Tema 1: Introducción a la compresión de datos

Rafael Molina
Depto. de Ciencias de la Computación
e Inteligencia Artificial
Universidad de Granada

Contenidos

- Compresión ¿por qué?. Algunos modelos.
- Técnicas de compresión
  - Compresión sin pérdida
  - Compresión con pérdida
  - Medidas de calidad
- Modelización y codificación
- Resumen
- Bibliografía
- Material sobre compresión
  - Libros
  - Cursos
  - Referencia histórica
  - Material didáctico
  - Sitios en Internet

0. Compresión ¿por qué?. Algunos modelos.

En los últimos años hemos visto una transformación (o revolución) en la forma que utilizamos para comunicarnos.

Esta transformación incluye: Internet, comunicaciones móviles y sin lugar a duda vídeo.

La compresión de datos es una de las llamadas tecnologías posibilitadoras (enabling technologies) para estos tres elementos que son parte de la revolución multimedia.

Sin compresión no tendría sentido poner imágenes, audio o vídeo en Internet, la calidad de las comunicaciones celulares no sería la misma y desde luego la TV digital no sería posible.

Podría decirse que la compresión de datos es "*El arte o la ciencia de representar información de una forma compacta*" [Sayood, 2000 página 1].

¿Por qué no nos centramos en el desarrollo de mejores técnicas de transmisión y almacenamiento?.

Mientras que podemos afirmar que la capacidad de transmisión y almacenamiento crece constantemente un corolario de la Primera Ley de Parkinson es que *las necesidades de transmisión y almacemamiento crecen a una velocidad que es el doble de la mejora en capacidad de transmisión y almacenamiento.*

Primera Ley de Parkinson: "el trabajo crece para llenar todo el tiempo disponible" [Sayood, 2000, página 2]

Uno de los primeros ejemplos de compresión de datos es el desarrollado por Samuel Morse a mediados del sigo XIX, (ver también el código Braille) en el que las letras enviadas por el telégrafo son codificadas utilizando puntos (.) y rayas (-) siendo la secuencia de símbolos más corta para las letras más frecuentes.

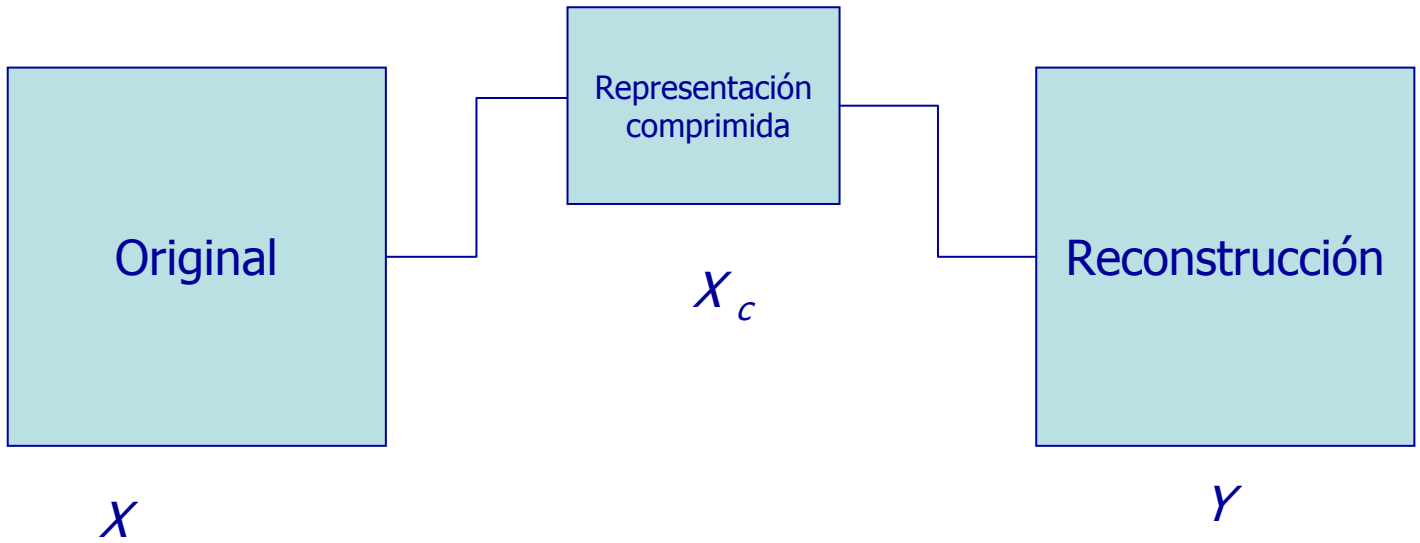
Este tipo de modelos para la compresión son llamados modelos estadísticos.

