Enl@ce: Revista Venezolana de Información, Tecnología y Conocimiento
ISSN: 1690-7515
Depósito legal pp 200402ZU1624
Año 9: No. 3, Septiembre-Diciembre, pp. 83-99

Cómo citar el artículo (Normas APA):
Rincón, C., Bracho, D. y Acurero, A. (2012). Efecto del tamaño del alfabeto en el rendimiento de un algoritmo de compresión probabilístico. *Enl@ce Revista Venezolana de Información, Tecnología y Conocimiento*, 9 (3), 83-99

Efecto del tamaño del alfabeto en el rendimiento de un algoritmo de compresión probabilístico

Carlos Rincón[1]
David Bracho[2]
Alfredo Acurero[3]

Resumen

El presente trabajo tuvo como finalidad determinar el efecto del tamaño del alfabeto de un mensaje, en el rendimiento del algoritmo de compresión probabilístico basado en la posición de los símbolos, el cual propusimos en un trabajo previo (Rincón, Acurero, Bracho y Jakymec, 2008). La metodología utilizada consistió en 7 etapas: (a) determinación de las variables dependientes e independientes a objeto de estudio, (b) desarrollo e implementación del algoritmo propuesto, (c) construcción de los archivos de prueba, (d) ejecución del algoritmo implementado sobre los archivos de prueba, (e) determinación del modelo matemático que explique el comportamiento de las variables dependientes, (f) aplicación del método estadístico análisis de varianza, (g) análisis de los resultados obtenidos. Las variables dependientes seleccionadas fueron el tiempo de compresión y la relación de compresión. El diseño del modelo estadístico seleccionado fue un totalmente aleatorizado con tratamiento en un arreglo factorial 4x2, con dos factores: tamaño del alfabeto (4,8,12 y 16 símbolos) y distribucción probabilística del alfabeto (aleatorio y equiprobable). Del análisis de varianza se obtuvo diferencias significativas para todas las variables independientes y su interacción en todas las variables dependientes, corroborando así el efecto que tiene el tamaño del alfabeto en el rendimiento del algoritmo de compresión estudiado. La prueba de Tukey determinó que para la variable tiempo de compresión el mejor rendimiento se obtiene con la distribución aleatoria y el mayor tamaño del alfabeto (12 y 16), mientras que para la variable relación de compresión, el mejor rendimiento se obtiene con la distribución aleatoria y el menor tamaño del alfabeto.

Palabras clave: compresión, posición, símbolos, tamaño, alfabeto

Recibido: 05-07-12 Aceptado: 02-11-12

1 Licenciado en Computación. Magister en Telemática. DEA en Informática. Profesor Asociado de la Licenciatura en Computación. Universidad del Zulia, Venezuela. Director del Departamento de Computación de la Facultad Experimental de Ciencias. Correo electrónico: crincon@fec.luz.edu.ve Departamento de Computación. Facultad Experimental de Ciencias. Universidad del Zulia

2 Ingeniero en Computación. Magister en Gerencia de Empresas. DEA en Informática. Doctorando en Ciencias Sociales. Profesor Titular de la Licenciatura en Computación. Universidad del Zulia. Correo electrónico: drbracho@fec.luz.edu.ve

3 Licenciado en Computación. Magister en Gerencia de Empresas. DEA en Informática. Doctorando en Ciencias Sociales. Profesor Asociado de la Licenciatura en Computación. Universidad del Zulia, Venezuela. Director de DICTICLUZ Correo electrónico: aacurero@fec.luz.edu.ve

Effect of Alphabet Size on the Performance of a Probabilistic Compression Algorithm

Abstract

The purpose of the present work was to determine the effect of alphabet size on the performance of a probabilistic compression algorithm based on symbol's position proposed by Rincón, Acurero, Bracho y Jakymec, 2008. The methodology used consisted of 7 stages: (a) determination of the independent and dependent variables under study, (b) implementation of the proposed algorithm, (c) test files construction, (d) execution of the implemented algorithm on test files, (e) determination of the mathematical model that explains the behavior of the dependent variables, (f) application of the anova procedure, (g) results analysis. The dependent variables used to measure the algorithm performance were compression time and compression ratio. The statistical model designed was a totally randomized with treatment on a factorial array 4x2, with 2 factors: alphabet size (4,8,12 and 16 symbols) and alphabet probabilistic distribution (random and equiprobable). The Results of the anova procedure showed significative differences for all independent variables and their interactions on all dependent variables, corroborating the effect of the alphabet size on the performance of a probabilistic compression algorithm based on symbols position. Tukey's media test determines that for compression time, the best performance was obtained with random distribution and the higher alphabet size (12 and 16), while for compression ratio the best performance was obtained with random distribution and the lowest alphabet size (4).

