CIENCIA 24(1), 41-53, 2016
Maracaibo, Venezuela

Optimización del algoritmo de compresión probabilístico basado en la posición de los símbolos

Carlos Rincón[1], Daylix Gómez[1], Aurely Leal[2,*], David Bracho[1] y Alfredo Acurero[1]

[1]Unidad de Redes e Ingeniería Telemática. [2]Unidad de Inteligencia Artificial y Computación Gráfica, Departamento de Computación, Facultad Experimental de Ciencias, Universidad del Zulia. Maracaibo, Venezuela

Recibido: 22-02-15 Aceptado: 12-01-16

Resumen

El objeto de estudio del presente trabajo tuvo como finalidad optimizar el rendimiento del algoritmo de compresión probabilístico basado en la posición de los símbolos. Las variables dependientes utilizadas para medir el rendimiento del algoritmo fueron el tiempo de compresión y la relación de compresión. El diseño del modelo estadístico seleccionado fue totalmente aleatorizado con tratamiento en un arreglo factorial 2 x 4 x 2, con tres factores: tipo de algoritmo (optimizado y original), tamaño del alfabeto a cuatro niveles (8, 12, 16, y 20 símbolos) y distribución probabilística del alfabeto a dos niveles (aleatorio y equiprobable). Al aplicar el análisis GLM a los resultados se obtuvo diferencias significativas para todas las variables independientes y su interacción en todas las variables dependientes, corroborando así el rendimiento alcanzado para el algoritmo de compresión probabilístico basado en la posición de los símbolos. La prueba de media 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 (16 y 20), mientras que para la variable relación de compresión, el mejor rendimiento se obtiene con la distribución aleatoria y el menor tamaño del alfabeto (8).

Palabras clave: compresión, posición, rendimiento, símbolos, alfabeto.

Optimization of a probabilistic compression algorithm based on the position of symbols

The purpose of the present work was to optimize the performance of a probabilistic compression algorithm based on the position of symbols. The dependent variables used to measure the algorithm performance were compression time and compression ratio. The statistical model designed was a totally randomized one with treatment on a factorial array 2 x 4 x 2 with three

*Autor para la correspondencia: aureleal.lozano@fec.luz.edu.ve

factors: type of algorithm (optimized and original), four leveled alphabet size (8, 12, 16, and 20 symbols) and alphabet probabilistic distribution at two levels (random and equiprobable). The results of the GLM procedure showed significant differences for all independent variables and their interactions on all dependent variables, corroborating the performance achieved for probabilistic compression algorithm based on the position of its symbols. Tukey's media test determined that for compression time, the best performance was obtained with random distribution and the higher alphabet size (16 and 20), while for compression ratio the best performance was obtained with random distribution and a lower alphabet size (8).

Key words: position, performance, symbols, alphabet.

Introducción

En la actualidad, la compresión de datos constituye un área de gran relevancia a nivel computacional, ya que aumenta en gran medida la capacidad de almacenamiento y transmisión de información de manera eficiente al reducir el tamaño de los datos. Wolff (1) propuso una teoría con la cual sugiere que la "computación es compresión", concretamente expone que compresión de datos puede interpretarse como un proceso de eliminación de la complejidad innecesaria (redundancia) de la información, y maximizando de esta manera la simplicidad, preservando al mismo tiempo tanto como sea posible de su poder descriptivo no redundante.

Rincón y colaboradores (2) diseñaron un algoritmo de compresión probabilístico sin pérdida basado en la teoría de la información propuesta por Shannon (3) en 1948. Sin embargo, este algoritmo presentaba un problema que incidía negativamente sobre la relación de compresión al aumentar el tamaño del alfabeto. Para alfabetos mayores a 15 símbolos no había compresión de los datos (la relación de compresión resultaba negativa).

Los autores, luego de un análisis matemático, identificaron la generación de una cadena de bits sparse como el problema de la pérdida de rendimiento del algoritmo. El propósito de esta investigación consistió en optimizar el algoritmo de compresión probabilístico basado en la posición de los símbolos, mediante la implementación del algoritmo LLRUN, por ser una herramienta efectiva para la compresión de este tipo de cadenas, logrando así que la relación de compresión sea siempre positiva, sin importar el tamaño del alfabeto.