Existen además modelos que tienen en cuenta el proceso físico de formación del dato (voz) o modelos que analizan la capacidad de percepción del receptor (imágenes).

I. Técnicas de compresión

Cualquier algoritmo o técnica de compresión tiene dos partes:

- Un algoritmo de compresión que toma una entrada X y genera una representación X_c que necesita menos bits.
- Un algoritmo de reconstrucción que trabaja en la representación comprimida X_c y genera la reconstrucción Y.



Un esquema de compresión consta tanto de la parte de compresión como de la de reconstrucción.

Normalmente se utiliza el término algoritmo de compresión para denotar el esquema completo de compresión.

Los algoritmos de compresión se dividen en dos grandes clases:

- Algoritmos sin pérdida, en los que la entrada al codificador, X, y la salida del decodificador, Y, coinciden.

- Algoritmos con pérdida, que suelen proporcionar mayor compresión que los sin pérdida, pero en los que X e Y no coinciden, aunque se parecen (concepto a definir).

I.1. Compresión sin pérdida

Como indica su nombre, no hay pérdida de información. Se utiliza en aplicaciones donde no se permite ninguna diferencia entre los datos originales y los reconstruidos.

Son campos de aplicación, entre muchos otros:

- Compresión de texto
- Compresión de datos bancarios
- Compresión de datos empresariales/financieros
- Compresión de binarios/ejecutables
- Compresión de imágenes médicas

Técnicas estadísticas

1. Código de Huffman.
2. Codigos aritméticos.
3. Código de Golomb.

Técnicas basadas en diccionarios

1. LZW, LZ77.

Técnicas predictivas

1. PPM, Método de Burrows-Wheeler.

Estándares: Morse, Braille, Unix compress, gzip, zip, bzip, gif, bmp, jbig, jpeg sin pérdida,...

I.2. Compresión con pérdida

Estas técnicas llevan asociadas una pérdida de información, los datos originales no pueden, normalmente, ser recuperados exactamente.

Voz e imágenes (vídeo) son ejemplos claros de campos que toleran pérdida en la compresión.

Incluye técnicas como:

1. Cuantificación de vectores.
2. Wavelets.
3. Transformaciones por bloques.
4. Estándares: JPEG, JPEG 2000, MPEG (1, 2, 4).

I.3. Medidas de Calidad

¿Cómo evaluamos la calidad de un algoritmo de compresión?:

1. Complejidad del algoritmo,
2. Necesidades de memoria,
3. Tiempo de ejecución en una determinada plataforma,
4. Cantidad de compresión,
5. Cuanto se parece la reconstrucción a los datos originales

En este curso utilizaremos fundamentalmente los dos últimos criterios.

Razón de compresión = cociente entre el número de bits necesarios para representar los datos antes de la compresión y el número de bits necesarios para representar los datos después de la compresión.

Ejemplo: dada una imagen de tamaño 256x256 con un byte de información por píxel, si tras la compresión ocupa 16.384 bytes su compresión será 65.536: 16.384=4:1.

También podríamos medir la compresión utilizando la reducción en la cantidad de datos expresada como porcentaje del tamaño de los datos originales. En nuestro ejemplo sería una reducción del 75%.

También podemos usar el número medio de bits necesarios para representar cada dato. En nuestro ejemplo: 2 bits/píxel.

Cuando la compresión es con pérdida tenemos que utilizar, además de la cantidad de compresión obtenida, una medida para determinar la diferencia entre los datos originales y reconstruidos. Esta diferencia recibe el nombre de distorsión.

Las medidas de distorsión podrían ser a su vez basadas en criterios "matemáticos" o perceptuales. Lo discutiremos cuando veamos la compresión con pérdida.

II. Modelización y Codificación

Uno de los aspectos más importantes de la compresión es la caracterización (modelización) de los datos a comprimir.

Cualquier algoritmo de compresión puede dividirse en dos fases:

Modelización, donde extraemos información sobre la redundancia en los datos y describimos la redundancia como un modelo y