Keywords: Compression, Position, Symbols, Size, Alphabet

Introducción

En 2009, presentamos un nuevo método de compresión basado en la posición de los símbolos en el mensaje denominado *Algoritmo de compresión probabilístico basado en la teoría de la información* (Rincón, Rodríguez, Acurero, Bracho y Jakymec, 2009). Estudios preliminares de rendimiento del algoritmo propuesto han mostrado un efecto adverso (generación de una cadena de bits *sparse*) sobre la relación de compresión del algoritmo a medida que aumenta el tamaño del alfabeto.

Aunque pareciera que la compresión probabilística sin pérdida es un campo poco estudiado como consecuencia de los avances en la compresión basada en diccionario, trabajos como: (Chan, 2008), (Wenjun; Weimin y Hui, 2006), (Ling; Qian y Wang, 2008), (Marcelloni y Vecchio, 2008), entre otros, permiten concluir que la compresión probabilística sin pérdida es utilizada en la actualidad para diferentes áreas de la compresión de datos.

Por lo antes expuesto, el propósito de la presente investigación consiste en analizar el problema que causa el incremento del tamaño del al-

fabeto del mensaje en la cadena de bits generada por el algoritmo, para luego proponer soluciones que permitan al algoritmo propuesto, aumentar su rendimiento (medido en función de la relación y tiempo de compresión).

Varios autores como: (Salomon, 2000), (Fraenkel y Klein, 1985), (Moffat y Zobel, 1992), entre otros, han propuesto distintas soluciones al problema de la compresión de cadenas de bits *sparse*. Luego de realizar un análisis del estado del arte sobre este problema, se implementarán las soluciones sobre el algoritmo propuesto, con la finalidad de obtener una modificación al mismo que mejore su rendimiento.

Aspectos teóricos

Compresión de datos

La compresión de datos es una técnica que permite representar la información generada por una fuente de datos con una menor cantidad de bits, utilizando para este fin, códigos óptimos.

Las soluciones de compresión se dividen en 2 pasos: (a) Algoritmo de compresión: es aquel que determina un código óptimo que permita representar los datos con la menor cantidad de bits y que genera un archivo comprimido basándose en los códigos óptimos encontrados. (b) Algoritmo de descompresión: es aquel que se encarga de transformar el archivo comprimido en la información original.

Relación de compresión

Es una métrica que permite determinar el rendimiento de un algoritmo de compresión. Matemáticamente puede definirse de dos formas: (a) Relación n a 1: RC = TO/TC y (b) Como porcentaje: RC = (TO-TC)/TO, donde TO = tamaño del archivo original y TC = tamaño del archivo comprimido. Si bien la relación de compresión es planteada como una función cuyas variables independientes son el tamaño del archivo original y el tamaño del archivo comprimido, esta métrica puede ser simplificada considerando que el tamaño del archivo original es igual al número de símbolos del archivo multiplicado por el tamaño promedio de los símbolos sin comprimir ($\overline{n}_{SC}$), y que el tamaño del archivo comprimido es igual al número de símbolos del archivo multiplicado por el tamaño promedio de los símbolos comprimidos ($\overline{n}_{C}$). Considerando lo antes expuesto, se puede concluir que RC = n_{SC}/n_{C} (relación n a 1) y RC = $(\overline{n}_{SC} - \overline{n}_{C})/\overline{n}_{C}$ (porcentaje).

Algoritmo de compresión probabilístico basado en la posición de los símbolos

Rincón y colaboradores (2009) propusieron un nuevo algoritmo de compresión probabilístico basado en la posición de los símbolos. Por ser este algoritmo el objeto de estudio de la presente investigación, se plantea a continuación el basamento teórico del mismo.

Diseño del Algoritmo

El diseño del algoritmo de compresión probabilístico propuesto se realizó utilizando como fundamento para la generación de los códigos, la posición de los símbolos como método para la creación de los nuevos códigos, evitando así la necesidad de escribir el mensaje nuevamente con diferentes códigos.

Como cualquier algoritmo probabilístico, el algoritmo propuesto se fundamenta en la generación de códigos óptimos, utilizando como parámetro la frecuencia de aparición de los símbolos en el mensaje. El concepto de información propuesto por Shannon permite asignar a los símbolos con mayor frecuencia, códigos de menor tamaño, permitiendo así minimizar la cantidad de bits para representar el mensaje, logrando algún grado de compresión.

El algoritmo de compresión propuesto utiliza la posición de los símbolos del mensaje en vez de generar un código de sustitución. Para tal fin se dispone de una cadena de bits la cual se encargará de mantener la posición de un determinado símbolo. Cada símbolo en el mensaje dispone de una cadena binaria, y es ésta información la que en último término se almacena.