Aspectos teóricos

Teoría de la Información: Un concepto fundamental en la teoría de la información es que la cantidad de

información contenida en un mensaje es un valor matemático bien definido y medible. En un artículo de 1948 titulado "Una teoría matemática de la comunicación", Claude E. Shannon (3) estableció un método matemático que permite cuantificar la información generada por una fuente de datos logrando determinar lo que hoy llamamos teoría de la información. Basándose en que todo evento tiene una probabilidad de ocurrencia, la cantidad de información de un mensaje, es inversamente proporcional a la probabilidad de ocurrencia ($I(x) = \log_u(1/P(x))$, siendo x el evento de estudio y u la base del logaritmo que representa la unidad de información (2=bit, e=nat, 10=hartley)). Por lo tanto, un evento con menor probabilidad de ocurrencia generará más información que un evento con mayor probabilidad de ocurrencia.

Compresión de Datos: Según Sayood (4), la compresión es "el arte o la ciencia de representar información de una forma compacta". Es un proceso que consiste en reducir, que a partir de fundamentos matemáticos o probabilísticos la repetición de los símbolos (redundancia), permitiendo así representar la información con una menor cantidad de bits, utilizando códigos óptimos.

Relación de Compresión: Salomon (5) define la relación de compresión como el tamaño total producido por la compresión de algún algoritmo entre el tamaño original del mensaje a comprimir. Permite determinar el rendimiento de un algoritmo de compresión.

Algoritmo de compresión probabilístico basado en la posición de los símbolos: En el 2009, Rincón y col. (2) plantearon un algoritmo de compresión que en lugar de generar códigos de sustitución, utiliza la posición de los símbolos y su frecuencia de aparición en el mensaje para la generación de cadenas de bits, manteniendo los principios fundamentales de la teoría de la información propuesta por Shannon (3), que permiten asignar a los símbolos de mayor frecuencia códigos de menor tamaño.

Rincón (6) expone: "A diferencia de los algoritmos de compresión tradicionales donde cada símbolo tiene su código basado en su frecuencia en el mensaje, el algoritmo de compresión probabilístico basado en la posición de los símbolos calcula los metacódigos considerando la posición en la que los símbolos aparecen en el mensaje. El concepto de metacódigo aplica a este algoritmo porque cada código asignado a un símbolo no sólo posee información sobre éste, sino que también contiene información sobre el contexto del mensaje".

Rendimiento del Algoritmo: Estudios anteriores a la investigación han comprobado que a medida que aumentaba el tamaño del alfabeto, la relación de compresión tendía a disminuir, siendo negativa cuando el tamaño de alfabeto sobrepasa los 15 símbolos.

Según Rincón y col. (7) "El comportamiento de la relación de compresión se explica por los conceptos asociados a la teoría de la información".

La pérdida de efectividad del algoritmo de compresión probabilístico basado en la posición de los símbolos, se debía al incremento progresivo del arreglo de bits resultantes que derivaba en la obtención de cadenas de bits sparse, como consecuencia del aumento del tamaño del alfabeto, lo que hacía de éste, un algoritmo poco eficiente para la compresión de cualquier fuente de información.

Análisis Teórico del LLRUN: El rendimiento de un algoritmo de compresión se mide determinando su mejor y peor caso. Estudios preliminares plantean un análisis teórico sobre el funcionamiento del LLRUN, con la finalidad de comprobar su posible uso en la compresión de cadenas de bits sparse. Los resultados arrojados por la investigación permitieron determinar que el método ofrece una relación de compresión teórica entre 87,5% y 50% (para el mejor y peor caso, respectivamente), obviando la información de cabecera necesaria para la reconstrucción del archivo, ya que estás representan un aporte mínimo de información adicional al tratarse de archivos de gran envergadura. El rendimiento del algoritmo analizado, permite catalogarlo como una herramienta efectiva para la compresión de cadenas de bits sparse.