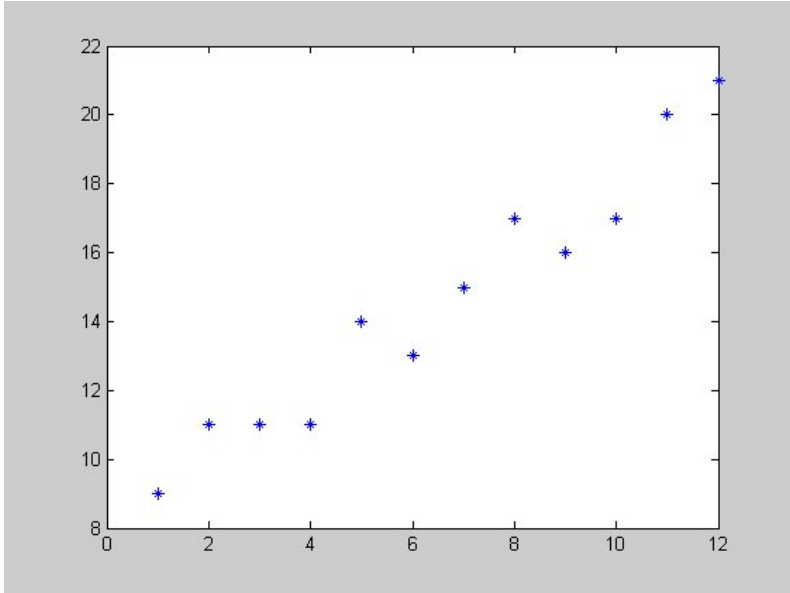
Codificación de la descripción del modelo y como los datos difieren del modelo.

Ejemplo II.1

Consideremos la secuencia (x_1,x_2,...

9 11 11 11 14 13 15 17 16 17 20 21

Si usamos la representación binaria de estos datos necesitaríamos 5 bits por dato.



Observando al gráfico un modelo aproximado para los datos sería

$$\hat{x}_n = n+8 \qquad n=1,2,...$$

Si consideramos los residuos

$$e_n = x - \hat{x}_n : \quad 0 \quad 1 \quad 0 \quad -1 \quad 1 \quad -1 \quad 0 \quad 1 \quad -1 \quad -1 \quad 1 \quad 1$$

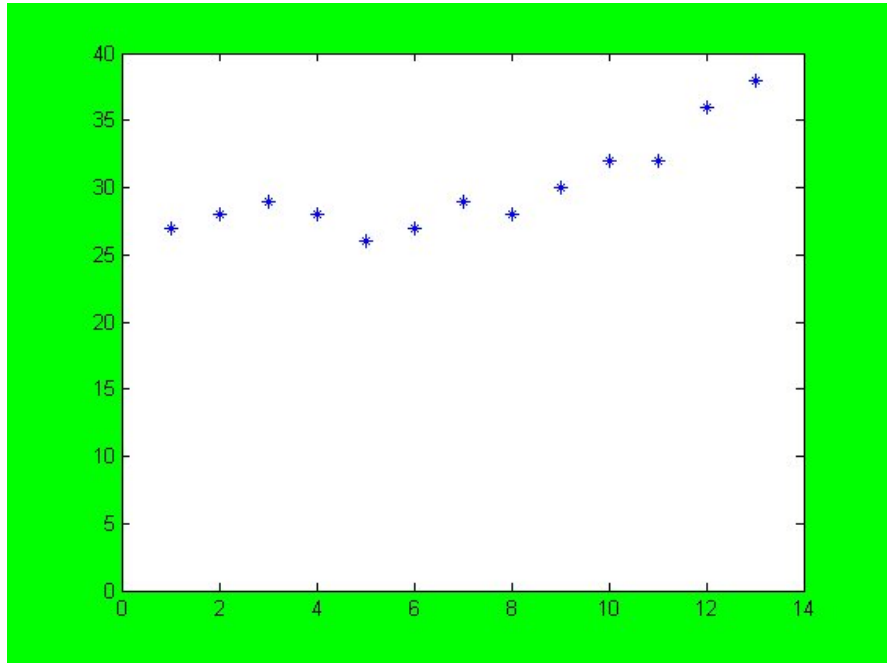
Podemos transmitir o almacenar el modelo y los residuos que pueden codificarse usando por ejemplo 00 para el -1, 01 para el 0 y el 10 para el 1.

Observemos que si enviamos o almacenamos sólo el modelo y los residuos los consideramos nulos tendríamos una codificación con pérdida.

Ejemplo II.2

Consideremos la secuencia (x_1,x_2,...

27 28 29 28 26 27 29 28 30 32 34 36 38



La sucesión no parece seguir una ley sencilla como en el ejemplo anterior.

Consideremos el modelo $x_{n+1} = x_n + d_{n+1}$ $n = 1,2,...$

Podemos transmitir o almacenar x_1 y todas las diferencias, es decir:

$$27 \quad 1 \quad 1 \quad -1 \quad -2 \quad 1 \quad 2 \quad -1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2$$

Discutiremos este tipo de técnicas que reciben el nombre de esquemas de codificación predictiva en el tema 6 para compresión sin pérdida y con posterioridad para compresión con pérdida.