El proceso de compresión (ver **Figura 1**) consiste en:

1) recorrer todo el mensaje realizando el cálculo de la frecuencia de los símbolos presentes en el mismo,

2) determinar cuál símbolo del mensaje tiene mayor frecuencia,

3) generar una cola de símbolos ordenándolos según su frecuencia de mayor a menor e inicializar la cadena de bits para el mensaje comprimido,

4) extraer de la cola el símbolo con mayor frecuencia,

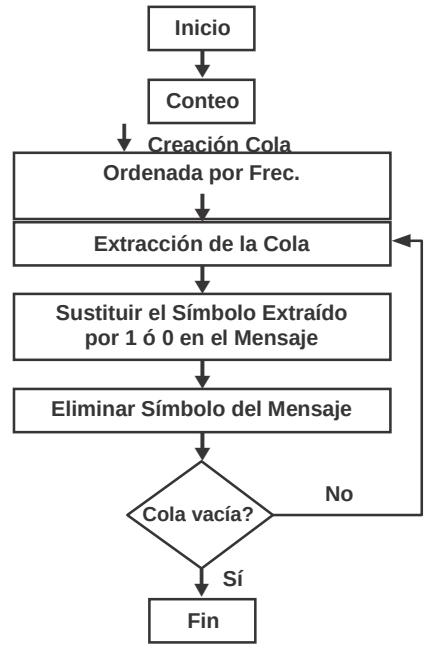
5) anexar a la cadena de bits del mensaje comprimido, una cadena de n bits (donde n es el tamaño del mensaje), la cual tendrá 1 en las posiciones del mensaje donde aparezca el símbolo extraído de la cola y 0 en donde éste no aparezca,

6) eliminar el símbolo extraído de la cola del mensaje original,

7) verificar si la cola de símbolos está vacía, en el caso de que sea cierto se finaliza el proceso y en el caso de ser falso ir al paso 4.

Figura 1

Proceso de compresión, algoritmo de compresión probabilístico basado en la posición de los símbolos

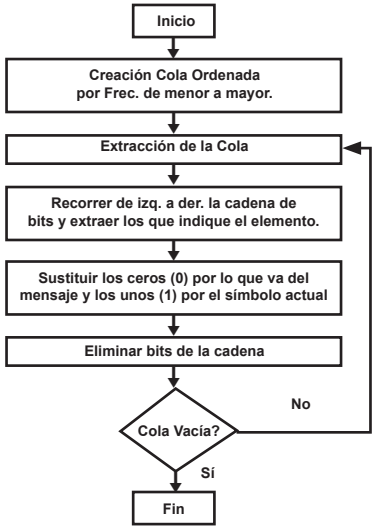


Proceso de Descompresión

Para retornar al mensaje original sólo es necesario conocer el tamaño en bits utilizado para representar cada símbolo presente en el mensaje y la cadena de bits con el mensaje comprimido. Como la secuencia binaria se generó en el orden de mayor a menor con respecto a las frecuencias, el proceso se realiza de forma inversa empezando por el último símbolo (el de menor frecuencia), sabiendo la cantidad de bits que ocupa, éstos se irán eliminando de la cadena binaria a medida que el proceso avance. Al mismo tiempo una cadena de caracteres se irá formando de acuerdo con el contenido de los bits correspondientes a cada símbolo (ver **Figura 2**).

Figura 2

Proceso de descompresión, algoritmo de compresión probabilístico basado en la posición de los símbolos



Análisis Matemático sobre el Rendimiento del Algoritmo

Para analizar el rendimiento del algoritmo, al igual que en la mayoría de los análisis matemáticos realizados a los algoritmos de compresión probabilísticos sin pérdida, existen 2 casos a considerar:

El mejor caso: cuando un mensaje está compuesto por n símbolos iguales, el número de bits que utiliza el algoritmo para generar el mensaje comprimido (sin considerar la información de control), será igual a n bits, dado a que el algoritmo representará el mensaje con n 1.

El peor caso: cuando un mensaje está compuesto por n símbolos tomados de un alfabeto conformado por m elementos y en dónde los símbolos aparecen de manera equiprobable en el mensaje, el rendimiento del algoritmo puede representarse mediante la siguiente serie matemática:

$$n+\left(n-\frac{n}{m}\right)+\left(n-\frac{2*n}{m}\right)+....+\left(n-\frac{(m-1)*n}{m}\right) \quad (1)$$

Simplificado tenemos:

$$\sum_{i=1}^{m} n-\sum_{i=o}^{m-1}\left(\frac{in}{m}\right)=n*m-\sum_{i=0}^{m-1}(i)*\frac{n}{m}=$$

$$n*m-\left(\frac{m*(m-1)}{2}\right)*\frac{n}{m}=$$

$$n*m-\left(\frac{n*(m-1)}{2}\right)=$$

$$n*\left(m-\frac{m-1}{2}\right)\ bits \quad (2)$$

Considerando que un archivo codificado en ASCII de n símbolos tiene n * 8 bits, podemos concluir que:

$$n*8 \geq n*\left(m-\frac{m-1}{2}\right)=$$

$$8 \geq \left(m-\frac{m-1}{2}\right)=8 \geq \left(\frac{2m-(m-1)}{2}\right)=$$

$$8 \geq \frac{m+1}{2}=m \leq 15 \quad (3)$$

El resultado de este análisis matemático permite establecer que para tamaños del alfabeto mayores a 15 símbolos, el algoritmo de compresión probabilístico basado en la posición de los símbolos ofrece una relación de compresión negativa (no comprime).