Metodología utilizada

Se utilizó una adaptación de la metodología propuesta por Rincón y colaboradores (8) en el trabajo titulado "Efecto del tamaño del archivo, la entropía y el tamaño del alfabeto en el rendimiento del algoritmo de Huffman". Las etapas que conforman la metodología utilizada son:

Definición de los parámetros de la experimentación: Se definieron los valores para las variables independientes tipo de algoritmo, tamaño del alfabeto y distribución probabilística del alfabeto para medir su efecto en las variables dependientes relación de compresión y tiempo de compresión.

Tamaño del Alfabeto: cantidad de símbolos utilizados para la construcción del archivo a comprimir. Los valores seleccionados fueron: 8, 12, 16 y 20 símbolos por alfabeto.

Distribución Probabilística: representa el comportamiento estadístico de la frecuencia de aparición de los símbolos en el mensaje. Los valores seleccionados fueron aleatorio (cuando los símbolos del alfabeto aparecen con frecuencias diferentes en el mensaje) y equiprobable (cuando los símbolos que conforman el alfabeto aparecen con la misma frecuencia).

Relación de Compresión: métrica que permite determinar el grado de minimización para la representación de un archivo. Indica que tanto comprime o no el algoritmo.

Tiempo de Compresión: es la cantidad de unidades de tiempo que tarda el algoritmo implementado en realizar el proceso de codificación de un archivo.

Desarrollo e Implementación del Algoritmo Propuesto: Se procedió a implementar el algoritmo de compresión probabilístico basado en la posición de los símbolos en un lenguaje de alto nivel (C++), en una computadora con las siguientes especificaciones: Intel(R) Pentium(R) CPU 2020M 2.40GHz, 6 GB de RAM, HHDD 700GB, con plataforma Windows.

Construcción de los archivos de prueba y ejecución del algoritmo implementado sobre los archivos de prueba generados: 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. Con éste fin, 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. Esta construcción consistió en escribir los símbolos de manera aleatoria a partir del alfabeto seleccionado, considerando la distribución probabilística de la aparición de los mismos. Los archivos de prueba tienen un tamaño de 1050000 bytes.

Determinación del Modelo Matemático que explique el Comportamiento de las Variables Dependientes y Aplicación del Método Estadístico general linear model (GLM).

En esta fase se estudió el comportamiento de dos variables aleatorias dependientes (tiempo de compresión y relación de compresión) y tres variables aleatorias independientes (tipo de algoritmo, tamaño del alfabeto y distribución probabilística del alfabeto). Éste consistió en manipular las variables independientes para evaluar los efectos de éstas sobre las variables dependientes.

Para el análisis del método GLM se decidió utilizar un diseño totalmente aleatorizado con tratamiento en un arreglo factorial 2 x 4 x 2, con tres factores: tipo de algoritmo (0 = optimizado, 1 = original), tamaño del alfabeto a cuatro niveles (8, 12, 16, y 20 símbolos) y distribución probabilística del alfabeto a dos niveles (0 = aleatorio y 1 = equiprobable).

Se realizaron 4 repeticiones para cada interacción de los niveles de las variables independientes, las cuales se ejecutaron sobre los dos tipos de algoritmos para un total de 64 repeticiones. El modelo matemático que permite explicar el comportamiento de las variables dependientes es:

$$Y_{ijkl}=\mu + A_i + T_j + P_k + (AT_{ij}) + (AP_{ik})$$

$$+ (TP_{jk}) + (ATP_{ijk}) + E_{ijkl}$$

Para todo:

i=1, 2 (niveles del tipo de algoritmo).

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

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

l=1, 2, 3, 4 (repeticiones).

Donde:

Y_{ijkl} = es la observación de la variable (tiempo de compresión o relación de compresión) en la l-esima repetición del k-ésimo nivel de la distribución probabilística en el j-ésimo nivel del tamaño del alfabeto en el i-ésimo nivel del tipo de algoritmo.

μ = promedio general de la variable.

A_i = efecto del i-ésimo nivel del tipo de algoritmo.

T_j = efecto del j-ésimo nivel del tamaño del alfabeto.

P_k = efecto del k-ésimo nivel de la distribución probabilística del alfabeto.

(AT_{ij}) = interacción del i-ésimo nivel del tipo de algoritmo con el j-ésimo nivel del tamaño del alfabeto.

