressor bzip2 or gzip is used for NCD algorithm on the basis of compression ratio or idempotency.

2. **Null Hypothesis, H2:** There is no difference on NCD’s classification performance by bzip2 & gzip when a block size of data is changed.

**Alternate Hypothesis, H3:** There will be difference on NCD’s classification performance by bzip2 & gzip when a block size of data is changed.

The hypotheses mean that we have to collect the following data:

- NCD’s classification results by bzip2 and gzip
- NCD’s classification results by bzip2 and gzip when block size is changed
- Results from experiment 5.1 of Idempotency values measured by the bzip2 and gzip

### 3.2.2.3 Experiment Variables

The experiment variables are:

- **Independent Variable**
  - Independent variable1: Compressor
  - Independent variable2: block size

- **Dependent Variable**
  - Dependent variable: Classification Performance (of NCD in terms of hit and miss)
3.2.2.4 Experiment Dataset

The experiment dataset used out of the corpus as described before follows the example given by [1] as we wanted to compare our results with the results of that work. The dataset used in this experiment is given in the table 4 below. The block size 512 is used as a referenced block size and the experiment is performed for block size 512 in first place, then for block size 1024, then for 1536 and then for 2048.

<table>
<thead>
<tr>
<th>File Ending</th>
<th>File Type</th>
<th>Number of Fragments (for block size in bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>512</td>
</tr>
<tr>
<td>.pdf</td>
<td>Portable document format</td>
<td>390</td>
</tr>
<tr>
<td>.html</td>
<td>Hypertext markup language</td>
<td>401</td>
</tr>
<tr>
<td>.jpg</td>
<td>Joint Photographic Experts Group Image</td>
<td>372</td>
</tr>
<tr>
<td>.txt</td>
<td>Text Files (Dos/Windows)</td>
<td>297</td>
</tr>
<tr>
<td>.doc</td>
<td>Microsoft Word (Various vs)</td>
<td>406</td>
</tr>
<tr>
<td>.xls</td>
<td>Microsoft Excel (various vs)</td>
<td>347</td>
</tr>
<tr>
<td>.ppt</td>
<td>Microsoft PowerPoint</td>
<td>322</td>
</tr>
<tr>
<td>.xml</td>
<td>Extensible Markup Language</td>
<td>387</td>
</tr>
<tr>
<td>.gif</td>
<td>Graphics Interchange Format</td>
<td>349</td>
</tr>
<tr>
<td>.ps</td>
<td>PostScript</td>
<td>196</td>
</tr>
<tr>
<td>.csv</td>
<td>Comma Separated Values</td>
<td>306</td>
</tr>
<tr>
<td>.gz</td>
<td>GNU zip compression</td>
<td>206</td>
</tr>
<tr>
<td>.eps</td>
<td>Encapsulated PostScript</td>
<td>474</td>
</tr>
<tr>
<td>.png</td>
<td>Portable Network Graphics</td>
<td>82</td>
</tr>
<tr>
<td>.swf</td>
<td>Shockwave Flash</td>
<td>248</td>
</tr>
<tr>
<td>.pps</td>
<td>Microsoft PowerPoint Show</td>
<td>448</td>
</tr>
<tr>
<td>.sql</td>
<td>Sql database Scripts and Dumps</td>
<td>248</td>
</tr>
<tr>
<td>.java</td>
<td>Java Source Code</td>
<td>317</td>
</tr>
<tr>
<td>.pptx</td>
<td>Microsoft PowerPoint (xml format)</td>
<td>280</td>
</tr>
</tbody>
</table>
3.2.2.5 Experiment Design

The experimental design can be understood from the figure 3 and figure 4 below:

![Figure 3: Experiment design of Main Experiment](image)

In this experiment we are studying the effects of idempotency of Compressors and Block size of data on NCD’s classification performance. There are two factors:

**Factor A:** Compressors

**Factor B:** Block size of Data

The compressors have two levels – Bzip2, gzip

The block size of data has four levels – 512, 1024, 1536, 2048
<table>
<thead>
<tr>
<th>Block size Of Data</th>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td>512 bytes</td>
<td>Treatment 1</td>
<td>Treatment 2</td>
</tr>
<tr>
<td>1024 bytes</td>
<td>Treatment 3</td>
<td>Treatment 4</td>
</tr>
<tr>
<td>1536 bytes</td>
<td>Treatment 5</td>
<td>Treatment 6</td>
</tr>
<tr>
<td>2048 bytes</td>
<td>Treatment 7</td>
<td>Treatment 8</td>
</tr>
</tbody>
</table>

Figure 4: Experiment design of Main Experiment

- Treatment1 (bzip2, 512), Treatment3 (bzip2, 1024), Treatment5 (bzip2, 1536), Treatment7 (bzip2, 2048)
- Treatment2 (gzip, 512), Treatment4 (gzip, 1024), Treatment6 (gzip, 1536), Treatment8 (gzip, 2048)
- All these treatments are given to the NCD (discussed in section 2.4) for calculating the distance matrices by the two compressors.
- ANOVA (ANalysis Of VAriance) is used for testing

3.2.2.6 Experiment Instrumentation

The following required instruments for the experiment are selected and made ready:

- Complearn NCD tool [19] for (bzip2, gzip)
- Dataset of Block sizes (512, 1024, 1536, 2048)
- Scripts are developed to perform experiment

3.2.2.7 Experiment Validity Evaluation

In our particular case, we have several levels of validity to consider.

**Internal Validity**: (treatment causes the outcome)

The experiment is run for three trails to ensure that the treatment causes the outcome.

**Conclusion Validity**: (relationship between treatment and outcome)

As our experimental design is two factors with more than two treatments so we have used ANOVA for statistical testing to draw the correct conclusion relationship between the treatment and outcome. And we have used proper instrumentation to get the reliable measures.
**Construct Validity:**

The theory is made clear and proper constructs are defined so that we they are translated to measure and treatments we can draw the treatment reflects the construct of the cause.