Es importante resaltar que el rendimiento (en número de bits para representar el mensaje comprimido) del algoritmo planteado en este análisis matemático, no considera el tamaño en bits de la cabecera que debe anexarse al mensaje comprimido para poder ser descomprimido.

Metodología utilizada

En la presente investigación utilizará una adaptación de la metodología propuesta en el trabajo titulado: Efecto del tamaño del archivo, la entropía y el tamaño del alfabeto en el rendimiento del algoritmo de Huffman (Rincón y col., 2008). Los pasos que conforman la metodología utilizada son:

1) Definición de los parámetros de la experimentación.

Se definieron los valores para las variables independientes tamaño del alfabeto y distribución probabilística del alfabeto para medir su efecto en las variables dependientes (medidas en segundos) relación de compresión y tiempo de compresión.

Tamaño del Alfabeto: cantidad de símbolos que conforman el mensaje a comprimir. Valores seleccionados: 4, 8, 12, 16 símbolos por alfabeto. Se seleccionaron valores menores o iguales a 16 a consecuencia del análisis matemático realizado por los investigadores que desarrollaron el algoritmo, donde concluyeron que para una tamaño del alfabeto mayor o igual a 16, la relación de compresión es negativa (no hay compresión).

Distribución probabilística de los símbolos en el alfabeto: comportamiento estadístico de la frecuencia de los símbolos en el mensaje. Valores seleccionados: equiprobable y aleatorio. Se seleccionaron estos valores porque el caso equiprobable es considerado como el peor caso posible (considerando su efecto en la entropía del mensaje) y aleatorio por ser la situación que normalmente se presenta al momento de comprimir un archivo. Para el caso aleatorio se utilizó la función random de C, que utiliza una distribución probabilística uniforme para generar los números.

2. Generación de los archivos de prueba:

Tomando en consideración el factor de variación de la variable independiente del estudio y el análisis estadístico a aplicar, se generaron la cantidad de archivos de prueba a comprimir. Se diseñó e implementó un generador de archivos donde se permite variar la cantidad de símbolos que componen el alfabeto y la distribución probabilística de los símbolos en el mensaje.

3. Ejecución del Algoritmo:

La implementación del algoritmo de compresión propuesto en un lenguaje de alto nivel, permitió aplicar el mismo a los archivos de prueba generados, con la finalidad de obtener los resultados para su posterior análisis. Se utilizó el lenguaje de programación C++, implementando el algoritmo de compresión en la plataforma Linux.

4. Análisis estadístico de los resultados:

En este trabajo se estudia el comportamiento de dos variables aleatorias dependientes (tiempo de compresión y relación de compresión) y dos variables aleatorias independientes (tamaño del alfabeto y distribución probabilística del alfabeto).

Para el análisis se decidió utilizar un diseño totalmente aleatorizado con tratamiento en un arreglo factorial 4 x 2, con dos factores: tamaño del alfabeto a cuatro niveles (4, 8, 12, y 16 símbolos) y distribución probabilística del alfabeto a dos niveles (0=aletorio y 1=equiprobable).

Se realizaron 5 repeticiones para cada interacción de los niveles de las variables independientes para un total de 40 repeticiones.

El modelo matemático que permite explicar el comportamiento de las variables dependientes es:

$$Y_{ijk}=\mu+T_i+D_j+TD_{ij}+E_{ijk} \qquad (4)$$

Para todo:

i=1,2,3,4 (niveles del tamaño del alfabeto)

j=1,2 (niveles de la distribución probabilística del alfabeto)

k=1,2,3,4,5 (repeticiones)

Donde:

Y_{ijk} es la observación de la variable (tiempo de compresión o relación de compresión) en la k-esima repetición del j-ésimo nivel de distancia en el i-ésimo nivel del tamaño.

μ es el promedio general de la variable

T_i es el efecto del i-ésimo nivel del tamaño del alfabeto

D_j es el efecto del j-ésimo nivel de la distribución probabilística del alfabeto.

TD_{ij} TD_{ij} es el efecto de la interacción del i-ésimo nivel del tamaño del alfabeto con el j-ésimo nivel de la distribución probabilística del alfabeto.

E_{ijk} es el error experimental

Los datos producto de las pruebas realizadas y el modelo estadístico diseñado fueron la base para ejecutar el procedimiento de análisis de varianza (ANADEVA), utilizando el paquete informático estadístico SAS.

Resultados obtenidos

La **Tabla No. 1**, presenta los resultados de todas las observaciones realizadas durante la presente investigación.