(AP_{ik}) = interacción del i-ésimo nivel del tipo de algoritmo con el k-ésimo nivel del tamaño de la distribución probabilistica del alfabeto.

(TP_{jk})= interacción del j-ésimo nivel del tamaño del alfabeto con el k-ésimo nivel del tamaño de la distribución probabilistica del alfabeto.

(ATP_{ijk}) = interacción del i-ésimo nivel del tipo de algoritmo con el j-ésimo nivel del tamaño del alfabeto con el k-ésimo nivel del tamaño de la distribución probabilística del alfabeto.

E_{ijkl} = error experimental (se asume la independencia entre los términos del error).

Resultados obtenidos

Después de obtener las 64 observaciones de las cuales se tienen 32 por tipo de algoritmo, 16 por cada nivel del tamaño del alfabeto (4 niveles) y 32 por cada nivel de la distribución probabilística del alfabeto (2 niveles), se obtuvieron los siguientes valores de las variables dependientes, reflejados en la tabla Nº 1.

Análisis y discusión de los resultados

Debido a que el presente trabajo tiene como finalidad presentar el redimiendo de la versión optimizada del algoritmo a objeto de estudio con respecto a la versión original, sólo se analizaron los parámetros del modelo estadístico propuesto (y sus interacciones) que incluyen la variable independiente tipo de algoritmo (ALG, ALG*TAM, ALG*DIST y ALG*TAM*DIST).

Variable Tiempo de Compresión: El análisis GLM para la variable dependiente tiempo de compresión revela diferencias significativas (alpha < 0.05) entre los niveles de tipo de algoritmo, los niveles del tamaño del alfabeto, los niveles de la distribución probabilística del alfabeto y en la interacción entre tamaño del alfabeto y la distribución probabilística, sin embargo para las interacciones ALG*TAM, ALG*DIST y ALG*TAM*DIST no existen diferencias significativas, debido a que la variable optimizada en este rendimiento fue la relación de compresión (ver figura 1).

Los resultados obtenidos a partir de las pruebas de media de Tukey realizadas sobre la variable dependiente tiempo de compresión en relación a las variables independientes que intervienen, muestran que existen diferencias significativas entre los dos tipos de algoritmo (ver figura 2A), obteniendo mejor rendimiento el método original ya que presenta un valor de media menor (81,00).

Para el caso del tiempo de compresión, mientras menor sea el tiempo mayor es el rendimiento de la técnica de compresión. A pesar de reflejar diferencias significativas, porcentualmente hablando la diferencia es mínima (2,04%). Con respecto al tipo de algoritmo y tamaño del alfabeto (ver figura 2C) y al tipo de algoritmo y distribución probabilística (ver figura 2B), el modelo muestra que no existen diferencias significativas.

En cuanto a la interacción de la variable dependiente tiempo de compresión en relación al tipo de algoritmo, tamaño del alfabeto y distribución probabilística (ver figura 3), a pesar que no existen diferencias significativas, se observa el mejor caso para el algoritmo original con tamaño 20 y distribución probabilística aleatoria.

Es importante resaltar que el objetivo principal de esta investigación fue optimizar el rendimiento del algoritmo propuesto en cuanto a la relación de compresión.

Fuente	GL	Suma de Cuadr	Prom Cuadrado	F Value	Pr > F
Modelo	15	314,3593750	20,9572917	9,60	<,0001
Error	48	104,7500000	2,1822917		
Total Corregido	63	419,1093750			

R-Square	Coeff Var	Root MSE	Prom TCOMP
0,750065	1,805318	1,477258	81,82813

Fuente	DF	Type I SS	Prom Cuadrado	F Value	Pr > F
ALG	1	43,8906250	43,8906250	20,11	<,0001
TAM	3	212,2968750	70,7656250	32,43	<,0001
DIST	1	21,3906250	21,3906250	9,80	0,0030
ALG*TAM	3	0,2968750	0,0989583	0,05	0,9870
ALG*DIST	1	0,0156250	0,0156250	0,01	0,9329
TAM*DIST	3	36,2968750	12,0989583	5,54	0,0024
ALG*TAM*DIST	3	0,1718750	0,0572917	0,03	0,9942