**External Validity:**

This threat is critical because as we are selecting the samples from the population and we have replicated the experiment three times so that we can generalize the results.

### 3.2.3 Experiment Operation

The experiment was executed on university’s server for over 40 minutes for each randomly selected dataset by the two compressors (bzip2 and gzip), during which the data is collected in text files and is validated properly.
3.2.4 Analysis and Interpretation

The experiment analysis and interpretation is presented below and then ANOVA test is performed for testing the hypotheses of the main experiment.

3.2.4.1 Results

3.2.4.1.1 Descriptive Statistics

As a first step in analyzing the data, we used descriptive statistics to visualize the data collected. We have start presenting the results individually by the four block sizes and then for the two compressors.

3.2.4.1.1.1 Results for Blocksize-512

The overall results of bzip2 & gzip for performance of NCD can be visualized from the Graph 16 below, the performance of both bzip2 and gzip for each file type is given there.

In graph, the x-axis is representing all the 28 file types and the y-axis is representing the percentage hit rate (hit rate is used for expressing the performance) and the colours are representing the two compressors. The gzip hit rate for the file type pdf is better than the hit rate of bzip2. Both the compressors are giving good results for csv, eps, java and bmp file types and they are showing hit rates more than 60 percent for these file types. But the hit rate of gzip for java & bmp file types is better than bzip2 and the hit rate of bzip2 for csv & eps file types is better than gzip.

The thing that should be kept in mind while reading the results from the graph is that the bzip2 is better than gzip in compression ratio and gzip is better than bzip2 in idempotency. As we can see that the performance of gzip is better than bzip2 for many file types so it seems that a compressor which is better in idempotency can produce good results for the NCD instead of compression ratio. The results of all the file types can be visualized in the similar way.

The results of bzip2 & gzip for different k values (0 to 10) are given in table 5 and the standard deviations of both the compressors are also given in table 6.
Graph 16: Hit rate for Blocksize-512 & averaged k (x-axis: File Types, y-axis: Hit rate)

<table>
<thead>
<tr>
<th></th>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>k</strong></td>
<td><strong>1</strong></td>
<td><strong>2</strong></td>
</tr>
<tr>
<td></td>
<td><strong>pdf</strong></td>
<td><strong>html</strong></td>
</tr>
<tr>
<td><strong>1</strong></td>
<td>25.1</td>
<td>26.9</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>28.8</td>
<td>29.1</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>32.8</td>
<td>33.8</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td>36.7</td>
<td>37.2</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>40.9</td>
<td>41.3</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>45.3</td>
<td>45.7</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>49.8</td>
<td>50.3</td>
</tr>
<tr>
<td><strong>8</strong></td>
<td>54.0</td>
<td>54.5</td>
</tr>
<tr>
<td><strong>9</strong></td>
<td>58.0</td>
<td>58.5</td>
</tr>
<tr>
<td><strong>10</strong></td>
<td>62.0</td>
<td>62.5</td>
</tr>
</tbody>
</table>

Legend: bzip2, gzip
<table>
<thead>
<tr>
<th>File Type</th>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF</td>
<td>10.4</td>
<td>3.8</td>
</tr>
<tr>
<td>HTML</td>
<td>10.9</td>
<td>5.1</td>
</tr>
<tr>
<td>JPG</td>
<td>0.43</td>
<td>7</td>
</tr>
<tr>
<td>TXT</td>
<td>5.47</td>
<td>7.48</td>
</tr>
<tr>
<td>DOC</td>
<td>16.5</td>
<td>1.51</td>
</tr>
<tr>
<td>XLS</td>
<td>0.64</td>
<td>1.69</td>
</tr>
<tr>
<td>PPT</td>
<td>3.35</td>
<td>3.52</td>
</tr>
<tr>
<td>XLM</td>
<td>15.0</td>
<td>8.54</td>
</tr>
<tr>
<td>GIF</td>
<td>8.64</td>
<td>8</td>
</tr>
<tr>
<td>PPS</td>
<td>19.2</td>
<td>5.68</td>
</tr>
<tr>
<td>CIV</td>
<td>12.8</td>
<td>9.64</td>
</tr>
<tr>
<td>GZ</td>
<td>2.46</td>
<td>4.2</td>
</tr>
<tr>
<td>EPS</td>
<td>17.0</td>
<td>5.63</td>
</tr>
<tr>
<td>PNG</td>
<td>5.22</td>
<td>6</td>
</tr>
<tr>
<td>SVG</td>
<td>9.93</td>
<td>5.4</td>
</tr>
<tr>
<td>PP</td>
<td>7.73</td>
<td>8.68</td>
</tr>
<tr>
<td>SQL</td>
<td>8.86</td>
<td>5.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>File Type</th>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDF</td>
<td>10.4</td>
<td>3.8</td>
</tr>
<tr>
<td>HTML</td>
<td>10.9</td>
<td>5.1</td>
</tr>
<tr>
<td>JPG</td>
<td>0.43</td>
<td>7</td>
</tr>
<tr>
<td>TXT</td>
<td>5.47</td>
<td>7.48</td>
</tr>
<tr>
<td>DOC</td>
<td>16.5</td>
<td>1.51</td>
</tr>
<tr>
<td>XLS</td>
<td>0.64</td>
<td>1.69</td>
</tr>
<tr>
<td>PPT</td>
<td>3.35</td>
<td>3.52</td>
</tr>
<tr>
<td>XLM</td>
<td>15.0</td>
<td>8.54</td>
</tr>
<tr>
<td>GIF</td>
<td>8.64</td>
<td>8</td>
</tr>
<tr>
<td>PPS</td>
<td>19.2</td>
<td>5.68</td>
</tr>
<tr>
<td>CIV</td>
<td>12.8</td>
<td>9.64</td>
</tr>
<tr>
<td>GZ</td>
<td>2.46</td>
<td>4.2</td>
</tr>
<tr>
<td>EPS</td>
<td>17.0</td>
<td>5.63</td>
</tr>
<tr>
<td>PNG</td>
<td>5.22</td>
<td>6</td>
</tr>
<tr>
<td>SVG</td>
<td>9.93</td>
<td>5.4</td>
</tr>
<tr>
<td>PP</td>
<td>7.73</td>
<td>8.68</td>
</tr>
<tr>
<td>SQL</td>
<td>8.86</td>
<td>5.56</td>
</tr>
</tbody>
</table>