Son 40 observaciones de las cuales se tienen 10 por cada nivel del tamaño del alfabeto (4 niveles) y 20 por cada nivel de la distribución probabilística del alfabeto (2 niveles).

Tabla 1

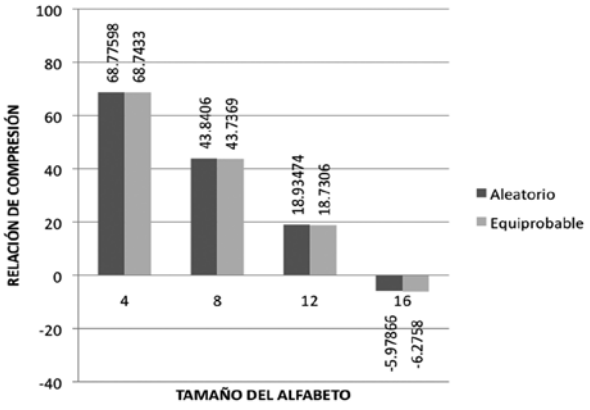
Resultados de las pruebas

OBS	TAMALF	DIST	TCOMP	RCOMP
1	4	0	11.91	687.740
2	4	0	11.92	687.621
3	4	0	11.92	687.769
4	4	0	11.92	687.825
5	4	0	11.89	687.844
6	8	0	10.85	438.269
7	8	0	10.81	438.088
8	8	0	10.82	439.029
9	8	0	10.82	438.350
10	8	0	10.83	438.294
11	12	0	10.54	189.219
12	12	0	10.50	189.362
13	12	0	10.52	189.525
14	12	0	10.50	189.633
15	12	0	10.52	188.998
16	16	0	10.40	-59.212
17	16	0	10.40	-60.365
18	16	0	10.39	-59.833
19	16	0	10.40	-59.810
20	16	0	10.41	-59.713
21	4	1	19.07	687.433
22	4	1	18.99	687.433
23	4	1	19.06	687.433
24	4	1	19.07	687.433
25	4	1	19.06	687.433
26	8	1	19.09	437.369
27	8	1	19.27	437.369
28	8	1	19.08	437.369
29	8	1	19.14	437.369
30	8	1	19.10	437.369
31	12	1	19.23	187.306
32	12	1	19.23	187.306
33	12	1	19.18	187.306
34	12	1	19.21	187.306
35	12	1	19.25	187.306
36	16	1	19.42	-62.758
37	16	1	19.32	-62.758
38	16	1	19.33	-62.758
39	16	1	19.32	-62.758
40	16	1	19.31	-62.758

En la **Figura 3**, se presentan los promedios de las observaciones discriminadas por el tamaño del alfabeto, considerando la variación de la distribución probabilística de los símbolos en el mensaje, para la variable dependiente Relación de Compresión.

Figura 3

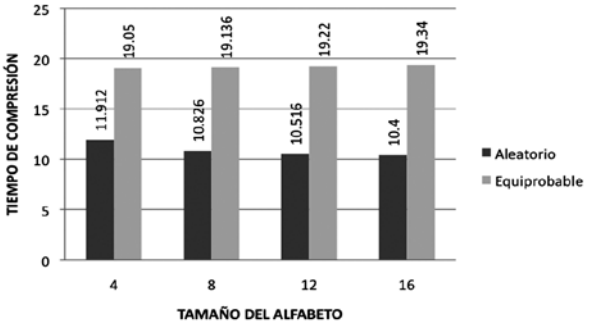
Valores promedio relación de compresión / Tamaño del alfabeto



En la **Figura 4**, presenta, los promedios de las observaciones discriminadas por el tamaño del alfabeto, considerando la variación de la distribución probabilística de los símbolos en el mensaje, para la variable dependiente Tiempo de Compresión.

Figura 4

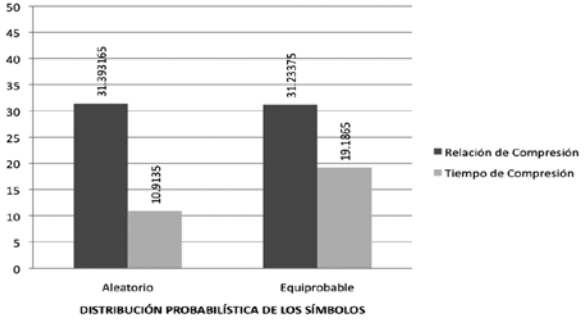
Valores promedio tiempo de compresión / tamaño del alfabeto



En la **Figura 5**, se muestran los promedios para las variables dependientes relación de compresión y tiempo de compresión, considerando la variación de la variable independiente distribución probabilística de los símbolos en el mensaje.

Figura 5

Valores promedio tiempo de compresión y relación de compresión / Distribución probabilística de los símbolos



Análisis de Varianza

La ejecución del análisis de varianza utilizando el modelo diseñado estableció la utilización de 40 observaciones y 4 niveles para la variable independiente tamaño del archivo (TAMALF) y 2 niveles para la variable independiente distribución probabilística de los caracteres (DIST) (ver **Figura 6**).

