Trabajo Práctico Nº2

Nombre, apellido y legajo de los integrantes: Mateo Etchepare (15093), Gregorio Firmani (15051), Franco Sardi([DNI]40831214).
Emails de contacto correspondientes: mateoetchepare@gmail.com , gregoriofirmani@gmail.com, fraansardi@gmail.com .
Número de grupo: 11
Universidad: Universidad Nacional de Mar del Plata.
URL GitHub: https://github.com/Greg1704/Teoria-de-la-Informacion/

Índice

Resumen

La Teoría de la Información es una propuesta teórica originada con un artículo de Claude E. Shannon, y trata de investigar y medir la información, su almacenamiento y comunicación, de una forma matemática y rigurosa. Una aplicación muy útil de ésta, especialmente hoy en día, es la compresión de información para su almacenamiento y transmisión.

En éste trabajo, a partir de un texto otorgado por la cátedra se comprimió un archivo de texto mediante los algoritmos de codificación de Huffman y Shannon-Fano, lo cual implicó generar un diccionario de palabras a partir del contenido de dicho archivo para usar como palabras origen. Luego de codificado el archivo hicimos el paso inverso, decodificando los archivos generados, y sacamos conclusiones en base a los resultados.

En la segunda parte, se analizaron tres canales de información para determinar si ese canal era confiable para transmitir la información, mediante el cálculo de ciertas propiedades particulares tales como el ruido, la pérdida, la información mutua, etc. Se pudo establecer el mejor canal dependiendo de la propiedad tenida en cuenta.

Introducción

Introducción a la codificación y compresión

La compresión de datos, también conocida como codificación de origen, es el proceso de codificación o conversión de datos de tal manera que consume menos espacio de memoria. La compresión de datos reduce la cantidad de recursos necesarios para almacenar y transmitir datos.

Se puede hacer de dos formas: compresión sin pérdida y compresión con pérdida. La compresión con pérdida reduce el tamaño de los datos al eliminar la información innecesaria, mientras que no hay pérdida de datos en la compresión sin pérdida.

En la primera parte de este trabajo práctico, se codificará un archivo de texto con más de 16000 palabras a código binario para así poder comprimir la información y conseguir un archivo que ocupe menos espacio en disco. Se compararán los dos métodos para poder comprobar su funcionalidad y determinar si uno de estos algoritmos es más eficiente que el otro.

Introducción a los canales de información

A diferencia de lo que se trabajó en el trabajo práctico anterior, en la segunda parte del trabajo práctico nª2 se modelará la transmisión y el procesamiento de la información, mediante el estudio de los canales de información. Un canal de información es un medio por el que se transmite la información desde la fuente de información al destino. Es importante analizarlos ya que en la transmisión de la información puede haber sucesos no deseados como la pérdida de información o el ruido.

La noción de los canales de información fue fundamental para el desarrollo de la telefonía fija moderna y las comunicaciones inalámbricas. Para las comunicaciones inalámbricas en particular, posteriormente se usaron mecanismos de corrección de ruido, tratando de hacer más confiable el canal de información.

Ahora, será posible que haya símbolos de distinta naturaleza en el centro emisor y en el centro receptor, y además, el tamaño del alfabeto de entrada puede o no ser igual al tamaño del alfabeto de salida.

Para analizar estos canales de información, se calcularán ciertas propiedades tales como la entropía "a priori", la entropía "a posteriori", la equivocación, información mutua, entre otros. A partir de los resultados se podrán obtener conclusiones sobre cada canal de información y se podrán caracterizar.

Desarrollo

Primera parte: Codificación y compresión

En esta primera parte, se nos presenta un archivo el cuál cuenta con un capítulo completo del Don Quijote, éste archivo fue comprimido a través de 2 algoritmos de compresión distintos, Huffman y Shannon-Fano. Una vez comprimidos, se guardaron en un nuevo archivo junto a la tabla necesaria para la traducción de éste, ya que posteriormente descomprimimos los mismos para demostrar la efectividad de los algoritmos, y para analizar si estos procesos generaron alguna pérdida en el archivo, y además, poder extraer conclusiones de si alguno de estos dos algoritmos es más eficiente que el otro.