Table 5: Hit rate for blocksize-512
df is better than the hit -n bzip2 and the hit rate of bzip2 for bmp file type is better than the deviations of both the compressors are also given in table 8. The results of bzip2 & gzip for different k values (0 to 10) are given in table 7 and the standard compression ratio. The results of all the file types can be visualized in the similar way.

As we can see that the performance of gzip is better than bzip2 for many file types so it seems that the performance which is better in idempotency should be selected for the NCD instead of compression ratio. The results of all the file types can be visualized in the similar way.

The results of bzip2 & gzip for different k values (0 to 10) are given in table 7 and the standard deviations of both the compressors are also given in table 8.

Table 6: Standard deviation for blocksize-512

3.2.4.1.1.2 Results for Blocksize-1024

The overall results of bzip2 & gzip on NCD’s performance can be visualized from the Graph 17 below, the performance of both bzip2 and gzip for each file type is given there.

In graph, the x-axis is representing all the 28 file types and the y-axis is representing the percentage hit rate (hit rate is: success rate used for expressing the performance) and the colours are representing the two compressors. The gzip hit rate for the file type pdf is better than the hit rate of bzip2. Both the compressors are giving good results for csv, eps, java and bmp file types and they are showing hit rates more than 60 percent for these file types. But the hit rate of gzip for java file types is better than bzip2 and the hit rate of bzip2 for bmp file type is better than gzip.

Once again the thing that should be kept in mind while reading the results from the graph is that the bzip2 is better than gzip in compression ratio and gzip is better than bzip2 in idempotency. As we can see that the performance of gzip is better than bzip2 for many file types so it seems that a compressor which is better in idempotency should be selected for the NCD instead of compression ratio. The results of all the file types can be visualized in the similar way.
Graph 17: Hit rate for Blocksize-1024 & averaged k (x-axis: File Types, y-axis: Hit rate)

<table>
<thead>
<tr>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>1</td>
</tr>
<tr>
<td>pdf</td>
<td>33.6</td>
</tr>
<tr>
<td>html</td>
<td>38.1</td>
</tr>
<tr>
<td>tml</td>
<td>58</td>
</tr>
<tr>
<td>js</td>
<td>37.1</td>
</tr>
<tr>
<td>dot</td>
<td>18.4</td>
</tr>
<tr>
<td>txt</td>
<td>35</td>
</tr>
<tr>
<td>html</td>
<td>53.4</td>
</tr>
<tr>
<td>js</td>
<td>59.9</td>
</tr>
<tr>
<td>xsl</td>
<td>78.3</td>
</tr>
<tr>
<td>po</td>
<td>50.1</td>
</tr>
<tr>
<td>xsl</td>
<td>54.8</td>
</tr>
<tr>
<td>po</td>
<td>48.6</td>
</tr>
<tr>
<td>gif</td>
<td>6.02</td>
</tr>
<tr>
<td>ps</td>
<td>20.9</td>
</tr>
<tr>
<td>csv</td>
<td>64.2</td>
</tr>
<tr>
<td>eg</td>
<td>59.1</td>
</tr>
<tr>
<td>poi</td>
<td>6.02</td>
</tr>
<tr>
<td>eg</td>
<td>11.2</td>
</tr>
<tr>
<td>sw</td>
<td>17.1</td>
</tr>
<tr>
<td>jp</td>
<td>15.9</td>
</tr>
<tr>
<td>sql</td>
<td>56.9</td>
</tr>
<tr>
<td>jsv</td>
<td>75.1</td>
</tr>
<tr>
<td>pp</td>
<td>15.2</td>
</tr>
</tbody>
</table>