Figura 6

Valores de los niveles de las variables independientes. Fuente propia

Analysis of Variance Procedure
Class Level Information

Class	Levels	Values
TAMALF	4	4 8 12 16
DIST	2	0 1

Number of observations in data set = 40

En la **Figura 7**, se muestran para las diferentes combinaciones de las variables independientes tamaño del alfabeto y distribución probabilística del alfabeto, los valores promedios y la desviación estándar para las variables dependientes tiempo de compresión y relación de compresión.

Figura 7

Promedios y desviación estándar de las variables tiempo de compresión y relación de compresión

Analysis of Variance Procedure

Level of TAMALF	Level of DIST	N	TCOMP Mean	TCOMP SD	RCOMP Mean	RCOMP SD
4	0	5	11.9120000	0.01303840	68.7759800	0.00881516
4	1	5	19.0500000	0.03391165	68.7433000	0.00000000
8	0	5	10.8260000	0.01516575	43.8406000	0.03618363
8	1	5	19.1360000	0.07829432	43.7369000	0.00000000
12	0	5	10.5160000	0.01673320	18.9347400	0.02509428
12	1	5	19.2200000	0.02645751	18.7306000	0.00000000
16	0	5	10.4000000	0.00707107	–5.9786660	0.04098515
16	1	5	19.3400000	0.04527693	–6.2758300	0.00000000

Variable dependiente tiempo de compresión

El análisis de varianza para la variable dependiente tiempo de compresión revela diferencias significativas ($P < 0.01$) entre los niveles del tamaño del alfabeto, entre los niveles de la distribución probabilística del alfabeto y diferencias significativas para la interacción entre el tamaño del alfabeto y la distribución probabilística del alfabeto, lo cual evidencia que ambos factores actúan o producen su efecto de manera conjunta; es decir, dependen en su acción el uno del otro (ver **Figuras 8 y 9**).

Figura 8

Análisis de Varianza para la Variable Tiempo de Compresión

Analysis of Variance Procedure

Dependent Variable: TCOMP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	691.7863600	98.8266229	73477.04	0.000
Error	32	0.0430400	0.0013450		
Corrected Total	39	691.8294000			

R-Square	C.V.	Root MSE	TCOMP Mean
0.999938	0.243683	0.036674	15.0500000

Figura 9

Efecto de las Variables Independientes en el análisis de Varianza para la Variable Tiempo de Compresión

Analysis of Variance Procedure

Dependent Variable: TCOMP

Source	DF	Anova SS	Mean Square	F Value	Pr > F
TAMALF	3	2.5604600	0.8534867	634.56	0.0001
DIST	1	684.4252900	684.4252900	99999.99	0.0
TAMALF*DIST	3	4.8006100	1.6002033	1189.74	0.0001

En resumen, el comportamiento de la variable Tiempo de Compresión es afectado (P < 0.01) por el tamaño del alfabeto y la distribución probabilística del alfabeto

Variable Dependiente Relación de Compresión

El análisis de varianza para la variable dependiente relación de compresión revela diferencias significativas (P < 0.01) entre los niveles del tamaño del alfabeto, entre los niveles de la distribución probabilística del alfabeto y diferencias significativas para la interacción entre el tamaño del alfabeto y la distribución probabilística del alfabeto, lo que indica una acción de dependencia entre el tamaño del alfabeto y la distribución probabilística del alfabeto sobre la variable relación de compresión (ver **Figuras 10 y 11**).

Figura 10

Análisis de Varianza para la Variable Relación de Compresión

Analysis of Variance Procedure

Dependent Variable: RCOMP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	31154.61442	4450.65920	99999.99	0.0
Error	32	0.01479	0.00046		
Corrected Total	39	31154.62921			

R-Square	C.V.	Root MSE	RCOMP Mean
1.000000	0.068646	0.021496	31.3134530

Figura 11

Efecto de las Variables Independientes en el análisis de Varianza para la Variable Relación de Compresión

Analysis of Variance Procedure

Dependent Variable: RCOMP

Source	DF	Anova SS	Mean Square	F Value	Pr > F
TAMALF	3	31154.25992	10384.75331	99999.99	0.0
DIST	1	0.25415	0.25415	550.04	0.0001
TAMALF*DIST	3	0.10035	0.03345	72.40	0.0001

En resumen, la variable relación de compresión es afectada (P < 0.01) por el tamaño del alfabeto, por la distribución probabilística del alfabeto y por su interacción.

Pruebas de media de TUKEY

Tiempo de compresión, distribución probabilística

Los resultados de la prueba de TUKEY para la variable dependiente tiempo de compresión en relación a la variable independiente Distribución Probabilística se presentan en la **Figura 12**.