Para analizar los resultados obtenidos, utilizamos los siguientes 3 métodos como criterio de comparación:

- Rendimiento

$$Rendimiento: \eta = \frac{H_r(S)}{L}$$

- Redundancia

$$Redundancia: 1 - \eta = \frac{L - H_r(S)}{L}$$

- Tasa de compresión de código

$$Tasa\ de\ compresión\ N: 1 = \frac{tamaño\ real}{tamaño\ comprimido}$$

Siendo:

- $H_r(S)$: entropía de la fuente S.
- L : longitud media del código.
- N : tamaño original / tamaño comprimido

Algoritmo de Huffman

El algoritmo de Huffman es un algoritmo para la construcción de códigos de Huffman, desarrollado por David A. Huffmann en 1952 y descrito en "A Method for the Construction of Minimum-Redundancy Codes". Éste algoritmo toma un alfabeto de "n" símbolos, junto con sus frecuencias de aparición asociadas, y produce un código de Huffman para ese alfabeto y esas frecuencias.

A continuación presentamos los resultados obtenidos luego de aplicar el algoritmo de Huffman.

Tamaño original del archivo	Tamaño del archivo comprimido (Sin la tabla)	Tamaño del archivo comprimido (Con la tabla)	Rendimiento	Redundancia	Tasa de compresión
93 KB (95,156 bytes)	19 KB (19578 bytes)	136 KB (139931 bytes)	0.997	0.003	4.86 : 1

(El peso del archivo comprimido con tabla se debe a que la tabla de traducciones de String a binario no pudo ser comprimida).

Algoritmo de Shannon-Fano

El algoritmo Shannon Fano es una técnica de codificación de entropía para la compresión de datos multimedia sin pérdidas. Creado por Claude Shannon y Robert Fano, asigna un código a cada símbolo en función de sus probabilidades de ocurrencia. Es un esquema de codificación de longitud variable, es decir, los códigos asignados a los símbolos serán de longitud variable.

A continuación presentamos los resultados obtenidos luego de aplicar el algoritmo de Shannon-Fano.

Tamaño original del archivo	Tamaño del archivo comprimido (Sin la tabla)	Tamaño del archivo comprimido (Con la tabla)	Rendimiento	Redundancia	Tasa de compresión
93 KB (95,156 bytes)	19 KB (19622 bytes)	136 KB (140095 bytes)	0.995	0.005	4.85 : 1

(El peso del archivo comprimido con tabla se debe a que la tabla de traducciones de String a binario no pudo ser comprimida).

Análisis de resultados de los algoritmos

A partir de los datos obtenidos, podemos deducir varias cosas. A primera vista, podemos ver que no existe una diferencia muy notable entre los resultados brindados por ambos algoritmos. Tanto Huffman como Shannon-Fano arrojan buenos resultados en general, teniendo Huffman resultados apenas superiores (ya sea por mayor rendimiento/tasa de compresión o que los archivos comprimidos pesan ligeramente menos) pero no lo suficientemente notables como para declararlo superior en este caso en particular.

Al tener un rendimiento tan alto, se pueden asegurar altas tasas de compresión, lo cual trae consigo una desventaja, que recae en el hecho de que se vuelve imposible detectar errores del código. Una redundancia baja también representa un desafío de gran magnitud a la hora de la detección y corrección de errores en la transmisión.

Ambos algoritmos comprimen el archivo en gran medida y posteriormente al descomprimirlo, se recupera correctamente el archivo original en ambos casos, como se explicará en la siguiente sección.

Decodificación

Usando la tabla de cada método, se logró descomprimir de vuelta los archivos comprimidos. Más allá de problemas al parsear los caracteres con tildes de parte del lenguaje de Java, ambos archivos decodificados poseían la misma información, ocupando el mismo espacio y poseyendo las mismas palabras en el orden que tenían en el archivo sin codificar. Ésto es consistente con el hecho de que ambos algoritmos sean algoritmos de compresión sin pérdida, caso contrario los archivos decodificados ocuparían menos espacio, habiendo perdido algo de información.

Segunda parte: Canales de Información

Para poder analizar los canales de información, se usarán las matrices brindadas por la cátedra y se completarán en base a nuestro número de grupo, que es el número 11.

El resultado final de la matriz de cada canal está en el anexo en **(1)**.

Tabla de resultados

Número de canal	Entropía "a priori" H(A)	Entropía de salida H(B)	Entropía "a posteriori" H(A/bj)	Entropía "afin" H(A, B)	Equivocación H(A/B)	H(B/A)	Información mutua I(A, B)
1	2,170	1,583	Tabla 1	3,737	2,154	1,566	0,017
2	1,948	1,993	Tabla 2	3,908	1,915	1,960	0,033
3	2,527	1,995	Tabla 3	4,490	2,496	1,963	0,031