Graph 17: Hit rate for Blocksize-1024 & averaged k (x-axis: File Types, y-axis: Hit rate)
<table>
<thead>
<tr>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdf</td>
<td>8.04</td>
<td>8.04</td>
<td>8.02</td>
<td>7.98</td>
<td>7.89</td>
<td>7.78</td>
<td>7.67</td>
<td>7.56</td>
<td>7.45</td>
<td>7.34</td>
</tr>
<tr>
<td>hit</td>
<td>8.20</td>
<td>8.20</td>
<td>8.20</td>
<td>8.20</td>
<td>8.20</td>
<td>8.20</td>
<td>8.20</td>
<td>8.20</td>
<td>8.20</td>
<td>8.20</td>
</tr>
<tr>
<td>jlg</td>
<td>1.53</td>
<td>1.53</td>
<td>1.53</td>
<td>1.53</td>
<td>1.53</td>
<td>1.53</td>
<td>1.53</td>
<td>1.53</td>
<td>1.53</td>
<td>1.53</td>
</tr>
<tr>
<td>txt</td>
<td>6.95</td>
<td>6.95</td>
<td>6.95</td>
<td>6.95</td>
<td>6.95</td>
<td>6.95</td>
<td>6.95</td>
<td>6.95</td>
<td>6.95</td>
<td>6.95</td>
</tr>
<tr>
<td>xpl</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
<td>14.5</td>
</tr>
<tr>
<td>glf</td>
<td>3.56</td>
<td>3.56</td>
<td>3.56</td>
<td>3.56</td>
<td>3.56</td>
<td>3.56</td>
<td>3.56</td>
<td>3.56</td>
<td>3.56</td>
<td>3.56</td>
</tr>
<tr>
<td>os</td>
<td>5.97</td>
<td>5.97</td>
<td>5.97</td>
<td>5.97</td>
<td>5.97</td>
<td>5.97</td>
<td>5.97</td>
<td>5.97</td>
<td>5.97</td>
<td>5.97</td>
</tr>
<tr>
<td>ciz</td>
<td>5.76</td>
<td>5.76</td>
<td>5.76</td>
<td>5.76</td>
<td>5.76</td>
<td>5.76</td>
<td>5.76</td>
<td>5.76</td>
<td>5.76</td>
<td>5.76</td>
</tr>
<tr>
<td>gr</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
<td>2.35</td>
</tr>
<tr>
<td>og</td>
<td>8.78</td>
<td>8.78</td>
<td>8.78</td>
<td>8.78</td>
<td>8.78</td>
<td>8.78</td>
<td>8.78</td>
<td>8.78</td>
<td>8.78</td>
<td>8.78</td>
</tr>
<tr>
<td>rfr</td>
<td>7.16</td>
<td>7.16</td>
<td>7.16</td>
<td>7.16</td>
<td>7.16</td>
<td>7.16</td>
<td>7.16</td>
<td>7.16</td>
<td>7.16</td>
<td>7.16</td>
</tr>
<tr>
<td>sqg</td>
<td>7.50</td>
<td>7.50</td>
<td>7.50</td>
<td>7.50</td>
<td>7.50</td>
<td>7.50</td>
<td>7.50</td>
<td>7.50</td>
<td>7.50</td>
<td>7.50</td>
</tr>
<tr>
<td>jav</td>
<td>7.53</td>
<td>7.53</td>
<td>7.53</td>
<td>7.53</td>
<td>7.53</td>
<td>7.53</td>
<td>7.53</td>
<td>7.53</td>
<td>7.53</td>
<td>7.53</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>1</td>
</tr>
<tr>
<td>pdf</td>
<td>8.04</td>
</tr>
<tr>
<td>hit</td>
<td>8.20</td>
</tr>
<tr>
<td>jlg</td>
<td>1.53</td>
</tr>
<tr>
<td>txt</td>
<td>6.95</td>
</tr>
<tr>
<td>xpl</td>
<td>14.5</td>
</tr>
<tr>
<td>glf</td>
<td>3.56</td>
</tr>
<tr>
<td>os</td>
<td>5.97</td>
</tr>
<tr>
<td>ciz</td>
<td>5.76</td>
</tr>
<tr>
<td>gr</td>
<td>2.35</td>
</tr>
<tr>
<td>og</td>
<td>8.78</td>
</tr>
<tr>
<td>rfr</td>
<td>7.16</td>
</tr>
<tr>
<td>sqg</td>
<td>7.50</td>
</tr>
<tr>
<td>jav</td>
<td>7.53</td>
</tr>
</tbody>
</table>

Table 7: Hit rate for blocksize-1024
The overall results of bzip2 & gzip on NCD’s performance can be visualized from the Graph 18 below, the performance of both bzip2 and gzip for each file type is given there.

In graph, the x-axis is representing all the 28 file types and the y-axis is representing the percentage hit rate (hit rate is: success rate used for expressing the performance) and the colours are representing the two compressors. This time the gzip hit rate for the file type pdf is not better than the hit rate of bzip2. Both the compressors are giving good results for csv, eps, java and bmp file types and they are showing hit rates more than 60 percent for these file types and the hit rate of gzip for java file types is better than bzip2 and the hit rate of bzip2 for bmp file type is better than gzip.

Once again the thing that should be kept in mind while reading the results from the graph is that the bzip2 is better than gzip in compression ratio and gzip is better than bzip2 in idempotency. As we can see that the performance of gzip is better than bzip2 for many file types so it seems that a compressor which is better in idempotency should be selected for the NCD instead of compression ratio. The results of all the file types can be visualized in the similar way.

The results of bzip2 & gzip for different k values (0 to 10) are given in table 9 and the standard deviations of both the compressors are also given in table 10.
Graph 18: Hit rate for Blocksize-1536 & averaged k (x-axis: File Types, y-axis: Hit rate)
Table 9: Hit rate for blocksize-1536

<table>
<thead>
<tr>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>k</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td><strong>pdf</strong></td>
<td>3.86</td>
</tr>
<tr>
<td><strong>hit</strong></td>
<td>3.86</td>
</tr>
<tr>
<td><strong>jpg</strong></td>
<td>2.70</td>
</tr>
<tr>
<td><strong>txt</strong></td>
<td>15.6</td>
</tr>
<tr>
<td><strong>doc</strong></td>
<td>15.5</td>
</tr>
<tr>
<td><strong>xls</strong></td>
<td>13.0</td>
</tr>
<tr>
<td><strong>ppt</strong></td>
<td>4.84</td>
</tr>
<tr>
<td><strong>gpl</strong></td>
<td>3.29</td>
</tr>
<tr>
<td><strong>ps</strong></td>
<td>11.2</td>
</tr>
<tr>
<td><strong>civ</strong></td>
<td>18.3</td>
</tr>
<tr>
<td><strong>gr</strong></td>
<td>6.34</td>
</tr>
<tr>
<td><strong>png</strong></td>
<td>1.08</td>
</tr>
<tr>
<td><strong>sw</strong></td>
<td>8.16</td>
</tr>
<tr>
<td><strong>sql</strong></td>
<td>22.5</td>
</tr>
<tr>
<td><strong>jav</strong></td>
<td>11.5</td>
</tr>
<tr>
<td><strong>cr</strong></td>
<td>32.3</td>
</tr>
<tr>
<td><strong>ps</strong></td>
<td>67.2</td>
</tr>
<tr>
<td><strong>bf</strong></td>
<td>86.4</td>
</tr>
<tr>
<td><strong>xb</strong></td>
<td>82.1</td>
</tr>
<tr>
<td><strong>jar</strong></td>
<td>1536</td>
</tr>
<tr>
<td><strong>zip</strong></td>
<td>49.1</td>
</tr>
</tbody>
</table>

49
The overall results of bzip2 & gzip on NCD’s performance can be visualized from the Graph 19 below, the performance of both bzip2 and gzip for each file type is given there.