Figura 12

Prueba de TUKEY para la variable dependiente Tiempo de Compresión en relación a la variable Independiente Distribución Probabilística

Analysis of Variance Procedure

Tukey's Studentized Range (HSD) Test for variable: TCOMP

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 32 MSE= 0.001345
Critical Value of Studentized Range= 2.881
Minimum Significant Difference= 0.0236

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	DIST
A	19.1865	20	1
B	10.9135	20	0

Se observa que existen diferencias significativas para todos los niveles de la variable independiente Distribución Probabilística (0 = aleaterio, 1= equiprobable), obteniendo el mejor rendimiento para la distribución ya que presenta un valor de media menor (10.9135). Para el caso de la variable dependiente tiempo de compresión, mientras menor sea el tiempo mayor es el rendimiento de la técnica de compresión. Se concluye de estos resultados que el menor tiempo de compresión se obtiene utilizando una distribución probabilística aleatoria. El basamento teórico de esta conclusión se fundamenta en el hecho del incremento de la información promedio producto de la distribución equiprobable.

Tiempo de compresión, tamaño del alfabeto

Los resultados de la prueba de TUKEY para la variable dependiente tiempo de compresión en relación a la variable independiente Tamaño del Alfabeto se presenta en la **Figura 13**.

Figura 13

Prueba de TUKEY para la variable dependiente Tiempo de Compresión en relación a la variable Independiente Tamaño del Alfabeto

Analysis of Variance Procedure

Tukey's Studentized Range (HSD) Test for variable: TCOMP

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 32 MSE= 0.001345
Critical Value of Studentized Range= 3.832
Minimum Significant Difference= 0.0444

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	TAMALF
A	15.4810	10	4
B	14.9810	10	8
C	14.8700	10	16
C	14.8680	10	12

Se observa que existen diferencias significativas para los niveles de la variable independiente tamaño 4, 8, y el grupo 12 y 16 (para el cual no hay diferencia significativa), obteniendo el mejor rendimiento para el tamaño 12 y 16 ya que presenta el valor de media menor (14.868 y 14.87 respecti-

vamente). Para el caso de la variable dependiente tiempo de compresión, mientras menor sea el tiempo mayor es el rendimiento de la técnica de compresión. Se concluye de estos resultados que mientras mayor sea el tamaño del alfabeto, menor es el tiempo de compresión y mayor es el rendimiento del algoritmo de compresión. El basamento teórico de esta conclusión se fundamenta en la complejidad computacional implícita en el proceso de eliminación de un símbolo ya codificado en el mensaje original. Se observa que este comportamiento sólo ocurre para el caso en el que la distribución probabilística del alfabeto es aleatoria, debido a que en este caso la cantidad de elementos a eliminar para calcular los metacódigos es mayor, lo que aumenta el tiempo de compresión.

Relación de Compresión, Tamaño del Alfabeto

Los resultados de la prueba de TUKEY para la variable dependiente relación de compresión en función a la variable independiente distribución probabilística del alfabeto se presenta en la **Figura 14**.

Figura 14

Prueba de TUKEY para la variable dependiente relación de compresión en función a la variable Independiente distribución probabilística

Analysis of Variance Procedure

Tukey's Studentized Range (HSD) Test for variable: RCOMP

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 32 MSE= 0.000462
Critical Value of Studentized Range= 2.881
Minimum Significant Difference= 0.0138

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	DIST
A	31.39316	20	0
B	31.23374	20	1

Se observa que existen diferencias significativas para todos los niveles de la variable independiente Distribución Probabilística (0 = aleaterio, 1= equiprobable), obteniendo el mejor rendimiento para la distribución ya que presenta un valor de media mayor (31.39316). Para el caso de la variable dependiente relación de compresión, mientras mayor sea la relación mayor es el rendimiento de

la técnica de compresión. Se concluye de estos resultados que la mayor relación de compresión se obtiene utilizando una distribución probabilística aleatoria. El basamento teórico de esta conclusión se fundamenta en el hecho del incremento de la información promedio producto de la distribución equiprobable.

Relación de Compresión, Tamaño del Alfabeto

Los resultados de la prueba de TUKEY para la variable dependiente relación de compresión en función a la variable independiente Tamaño del Alfabeto se presenta en la **Figura 15.**

Figura 15

Prueba de TUKEY para la variable dependiente Relación de Compresión en función a la variable Independiente Tamaño del Alfabeto

Analysis of Variance Procedure

Tukey's Studentized Range (HSD) Test for variable: RCOMP

NOTE: This test controls the type I experimentwise error rate, but generally has a higher type II error rate than REGWQ.

Alpha= 0.05 df= 32 MSE= 0.000462
Critical Value of Studentized Range= 3.832
Minimum Significant Difference= 0.026

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	TAMALF
A	68.75964	10	4
B	43.78875	10	8
C	18.83267	10	12
C	–6.12725	10	16

Se observa que existen diferencias significativas para los niveles de la variable independiente tamaño 4, 8, 12 y 16, obteniendo el mejor rendimiento para el tamaño 4 ya que presenta el valor de media mayor (68.75964). Para el caso de la variable dependiente relación de compresión, mientras mayor sea la relación mayor es el rendimiento de la técnica de compresión. Se concluye de estos resultados que mientras menor sea el tamaño del alfabeto, mayor es la relación de compresión y mayor es el rendimiento del algoritmo de compresión.