Figura 1. Resultado de la ejecución del procedimiento GLM para la variable dependiente Tiempo de Compresión
Fuente: Elaboración propia

Tukey Líneas de comparación para medias de mínimos cuadrados de ALG

LS-medias con la misma letra no son significativamente diferente

	TCOMP LSMEAN	ALG*	Número LSMEAN
A	82,65625	**0**	1
B	81,00000	**1**	2

A

Tukey Líneas de comparación para medias de mínimos cuadrados de ALG*DIST

LS-medias con la misma letra no son significativamente diferente

		TCOMP LSMEAN	ALG*	DIST**	Número LSMEAN
	A	83,2500	**0**	**0**	1
	A				
B	A	82,0625	**0**	**1**	2
B					
B	C	81,5625	**1**	**0**	3
	C				
	C	80,4375	**1**	**1**	4

B

Tukey Líneas de comparación para medias de mínimos cuadrados de ALG*TAM

LS-medias con la misma letra no son significativamente diferente

		TCOMP LSMEAN	ALG*	TAM	Número LSMEAN
	A	85,000	**0**	**8**	1
	A				
B	A	83,625	**0**	**12**	2
B	A				
B	A	83,500	**1**	**8**	5
B					
B	C	82,000	**1**	**12**	6
B	C				
B	C	81,500	**0**	**16**	3
	C				
D	C	80,500	**0**	**20**	4
D	C				
D	C	79,875	**1**	**16**	7
D					
D		78,625	**1**	**20**	8

C

*Tipo de Algoritmo (0 = optimizado, 1 = original)

**Distribución Probabilística (0 = aleatorio, 1 = equiprobable)

Figura 2. Prueba de Tukey para la variable dependiente Tiempo de Compresión en relación a las variables independientes Tipo de Algoritmo, Tipo de Algoritmo - Tamaño del Alfabeto y Tipo de Algoritmo – Distribución Probabilística
Fuente: Elaboración propia

Tukey Líneas de comparación para medias de mínimos cuadrados de ALG*TAM*DIST							
LS-medias con la misma letra no son significativamente diferente							
			TCOMP LSMEAN	ALG	TAM	DIST	Número LSMEAN
	A		85,50	**0**	**12**	**0**	3
	A						
	A		85,00	**0**	**8**	**1**	2
	A						
	A		85,00	**0**	**8**	**0**	1
	A						
B	A		83,75	**1**	**12**	**0**	11
B	A						
B	A		83,50	**1**	**8**	**1**	10
B	A						
B	A		83,50	**1**	**8**	**0**	9
B	A						
B	A	C	82,00	**0**	**16**	**0**	5
B	A	C					
B	A	C	81,75	**0**	**12**	**1**	4
B		C					
B		C	81,00	**0**	**16**	**1**	6
B		C					
B		C	80,50	**0**	**20**	**1**	8
B		C					
B		C	80,50	**0**	**20**	**0**	7
B		C					
B		C	80,50	**1**	**16**	**0**	13
B		C					
B		C	80,25	**1**	**12**	**1**	12
		C					
		C	79,25	**1**	**16**	**1**	14
		C					
		C	78,75	**1**	**20**	**1**	16
		C					
		C	78,50	**1**	**20**	**0**	15

Figura 3. Prueba de Tukey para la variable Tiempo de Compresión en relación a las variables independientes Tipo de Algoritmo, Tamaño del Alfabeto y Distribución Probabilística
Fuente: Elaboración propia

Variable relación de compresión

El análisis GLM para esta variable revela diferencias significativas (alpha=0.05) entre todas las variables independientes (y sus interacciones) del modelo estadistico propuesto, lo que indica una acción de dependencia entre tipo de algoritmo, el tamaño del alfabeto y la distribución probabilística

A partir de los resultados de la prueba de Tukey, se observa que para la variable dependiente en cuestión, existen diferencias significativas entre los dos tipos de algoritmo (ver figura 5A), obteniendo el mejor rendimiento para el algoritmo optimizado ya que presenta un valor de media mayor (53.06484). 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 se incrementó considerablemente la efectividad del algoritmo (194.20%), logrando los objetivos planteados en la investigación. Para los niveles de las variables independientes tipo de algoritmo y tamaño de alfabeto también existen diferencias significativas (ver figura 5C). La prueba de media de Tukey, establece 4 grupos, obteniendo el mejor rendimiento para el tamaño 8 del algoritmo optimizado. De estos resultados se concluye que mientras menor sea el tamaño del alfabeto, mayor es la relación de compresión y el rendimiento del algoritmo.