In graph, the x-axis is representing all the 28 file types and the y-axis is representing the percentage hit rate (hit rate is: success rate used for expressing the performance) and the colours are representing the two compressors. This time again the gzip hit rate for the file type pdf is better than the hit rate of bzip2. Both the compressors are giving good results for csv, eps, java and bmp file types and they are showing hit rates more than 60 percent for these file types and the hit rate of gzip for java file types is better than bzip2 and the hit rate of bzip2 for bmp file type is better than gzip.

Once again the thing that should be kept in mind while reading the results from the graph is that the bzip2 is better than gzip in compression ratio and gzip is better than bzip2 in idempotency. As we can see that the performance of gzip is better than bzip2 for many file types so it seems that a compressor which is better in idempotency should be selected for the NCD instead of compression ratio. The results of all the file types can be visualized in the similar way.

The results of bzip2 & gzip for different k values (0 to 10) are given in table 11 and the standard deviations of both the compressors are also given in table 12.

**Table 10: Standard deviation for blocksize-1536**

| pp | tx | 15.9 | 15.3 | 17.0 | 19.3 | 21.7 | 22.5 | 21.7 | 22.3 | 21.3 | 22.0 | 19.2 | 22.8 | 22.2 | 23.0 | 22.0 | 21.9 | 22.8 | 23.2 | 22.3 | 22.7 |
|----|----|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| do | cr | 2.35 | 4.72 | 3.28 | 3.13 | 2.81 | 3.81 | 2.81 | 3.21 | 5.43 | 3.52 | 8.76 | 4.46 | 10.0 | 7.72 | 7.61 | 7.84 | 11.0 | 5.96 | 7.94 | 10.0 |
| tf | 4.15 | 3.24 | 3.33 | 6.53 | 6.26 | 5.25 | 5.02 | 8.40 | 7.38 | 3.46 | 5.89 | 4.42 | 7.54 | 5.97 | 7.64 | 9.38 | 13.1 | 9.31 |
| js | 13.9 | 14.8 | 19.9 | 76 | 19.9 | 76 | 44 | 19 | 28 | 41 | 64 | 81 | 96 | 48 | 94 | 18 | 01 | 29 | 18 | 35 | 34 |
| zip | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| bm | 5.70 | 7.19 | 5.50 | 5.16 | 4.87 | 5.81 | 3.21 | 2.52 | 2.71 | 7.60 | 8.73 | 10.8 | 9.87 | 11.9 | 12.1 | 10.4 | 11.6 | 10.2 | 11.4 |
| xb | 4.27 | 6.87 | 5.29 | 7.84 | 7.84 | 4.25 | 4.36 | 4.42 | 4.71 | 3.69 | 4.93 | 4.99 | 5.20 | 13.7 | 11.2 | 9.64 | 9.64 | 10.7 | 03 |
| x | 2.04 | 0.92 | 4.12 | 3.28 | 3.64 | 3.27 | 3.36 | 3.82 | 7.40 | 3.51 | 4.92 | 8.48 | 8.98 | 9.74 | 9.38 | 8.35 | 9.83 | 10.8 | 10.9 |
| jar | 23.7 | 16.1 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 | 87 |
| zip | 11.9 | 7.26 | 13.1 | 99 | 13.4 | 81 | 14.9 | 12.8 | 7.66 | 4.50 | 10.5 | 10.0 | 12.5 | 10.5 | 13.3 | 14.9 | 15.4 | 14.0 | 9.02 | 3.21 | 13.0 | 9.85 |

3.2.4.1.1.4 Results for Blocksize-2048
Graph 19: Hit rate for Blocksize-2048 & averaged k (x-axis: File Types, y-axis: Hit rate)

<table>
<thead>
<tr>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>k</strong></td>
<td><strong>1</strong></td>
</tr>
<tr>
<td><strong>pdf</strong></td>
<td>46.2</td>
</tr>
<tr>
<td><strong>txt</strong></td>
<td>8.9</td>
</tr>
<tr>
<td><strong>xls</strong></td>
<td>66.1</td>
</tr>
<tr>
<td><strong>ps</strong></td>
<td>76</td>
</tr>
<tr>
<td><strong>doc</strong></td>
<td>59</td>
</tr>
<tr>
<td><strong>ppt</strong></td>
<td>93</td>
</tr>
<tr>
<td><strong>xml</strong></td>
<td>16</td>
</tr>
<tr>
<td><strong>pptx</strong></td>
<td>97</td>
</tr>
<tr>
<td><strong>jar</strong></td>
<td>35</td>
</tr>
<tr>
<td><strong>html</strong></td>
<td>55</td>
</tr>
<tr>
<td><strong>zip</strong></td>
<td>50.5</td>
</tr>
<tr>
<td><strong>jar</strong></td>
<td>38</td>
</tr>
<tr>
<td><strong>xml</strong></td>
<td>72.2</td>
</tr>
<tr>
<td><strong>docx</strong></td>
<td>59</td>
</tr>
<tr>
<td><strong>xls</strong></td>
<td>12.4</td>
</tr>
</tbody>
</table>