Tabla 1

H(A/b = 1)	H(A/b = 2)	H(A/b = 3)
2,202	2,185	2,079

Tabla 2

H(A/b = 1)	H(A/b = 2)	H(A/b = 3)	H(A/b = 4)
1,925	1,934	1,985	1,820

Tabla 3

H(A/b = 1)	H(A/b = 2)	H(A/b = 3)	H(A/b = 4)
2,476	2,526	2,540	2,430

Análisis a partir de la tabla

1. El mejor canal con respecto a la pérdida de información es el canal 2, ya que su equivocación es menor respecto a los otros dos canales.
2. El mejor canal con respecto a la información mutua es el canal 2, ya que al menor ser su valor, más independientes son los símbolos de entrada con los símbolos de salida, ya que la gran cantidad de ruido del canal da incertidumbre de cuál es la entrada al observarse una salida determinada. Aún así, para que el canal sea **confiable**, el valor de la información mutua debería ser mayor.
3. La incertidumbre de los sucesos simultáneos H(A, B) son bastante mayor a la entropía "a priori" y la entropía de salida. Esto es lógico ya que en éstos canales el valor de la información mutua es muy bajo, y por definición esta propiedad es dependiente de la información mutua.
4. El canal con mayor entropía "a priori" es el canal 3. Esto se debe a que tiene 6 salidas (más que el canal 1 y el canal 2) y éstas 6 posibles salidas tienen probabilidades similares de ocurrencia.

5. En promedio, en ninguno de los canales se pierde información ya que H(A/B) < H(A).
6. En el canal 2, la pérdida es mayor al ruido, a diferencia de los canales 1 y 3.

Propiedades de la Información Mutua

1. I(A, B) >= 0

Al observar la tabla se puede ver el cumplimiento de esta propiedad para los 3 canales. La condición para que la información mutua sea 0; es que los símbolos de entrada y de salida sean estadísticamente independientes. Cuanto mayor sea el valor de la información mutua, mayor será el canal de información ya que menos incertidumbre habrá en la entrada y salida de la información ya que los símbolos de entrada y de salida habrá dependencia estadística.

2. I(A,B) = I(B, A) (reciprocidad de la información mutua).

Los resultados obtenidos con el programa (son verificables) son los siguientes:

Número de canal	I(A, B)	I(B, A)
1	0,017	0,017
2	0,033	0,033
3	0,031	0,031

Por lo que se ver a simple vista que la propiedad se cumple.

Lo que esta propiedad establece es que la cantidad de información que se obtiene de A gracias al conocimiento de B, es igual a la cantidad de información que se obtiene de B gracias al conocimiento de A.

3. I (A,B) = H (B) - H (B/A)

H(A,B)=H(A)+H(B)-I(A,B)

Para la primera ecuación (y primer canal): I(A, B) = 1,583 - 1,566 → I(A, B) = 0,017

Se observa que el resultado es correcto

Para la segunda ecuación (y primer canal): H(A,B) = 2,170 + 1,583 - 0,017 → H(A, B) = 1,566. Se observa que el resultado es correcto.

Esta propiedad se cumple para todos los canales, ya que las cuentas realizadas para todos los canales son las mismas ya que se hizo un programa lo más genérico posible para que funcione para cualquier canal de información.

Para la primera ecuación, se puede concluir que cuánta más pérdida haya, menos cantidad de información se obtendrá de A gracias al conocimiento de B, lo cuál es lógico.

Para la segunda ecuación, se puede concluir que la incertidumbre de sucesos simultáneos es la suma de la entropía "a priori" más la entropía de salida menos la información mutua. Es decir, que si el valor de la información mutua es alto, se puede obtener mucho conocimiento de la entrada o de la salida a partir del conocimiento de una de ellas. Si este valor es bajo, los sucesos son poco dependientes, lo cuál genera mucha incertidumbre que se den dos sucesos de manera simultánea. Esto explica lo mencionado en el ítem 3 de las conclusiones a partir de la tabla.