Es importante resaltar que para el algoritmo optimizado la relación de compresión es similar en todos los casos, mientras que para el original la relación de compresión disminuye considerablemente a medida que aumenta el tamaño del alfabeto. En la figura 5B, se observa que existen diferencias significativas para todos los niveles de las variables independientes tipo de algoritmo y distribución probabilística. La prueba de media de Tukey, establece 3 grupos, obteniendo el mejor rendimiento para la distribución aleatoria del algoritmo optimizado.

El basamento teórico de este resultado se fundamenta en el hecho del incremento de la información promedio producto de la distribución equiprobable. Es importante destacar que a pesar de este fundamento, para el algoritmo optimizado se logró obtener resultados similares en cuanto a la relación de compresión de ambos casos (aleatorio y equiprobable).

En la interacción de la variable dependiente relación de compresión en función al tipo de algoritmo, tamaño del alfabeto y distribución probabilística (ver figura 6), la prueba de Tukey, establece 6 grupos, siendo el mejor caso el algoritmo original con tamaño 8 y distribución probabilística aleatoria.

Es importante resaltar que todas las interacciones realizadas en relación al algoritmo optimizado, se encuentran en el grupo A de la prueba de media de Tukey. Esto implica que no existen diferencias significativas entre el mejor caso del algoritmo original y todos los casos del algoritmo optimizado, razón por la cual se concluye que se logró el objetivo principal de la investigación que fue optimizar el algoritmo propuesto en cuanto a la relación de compresión.

Fuente	GL	Sum Cuadrados	Prom Cuadrado	F Value	Pr > F
Modelo	15	4417,.71658	2944,98111	247,37	<,0001
Error	48	571,44587	11,90512		
Total Corregido	63	44746,16245			

R-Square	Coeff Var	Root MSE	Prom RCOMP
0,987229	9,705425	3,450380	35,55105

Fuente	DF	Type I SS	Prom Cuadrado	F Value	Pr > F
ALG	1	19630,91088	19630,91088	1648,95	<,0001
TAM	3	10483,94215	3494,64738	293,54	<,0001
DIST	1	2328,39761	2328,39761	195,58	<,0001
ALG*TAM	3	9165,12273	3055,04091	256,62	<,0001
ALG*DIST	1	2120,10035	2120,10035	178,08	<,0001
TAM*DIST	3	215,58643	71,86214	6,04	0,0014
ALG*TAM*DIST	3	230,65643	76,88548	6,46	0,0009

Figura 4. Resultado de la ejecución del procedimiento GLM para la variable dependiente Relación de Compresión
Fuente: Elaboración propia

Tukey Líneas de comparación para medias de mínimos cuadrados de ALG			
LS-medias con la misma letra no son significativamente diferente			
	RCOMP LSMEAN	ALG*	Número LSMEAN
A	53,06484	**0**	1
B	18,03725	**1**	2

A

Tukey Líneas de comparación para medias de mínimos cuadrados de ALG*DIST				
LS-medias con la misma letra no son significativamente diferente				
	RCOMP LSMEAN	ALG*	DIST**	Número LSMEAN
A	53,34096	**0**	**0**	1
A				
A	52,78873	**0**	**1**	2
B	29,82450	**1**	**0**	3
C	6,25000	**1**	**1**	4

C

Tukey Líneas de comparación para medias de mínimos cuadrados de ALG*TAM				
LS-medias con la misma letra no son significativamente diferente				
	RCOMP LSMEAN	ALG*	TAM	Número LSMEAN
A	54,269875	**0**	**8**	1
A				
A	53,441337	**0**	**12**	2
A				
A	52,421213	**0**	**16**	3
A				
A	52,126938	**0**	**20**	4
A				
A	49,544300	**1**	**8**	5
B	32,118075	**1**	**12**	6
C	5,806337	**1**	**16**	7
D	-15,319703	**1**	**20**	8