Legend:
- bzip2
- gzip

51
Table 1: Hit rate for blocksize-2048

<table>
<thead>
<tr>
<th>Blocksize</th>
<th>Bzip2</th>
<th>Gzip</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9.43</td>
<td>1.17</td>
</tr>
<tr>
<td>2</td>
<td>8.02</td>
<td>0.95</td>
</tr>
<tr>
<td>3</td>
<td>6.90</td>
<td>1.24</td>
</tr>
<tr>
<td>4</td>
<td>8.85</td>
<td>2.23</td>
</tr>
<tr>
<td>5</td>
<td>9.62</td>
<td>7.08</td>
</tr>
<tr>
<td>6</td>
<td>9.61</td>
<td>7.49</td>
</tr>
<tr>
<td>7</td>
<td>9.39</td>
<td>7.72</td>
</tr>
<tr>
<td>8</td>
<td>9.67</td>
<td>6.03</td>
</tr>
<tr>
<td>9</td>
<td>9.66</td>
<td>7.74</td>
</tr>
<tr>
<td>10</td>
<td>9.66</td>
<td>7.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pdf</td>
<td>10.0</td>
<td>12.0</td>
</tr>
<tr>
<td>1</td>
<td>8.40</td>
<td>7.09</td>
</tr>
<tr>
<td>2</td>
<td>9.04</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>8.74</td>
<td>5.12</td>
</tr>
<tr>
<td>4</td>
<td>9.65</td>
<td>7.42</td>
</tr>
<tr>
<td>5</td>
<td>9.87</td>
<td>8.29</td>
</tr>
<tr>
<td>6</td>
<td>9.87</td>
<td>3.29</td>
</tr>
<tr>
<td>7</td>
<td>11.3</td>
<td>10.4</td>
</tr>
<tr>
<td>8</td>
<td>11.8</td>
<td>11.0</td>
</tr>
<tr>
<td>9</td>
<td>11.8</td>
<td>11.0</td>
</tr>
<tr>
<td>10</td>
<td>11.4</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ht</td>
<td>15.2</td>
<td>1.17</td>
</tr>
<tr>
<td>1</td>
<td>13.7</td>
<td>0.95</td>
</tr>
<tr>
<td>2</td>
<td>11.0</td>
<td>1.24</td>
</tr>
<tr>
<td>3</td>
<td>13.6</td>
<td>2.23</td>
</tr>
<tr>
<td>4</td>
<td>15.2</td>
<td>7.08</td>
</tr>
<tr>
<td>5</td>
<td>16.6</td>
<td>7.49</td>
</tr>
<tr>
<td>6</td>
<td>17.6</td>
<td>7.72</td>
</tr>
<tr>
<td>7</td>
<td>16.9</td>
<td>6.03</td>
</tr>
<tr>
<td>8</td>
<td>16.8</td>
<td>7.74</td>
</tr>
<tr>
<td>9</td>
<td>17.8</td>
<td>7.37</td>
</tr>
<tr>
<td>10</td>
<td>17.8</td>
<td>7.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ml</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jpg</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>txt</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>do</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xz</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tar</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xz</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xz</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>xz</td>
<td>1.17</td>
<td>5.12</td>
</tr>
<tr>
<td>1</td>
<td>0.95</td>
<td>7.42</td>
</tr>
<tr>
<td>2</td>
<td>1.24</td>
<td>5.63</td>
</tr>
<tr>
<td>3</td>
<td>4.65</td>
<td>7.09</td>
</tr>
<tr>
<td>4</td>
<td>3.99</td>
<td>5.12</td>
</tr>
<tr>
<td>5</td>
<td>2.55</td>
<td>7.42</td>
</tr>
<tr>
<td>6</td>
<td>2.33</td>
<td>5.63</td>
</tr>
<tr>
<td>7</td>
<td>2.87</td>
<td>7.09</td>
</tr>
<tr>
<td>8</td>
<td>2.62</td>
<td>5.12</td>
</tr>
<tr>
<td>9</td>
<td>2.75</td>
<td>7.42</td>
</tr>
<tr>
<td>10</td>
<td>2.96</td>
<td>5.63</td>
</tr>
<tr>
<td>file type</td>
<td>standard deviation for blocksize-2048</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------</td>
<td></td>
</tr>
<tr>
<td>sql</td>
<td>26.8 06</td>
<td></td>
</tr>
<tr>
<td>jav</td>
<td>14.8 12</td>
<td></td>
</tr>
<tr>
<td>ps</td>
<td>17.2 62</td>
<td></td>
</tr>
<tr>
<td>do</td>
<td>4.73 3</td>
<td></td>
</tr>
<tr>
<td>txt</td>
<td>4.21 1</td>
<td></td>
</tr>
<tr>
<td>jar</td>
<td>21.3 35</td>
<td></td>
</tr>
<tr>
<td>zip</td>
<td>14.0 34</td>
<td></td>
</tr>
<tr>
<td>av</td>
<td>14.3 11</td>
<td></td>
</tr>
<tr>
<td>pp</td>
<td>17.5 78</td>
<td></td>
</tr>
<tr>
<td>co</td>
<td>5.27 2</td>
<td></td>
</tr>
<tr>
<td>ttf</td>
<td>7.57 4</td>
<td></td>
</tr>
<tr>
<td>js</td>
<td>17.4 43</td>
<td></td>
</tr>
<tr>
<td>pb</td>
<td>4.90 6</td>
<td></td>
</tr>
<tr>
<td>bm</td>
<td>1.70 3</td>
<td></td>
</tr>
<tr>
<td>xb</td>
<td>2.08 5</td>
<td></td>
</tr>
<tr>
<td>js</td>
<td>14.5 17</td>
<td></td>
</tr>
<tr>
<td>xj</td>
<td>21.8 35</td>
<td></td>
</tr>
<tr>
<td>xsl</td>
<td>14.0 34</td>
<td></td>
</tr>
<tr>
<td>xpl</td>
<td>14.0 34</td>
<td></td>
</tr>
<tr>
<td>xbs</td>
<td>14.0 34</td>
<td></td>
</tr>
<tr>
<td>xbs</td>
<td>14.0 34</td>
<td></td>
</tr>
<tr>
<td>jav</td>
<td>14.3 11</td>
<td></td>
</tr>
<tr>
<td>ps</td>
<td>17.2 62</td>
<td></td>
</tr>
<tr>
<td>do</td>
<td>4.73 3</td>
<td></td>
</tr>
<tr>
<td>txt</td>
<td>4.21 1</td>
<td></td>
</tr>
<tr>
<td>jar</td>
<td>21.3 35</td>
<td></td>
</tr>
<tr>
<td>zip</td>
<td>14.0 34</td>
<td></td>
</tr>
<tr>
<td>av</td>
<td>14.3 11</td>
<td></td>
</tr>
<tr>
<td>pp</td>
<td>17.5 78</td>
<td></td>
</tr>
<tr>
<td>co</td>
<td>5.27 2</td>
<td></td>
</tr>
<tr>
<td>ttf</td>
<td>7.57 4</td>
<td></td>
</tr>
<tr>
<td>js</td>
<td>17.4 43</td>
<td></td>
</tr>
<tr>
<td>pb</td>
<td>4.90 6</td>
<td></td>
</tr>
<tr>
<td>bm</td>
<td>1.70 3</td>
<td></td>
</tr>
<tr>
<td>xb</td>
<td>2.08 5</td>
<td></td>
</tr>
<tr>
<td>xj</td>
<td>21.3 35</td>
<td></td>
</tr>
<tr>
<td>xsl</td>
<td>14.0 34</td>
<td></td>
</tr>
<tr>
<td>xpl</td>
<td>14.0 34</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Standard deviation for blocksize-2048