Conclusiones

Como resultado del análisis de los datos obtenidos mediante la aplicación del algoritmo de compresión probabilístico basado en la posición a los diferentes archivos de texto generados variando los parámetros tamaño del alfabeto, y la distribución probabilística de los símbolos en el mensaje, se puede concluir lo siguiente:

La observación directa de los resultados permite determinar:

A medida que aumenta el tamaño del alfabeto, disminuyen los valores las variables independientes relación de compresión y tiempo de compresión. El comportamiento de la relación de compresión se explica por los conceptos asociados a la teoría de la información, mientras que el del tiempo de compresión se explica por la complejidad computacional del algoritmo.

El mejor rendimiento del algoritmo (mayor relación de compresión y menor tiempo de compresión) se obtuvo cuanto se trabajan con archivos con una distribución aleatoria de los símbolos en el mensaje. Este comportamiento se explica por los conceptos asociados a la teoría de la información.

Del análisis de varianza aplicado a los resultados se obtiene:

El modelo estadístico formulado refleja de manera correcta el comportamiento de las variables dependientes en función de las variables independientes.

El análisis de varianza para las variable dependientes revelaron diferencias significativas ($P < 0.01$) entre los niveles del tamaño del alfabeto, entre los niveles de la distribución probabilística del alfabeto y diferencias significativas para la interacción entre el tamaño del alfabeto y la distribución probabilística del alfabeto, lo cual evidencia que ambos factores actúan o producen su efecto de manera conjunta.

En la aplicación de la prueba de media de Tukey para la variable dependiente tiempo de compresión, se observó el mejor rendimiento para los mayores valores del tamaño del alfabeto y una distribución aleatoria de los símbolos, mientras que para la variable dependiente relación de compresión, el mejor rendimiento se obtuvo con el menor tamaño del alfabeto y una distribución aleatoria de los símbolos.

La investigación corroboró experimentalmente el análisis matemático propuesto por Rincón y colaboradores (2009), por lo que se avanza en la prosecución de una solución al problema de rendimiento del algoritmo de compresión probabilístico basado en la posición de los símbolos.

La generación de una cadena de bits *sparse*, fue identificada como el problema de la pérdida de rendimiento del algoritmo de compresión probabilístico basado en la posición de los símbolos producto del aumento del tamaño del alfabeto.

Bibliografía

Chan, Y. (2008). *A lossless coding scheme for encoding color-indexed video sequences*. 2008 International Conference on Neural Networks and Signal Processing. 676-681. doi: 10.1109/ICNNSP.2008.4590436

Fraenkel, A. y Klein S. (1985). Novel Compression of Sparse Bit-Strings-Preliminary Report. *A. Apostolico and Z. Galil, eds., Combinatorial Algorithms on Words*. Vol. 12, 169-183. New York, Springer-Verlag

Ling, Y.; Qian, C. y Wang, X. (2008). *A new lossless compression algorithm for vector maps*. International Symposium on Computer Science and Computational Technology, ISCSCT '08. 1:347-351, doi:10.1109/ISCSCT.2008.237

Marcelloni, F. y Vecchio, M. (2008). A simple algorithm for data compression in wireless sensor networks. *Communications Letters, IEEE*. 12(6):411-413. ISSN 1089-7798, doi:10.1109/LCOMM.2008.080300

Moffat, A. y Zobel, J. (1992). *Parameterised compression for sparse bitmaps*. In Proceedings of the 15th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Copenhagen, Denmark, June 21 - 24, 1992)

Rincón, C.; Acurero, A.; Bracho, D. y Jakymec, J. (2008). Efecto del tamaño del archivo, la entropía y el tamaño del alfabeto en el rendimiento del algoritmo de Huffman, *Revista CIENCIA*. Vol. 16, No. 2, 76 – 185

Rincón, C.; Rodríguez, D.; Acurero, A.; Bracho, D. y Jakymec, J. (2009). *Algoritmo de Compresión probabilistico basado en la teoría de la información*. Octava Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISCI 2009, Orlando – FL, US

Salomon, D. (2000). Prefix compression of sparse binary strings, *Crossroads*. Vol. 6, No. 3 (Mar. 2000), 22-25, DOI= http://doi.acm.org/10.1145/331624.331631. Más reciente consulta 55/10/2012

Wenjun, H.; Weimin, W. y Hui, X. (2006). *A loss less data compression algorithm for real-time database*. The Sixth World Congress on Intelligent Control and Automation, WCICA 2006, 2:6645-6648, doi:10.1109/WCICA.2006.1714368