B

*Tipo de Algoritmo (0 = optimizado, 1 = original)

**Distribución Probabilística (0 = aleatorio, 1 = equiprobable)

Figura 5. Prueba de Tukey para la variable dependiente Relación de Compresión en relación al Tipo de Algoritmo, Tipo de Algoritmo - Tamaño del Alfabeto, Tipo de Algoritmo -Distribución Probabilística
Fuente Propia

Tukey Líneas de comparación para medias de mínimos cuadrados de ALG*TAM*DIST							
LS-medias con la misma letra no son significativamente diferente.							
			RCOMP LSMEAN	ALG	TAM	DIST	Número LSMEAN
	A		55,338600	**1**	**8**	**0**	9
	A						
	A		54,606575	**0**	**8**	**0**	1
	A						
B	A		53,971525	**0**	**12**	**0**	3
B	A						
B	A		53,933175	**0**	**8**	**1**	2
B	A						
B	A		52,911150	**0**	**12**	**1**	4
B	A						
B	A		52,613900	**0**	**16**	**0**	5
B	A						
B	A	C	52,228525	**0**	**16**	**1**	6
B	A	C					
B	A	C	52,171825	**0**	**20**	**0**	7
B	A	C					
B	A	C	52,082050	**0**	**20**	**1**	8
B		C					
B		C	45,486150	**1**	**12**	**0**	11
		C					
		C	43,750000	**1**	**8**	**1**	10

	D		18,750000	**1**	**12**	**1**	12
	D						
	D		17,862675	**1**	**16**	**0**	13
	E		0,610595	**1**	**20**	**0**	15
	E						
	E		-6,250000	**1**	**16**	**1**	14
	F		-31,250000	**1**	**20**	**1**	16

Figura 6. Prueba de Tukey para la variable Relación de Compresión en función de las variables independientes Tipo de Algoritmo, Tamaño del Alfabeto y Distribución Probabilística

Tabla No. 1
Resultados de las Pruebas

Obs	ALG	TAM	DIST	TCOMP	RCOMP
1	0	8	0	86	54.8009
2	0	8	0	84	54.4368
3	0	8	0	84	54.5584
4	0	8	0	86	54.6302
5	0	8	1	86	53.9268
6	0	8	1	85	53.9342
7	0	8	1	84	53.9342
8	0	8	1	85	53.9375
9	0	12	0	84	53.6171
10	0	12	0	84	53.3516
11	0	12	0	90	55.0426
12	0	12	0	84	53.8748
13	0	12	1	82	52.9133
14	0	12	1	81	52.9128
15	0	12	1	82	52.9154
16	0	12	1	82	52.9031
17	0	16	0	82	52.5658
18	0	16	0	84	52.7512
19	0	16	0	81	52.5655
20	0	16	0	81	52.5731
21	0	16	1	81	52.2179
22	0	16	1	81	52.2370
23	0	16	1	81	52.2299
24	0	16	1	81	52.2293
25	0	20	0	80	52.1026
26	0	20	0	80	52.1711
27	0	20	0	81	52.1453
28	0	20	0	81	52.2683
29	0	20	1	81	52.0936
30	0	20	1	80	52.0853

Tabla No. 1
(Continuación)

Obs	ALG	TAM	DIST	TCOMP	RCOMP
31	0	20	1	80	52.0688
32	0	20	1	81	52.0805
33	1	8	0	86	58.3426
34	1	8	0	82	52.0465
35	1	8	0	82	55.0564
36	1	8	0	84	55.9089
37	1	8	1	85	43.7500
38	1	8	1	83	43.7500
39	1	8	1	83	43.7500
40	1	8	1	83	43.7500
41	1	12	0	82	41.1727
42	1	12	0	82	35.3565
43	1	12	0	89	60.6034
44	1	12	0	82	44.8120
45	1	12	1	81	18.7500
46	1	12	1	80	18.7500
47	1	12	1	80	18.7500
48	1	12	1	80	18.7500
49	1	16	0	80	16.0532
50	1	16	0	83	22.9800
51	1	16	0	80	16.7836
52	1	16	0	79	15.6339
53	1	16	1	79	-6.2500
54	1	16	1	79	-6.2500
55	1	16	1	79	-6.2500
56	1	16	1	80	-6.2500
57	1	20	0	78	-8.3081
58	1	20	0	78	2.1091
59	1	20	0	79	-0.8894
60	1	20	0	79	9.5308
61	1	20	1	79	-31.2500
62	1	20	1	79	-31.2500
63	1	20	1	79	-31.2500
64	1	20	1	78	-31.2500

Fuente: Elaboración propia

Conclusiones

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

1. La observación directa de los resultados permite determinar:

a. 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.

b. 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.