3.2.4.1.1.5 For Individual Bzip2

The overall results of bzip2 on NCD’s performance can be visualized from the Graph 20 below. The purpose of this graph is to give an overview of the effect of increasing the block size of data.

In graph, the x-axis is representing all the 28 file types along with the block sizes and the y-axis is representing the percentage hit rate (hit rate is: success rate used for expressing the performance) and this time the colours are representing the four block sizes. The hit rate of bzip2 for the pdf file type is improving up by increasing the block size from 512 bytes to 1024 bytes, then to 1536 bytes and then to 2048 bytes. The results of all the file types can be visualized in the similar way.

The thing that should be kept in mind while reading the results from the graph is that for some file types like pdf, html, doc the hit rate of bzip2 is improving up by increasing the block size from 512 bytes to 1024 bytes, then to 1536 bytes and then to 2048 bytes but for some file types like zip, xsl the hit rate of bzip2 is decreasing down also by increasing the block size. The results of all the file types can be visualized in the similar way.

53
3.2.4.1.1.6 For Individual Gzip

The overall results of gzip on NCD’s performance can be visualized from the Graph 21 below. The purpose of this graph is to give an overview of the effect of increasing the block size of data.

In graph, the x-axis is representing all the 28 file types along with the block sizes and the y-axis is representing the percentage hit rate (hit rate is: success rate used for expressing the performance) and the colours are representing the four block sizes. The hit rate of gzip for the doc file type is improving up by increasing the block size from 512 bytes to 1024 bytes, then to 1536 bytes and then to 2048 bytes. The results of all the file types can be visualized in the similar way.

The thing that should be kept in mind while reading the results from the graph is that for some file types like doc, xbm the hit rate of gzip is improving up by increasing the block size from 512 bytes to 1024 bytes, then to 1536 bytes and then to 2048 bytes but for some file types like csv, bmp the hit rate of bzip2 is decreasing down also by increasing the block size. The results of all the file types can be visualized in the similar way.
Graph 21: Hit rate for All Block size with averaged k by Gzip (x-axis: File Types with block sizes, y-axis: Hit rate)
3.2.4.2 Hypothesis Testing

The hypothesis testing is performed on the collected data for main experiment. The first hypothesis H0 regarding there is no difference on NCD’s classification performance when a compressor is used for NCD algorithm on the basis of compression ratio or idempotency is evaluated using ANOVA test. The second hypothesis H2 regarding there is no difference on NCD’s classification performance when a block size of data is changed is evaluated using ANOVA two factor test. The reason of using ANOVA is that as we have to independent variables so it is more appropriate to test the hypotheses using ANOVA test.

First Hypothesis, H0: There is no difference on NCD’s classification performance when a compressor is used for NCD algorithm on the basis of compression ratio or idempotency.

The results from ANOVA test for this hypothesis are shown below in table 13:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>2794.876</td>
<td>19</td>
<td>147.0988</td>
<td>0.286701</td>
<td>0.998864</td>
<td>1.60583</td>
</tr>
<tr>
<td>Within Groups</td>
<td>277059.8</td>
<td>540</td>
<td>513.0738</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>279854.7</td>
<td>559</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 13: Results from ANOVA Test

- The results of ANOVA test from table 13 are not in favour of rejecting the null hypothesis, as the F is less than F crit so we cannot reject the null hypothesis and as P-value is greater than 0.05 which is not significant. I.e. there is no significant difference on NCD’s classification performance when a compressor is used for NCD algorithm on the basis of compression ratio or idempotency.

Second Hypothesis, H2: There is no difference on NCD’s classification performance by bzip2 & gzip when a block size of data is changed.