Conclusión

Conclusión de codificación y compresión

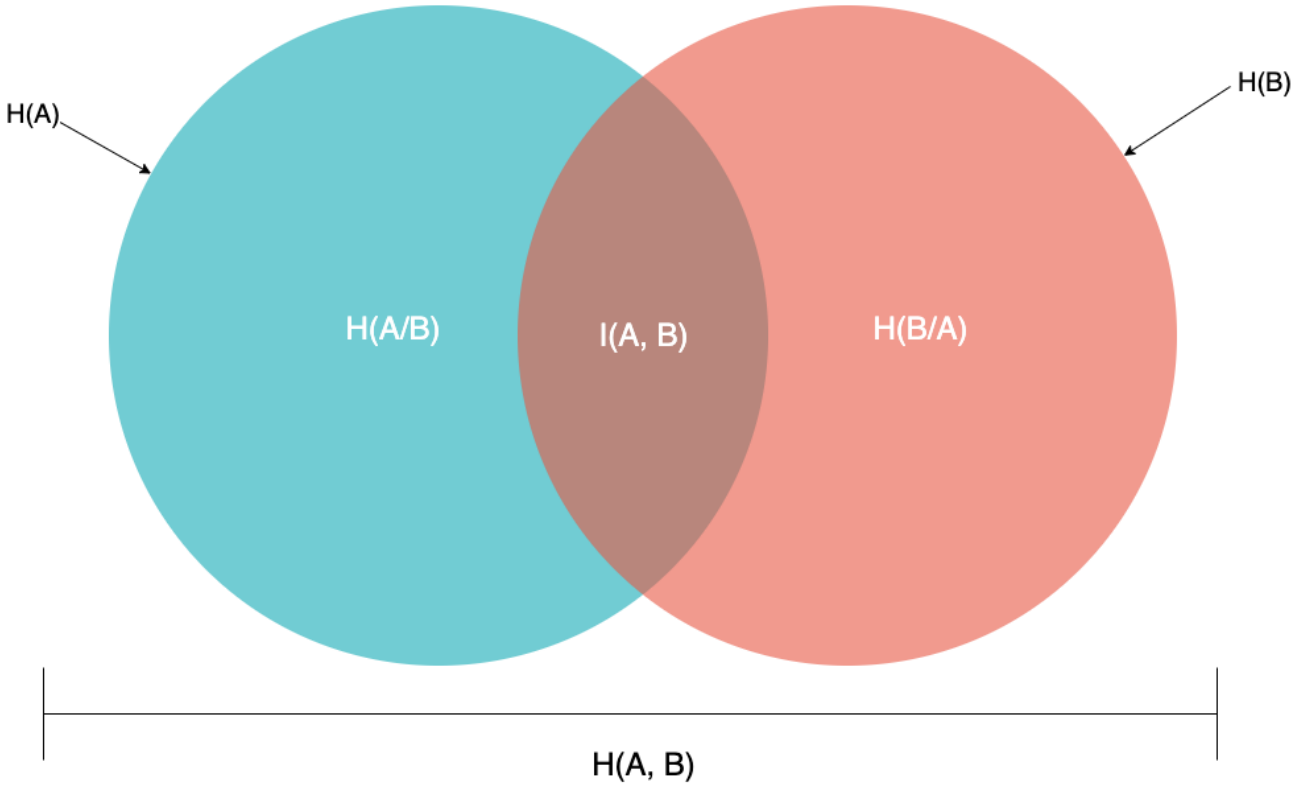
Para esta primera parte, podemos concluir que, tanto el Algoritmo de Huffman como el de Shannon-Fano son efectivos a la hora de comprimir alfabetos de "n" símbolos con probabilidades de aparición asociadas.

También se pudo demostrar que ambos algoritmos tienen un alto rendimiento, y consecuentemente, altas tasas de compresión, lo cuál refleja la efectividad de los algoritmos.

Por último, se pudo demostrar que ambos algoritmos son de compresión sin pérdida, lo cuál se refleja en el peso de los archivos decodificados, los cuales son casi idénticos al del original.

Conclusión de canales de información

Para concluir con la sección de Canales de Información, se puede resumir todo en el diagrama de Venn siguiente:



Como se explicó en la parte del desarrollo, el mejor canal de información será el que tenga un valor de información mutua más alto, ya que esto implica en el canal menos ruido y menos pérdida, por lo que hay menos incertidumbre en la transmisión de la información. Aparte de buscar el valor más alto de información mutua, se deben cumplir las propiedades mencionadas en el cuerpo del informe, que en el caso de los canales dados se cumplieron correctamente.

Anexo

(1):

Matriz de canal 1	B1	B2	B3
S1	0,3	0,3	0,4
S2	0,4	0,4	0,2
S3	0,3	0,3	0,4
S4	0,3	0,4	0,3
S5	0,3	0,4	0,3

Matriz de canal 2	B1	B2	B3	B4
S1	0,2	0,3	0,2	0,3
S2	0,3	0,3	0,2	0,2
S3	0,3	0,2	0,2	0,3
S4	0,3	0,3	0,3	0,1

Matriz de canal 3	B1	B2	B3	B4
S1	0,2	0,3	0,2	0,3
S2	0,3	0,3	0,3	0,1
S3	0,2	0,2	0,3	0,3
S4	0,3	0,3	0,2	0,2
S5	0,2	0,3	0,3	0,2
S6	0,2	0,3	0,3	0,2