2. Del análisis GLM aplicado a los resultados se obtiene:

a. El modelo estadístico formulado b. refleja fielmente el comportamiento de las variables dependientes en función de las variables independientes.

c. El análisis GLM para las variables dependientes revelaron diferencias significativas entre los niveles del tipo de algoritmo, tamaño del alfabeto, distribución probabilística, lo cual evidencia que los tres factores producen su efecto de manera conjunta. Para la variable dependiente tiempo de compresión sólo los efectos individuales de las variables independientes y la interacción del tamaño del alfabeto y la distribución probabilística fueron significativos, mientras que para la variable relación de compresión, tanto los efectos individuales de las variables independientes como las interacciones definidas en el modelo, resultaron significativas.

d. En la aplicación de la prueba de media de Tukey para la variable dependiente tiempo de compresión con respecto a la variable independiente tipo de algoritmo, se observó un mejor rendimiento por parte del algoritmo original (sólo 2.04% con respecto a la versión optimizada del algoritmo), mientras que para la variable dependiente relación de compresión, se observó un mejor rendimiento para la versión optimizada del algoritmo (194.20% con respecto a la versión original).

e. La prueba de media de Tukey realizada sobre la interacción del tipo de algoritmo, con el tamaño del alfabeto y la distribución probabilística de los símbolos,

permitió determinar que para la variable dependiente relación de compresión todas las interacciones del algoritmo optimizado se encuentran estadísticamente en el mismo nivel que el mejor de los casos del algoritmo original. La investigación corroboró experimentalmente el alcance de los objetivos planteados para este trabajo, al optimizar el algoritmo propuesto, logrando un rendimiento similar para todos los niveles del alfabeto del algoritmo optimizado.

Referencias bibliográficas

1. WOLFF, G., ***Computing as compression: the SP theory of intelligence***. CognitionResearch.org, UK. 2013.
2. RINCÓN, C., RODRÍGUEZ, D., ACURERO, A., BRACHO, D., JAKYMEC, J. ***Algoritmo de Compresión probabilístico basado en la teoría de la información.*** Octava Conferencia Iberoamericana en Sistemas, Cibernética e Informática, volumen I. pp. 320-326. CISCI 2009. Orlando – FL, US. (2009).
3. SHANNON, C. ***E. A mathematical theory of communication. Bell Systems Technical Journal,*** Revista Bell Systems Technical Journal volume 27:379-423, pp. 623-656. 1948.
4. SAYOOD, K. ***Introduction to Data Compression.*** Editorial Morgan Kaufmann Publishers, INC. San Francisco, California, USA. 1996.
5. SALOMON, D. ***DataCompression: The Complete Reference.*** 4ta Edición. Springer. Secaucus, NJ, USA. 2006.
6. RINCÓN, C. ***Efecto del tamaño del alfabeto en el rendimiento del algoritmo de compresión basado en la posición de los símbolos.*** Trabajo de Ascenso para profesor asociado. Facultad Experimental de Ciencias. Universidad del Zulia. Maracaibo, (Venezuela). (2011).
7. RINCÓN, C., BRACHO, D., ACURERO, A. ***Efecto del tamaño del alfabeto en el rendimiento de un algoritmo de compresión probabilístico***.Revista Enlace, Volumen 13, no. 3, pp. 83-99. 2012.
8. RINCÓN, C., ACURERO, A., BRACHO, D., JAKYMEC, J. ***Efecto del tamaño del archivo, la entropía y el tamaño del alfabeto en el rendimiento del algoritmo de Huffman.*** Revista CIENCIA, Volumen 16, No. 2, páginas 176 – 185. 2008.