As here we have two factors with different levels of treatments so we will use factorial ANOVA and the hypothesis will be break down as:

Case-1) \( H0 \), Block size original = Block size Changed

\( H1 \), Block size original ≠ Block size Changed
Case-2) $H_2$, Bzip2 = Gzip

$H_3$, Bzip2 ≠ Gzip

Case-3) $H_4$, an interaction is absent

$H_5$, an interaction is present

The results of ANOVA test for these hypotheses are shown below in table 14:

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>972.2505</td>
<td>3</td>
<td>324.0835</td>
<td>0.627414</td>
<td>0.597286</td>
<td>2.609022</td>
</tr>
<tr>
<td>Columns</td>
<td>6653.953</td>
<td>19</td>
<td>350.2081</td>
<td>0.67799</td>
<td>0.843952</td>
<td>1.59133</td>
</tr>
<tr>
<td>Interaction</td>
<td>874.7124</td>
<td>57</td>
<td>15.34583</td>
<td>0.029709</td>
<td>1</td>
<td>1.333049</td>
</tr>
<tr>
<td>Within</td>
<td>1115723</td>
<td>2160</td>
<td>516.5386</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1124224</td>
<td>2239</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 14: Results from ANOVA two factor Test

The results of analysis are again not very significant as:

Case-1) $H_0$ is retained.

As in above table 14 in sample row, $F$ is less than $F$ crit and p-value is greater than 0.05, $H_0$ will retain. I.e. there is no significant difference on NCD’s classification performance by bzip2 & gzip when a block size of data is changed.

Case-2) $H_2$ will retain.

As in above table 14 in Columns row, $F$ is less than $F$ crit and p-value is greater than 0.05, $H_0$ will retain. So we cannot accept that there is any significant difference on NCD’s performance by bzip2 or gzip.

Case-3) $H_4$ will retain

As in above table 14 in Interaction row, $F$ is less than $F$ crit and p-value is greater than 0.05, $H_0$ will retain. I.e. Interaction is absent.
4 DISCUSSION

Our results have shown that there is no significant effect on the performance of NCD algorithm by using different block sizes of data and also there is no significant effect on the NCD’s performance either a compressor for NCD is selected on the basis of better compression ratio or idempotency.

The purpose of using different block sizes (512 bytes, 1024 bytes, 1536 bytes, 2048 bytes) was to examine that either the performance of NCD will improve or not, either one should take care of the block sizes while making the dataset for NCD or not. Although the results from section 3.2 show that the performance do improves when the block size is changed from 512 bytes to 1024 bytes, 1536 bytes and 2048 bytes for some file types but this improvement in performance of the obtained results is not sufficient enough to reject the null hypothesis.

The second parameter idempotency versus compression ratio was evaluated with intension that if someone uses a compressor for NCD which is better in idempotency can produce better performance as compared to compression ratio. Although the results from section 3.2 show that for some file types the performance of NCD with gzip compressor (which is better in idempotency than bzip2) is better than the performance of NCD with bzip2 compressor (which is better in compression ratio than gzip) but this improvement in performance of the obtained results is not sufficient enough to reject the null hypothesis.
5 Conclusion and Future Work

The effect of idempotency and block size of data on the performance of the NCD algorithm is studied by performing the two experiments namely:

1. Supporting Experiment for evaluating Idempotencies
2. Main Experiment for evaluating NCD’s performance

The null hypothesis of supporting experiment is rejected which concludes that there exists statistically significant effect in the measured idempotencies by the three compressors bzip2, gzip and ppmd, and the measured idempotencies of gzip are better than ppmd and the measured idempotencies of ppmd are better than bzip2.

The main experiment was then performed to examine the performance of NCD algorithm in perspective of compression ratio versus idempotency of two compressors (bzip2 & gzip). The bzip2 is better than gzip in compression ratio and the gzip is better than bzip2 in idempotency. As the null hypothesis of main experiment is retained which concludes that there is no statistical significant difference on NCD’s performance either a compressor is used for NCD algorithm on the basis of compression ratio or idempotency.

In the main experiment the effect of changing block size of data is also evaluated and the results of hypothesis testing shows that there is no statistically significant effect on the NCD’s performance when the block size of data is changed from 512 bytes to 1024 bytes, 1536 bytes and 2048 bytes.

From the results of both supporting and main experiment it can also be concluded that although idempotency of gzip is better than bzip2 but when we came to measure this effect of better idempotency of gzip on the performance of NCD it was not statistically significantly better than bzip2 which is better than gzip in compression ratio. In other words the performance of NCD does not improve statistically significantly either a compressor is selected on the basis of idempotency or compression ratio.

In future a similar combination of experiment could be conducted for NCD’s performance by bzip2 and gzip with a different classification algorithm to ensure the results of hypotheses. And also a further study could be conducted in future to find the reasons behind non significant relationship between the compression ratio and idempotency.
6 **Related Work**

The most related work is done by Axelsson [1] where he used the similar data corpus and NCD algorithm as we did here and he performed the experiment for only bzip2 with the block size of 512 bytes. And here we performed experiment with gzip and bzip2 along with the four different block sizes 512 bytes, 1024 bytes, 1536 bytes and 2048 bytes and we have compared the performance of both gzip and bzip2. And also he run the NCD algorithm once for ten trial sets and here we did for three trial sets. The NCD algorithm in that experiment did the best for the file types eps, java, csv and ttf. It also did well for xml, js, bmp and doc file types while the performance of ppt, swf, jpg, gif, pptx, pps, gz, png file types is poor.

Karresand and Shahmehri [9] developed an Oscar method [8] for the classification of file fragments which is entirely based on the structure of the file fragments rather than using the meta data. They reported a detection rate of 99.2 % for jpg file type without any false positives. As their work is limited to one file type so it is hard to compare their results with ours as we here address the n-valued problem (for more than one file types).

Veenman [14] also tackled the n-valued problem with the value of n (is equal to 11). He has used statistical content of the clusters of the specific file types and he used block size of 4096 KB. He reported an overall accuracy of 45 % with low false positive rate (as he reported the results in graphical form so it is difficult to compare his results with ours)

Calhoun and Coles [2] investigated two algorithms for classifying the type of file fragments, one based on Fisher’s linear discriminant and the other based on longest common subsequences of the fragment. They used four file types jpg, pdf, gif and bmp. They did a pair wise comparison of the classification of file fragments having accuracy of more than 80 %.
7 REFERENCES