Ejemplo II.3

Consideremos la sentencia siguiente:

abbarayaranbarraybranbfarbfaarbfaaarbaway

donde b denota espaco en blanco. Podemos usar tres bits por símbolo para codificarla. También podemos usar la siguiente tabla para codificarla con longitud variable:

a	1
b	001
b	01100
f	0100
n	0111
r	000
w	01101
y	0101

Si usamos estos códigos la secuencia será codificada usando 106 bits. Puesto que tenemos 41 símbolos el modelos utiliza 2.58 bits por símbolo. La razón de compresión es 3:2.56=1.16:1. Estos modelos que se basan en la redundancia estadística serán estudiados en los temas 3 y 4.

Usando texto hay palabras que se repiten frecuentemente, podemos construir una lista con ellas y representarlas por su posición en la lista. Estamos ante los esquemas de compresión basados en diccionarios que veremos en el tema 5.

A veces la redundancia es más evidente cuando miramos a grupos de símbolos. Estos modelos serán discutidos en el capítulo 4.

Por último, en determinadas situaciones será más conveniente descomponer los datos en un conjunto de componentes, podemos estudiar cada componente separadamente y usar un modelo para cada una de las componentes.

III. Resumen del tema

1. Introducción al problema de la compresión de datos y su necesidad,
2. Algo de terminología,
3. Distinción entre compresión sin pérdida y con pérdida,
4. Hemos definido la razón de compresión y comentado la existencia de criterios de similitud matemáticos y perceptuales,
5. Por último, hemos descrito algunos modelos que se utilizarán para comprimir los datos: estadísticos, diccionarios y predictivos.

IV Bibliografía

K. Sayood, "Introduction to Data Compression", Morgan and Kaufmann, 2000.

Material Complementario

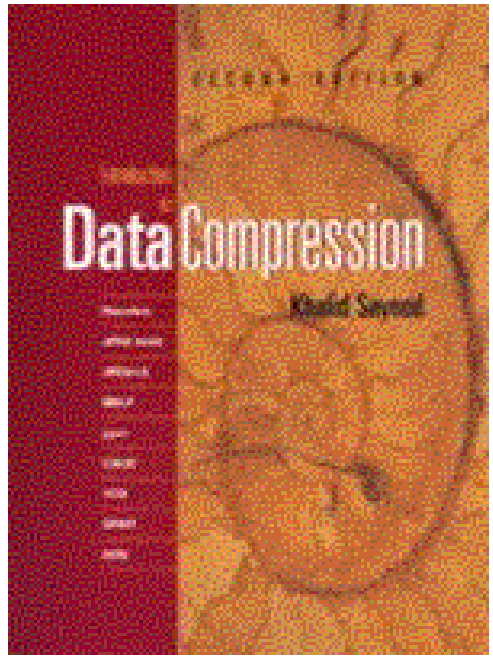
Tema 1 del curso de compresión de datos impartido en Chalmers University of Technology (Suecia), curso 2003-2004. (tema1_chalmers.pdf)

Tema 1 del curso de compresión de datos impartido en Stony Brook University (NY, USA), 2002-2003. (tema1_stony_univ.pdf).

C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal,* vol. 27, pp. 379-423 and 623-656, July and October, 1948.

V. Material sobre compresión

LIBROS



El material básico del curso será el libro:

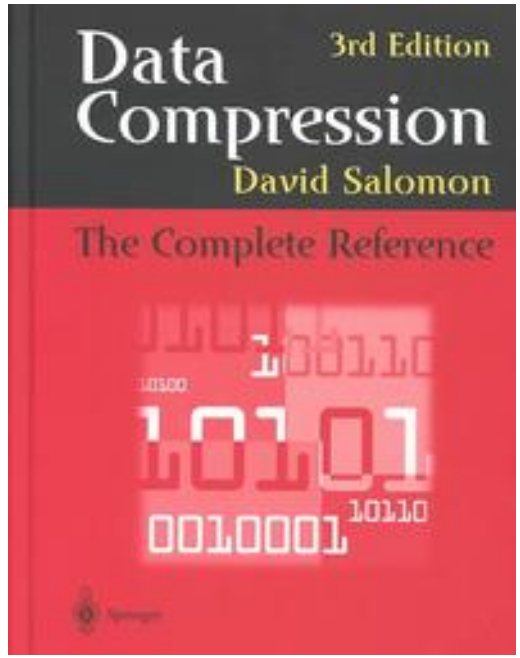
Introduction to Data Compression, 2nd edition,

by Khalid Sayood. ISBN 1-55860-558-4.

En la página web http://www.mkp.com buscar el libro (usando el nombre del autor) y seleccionar el título del libro.

El "companion site" en la página del libro será utilizado frecuentemente en las prácticas del curso.

Data Compression: The Complete Reference
3rd Edition
By David Salomon. Published by Springer (2004). ISBN 0-387-40697-2. LCCN QA76.9 D33S25 2004. xx+899 pages.

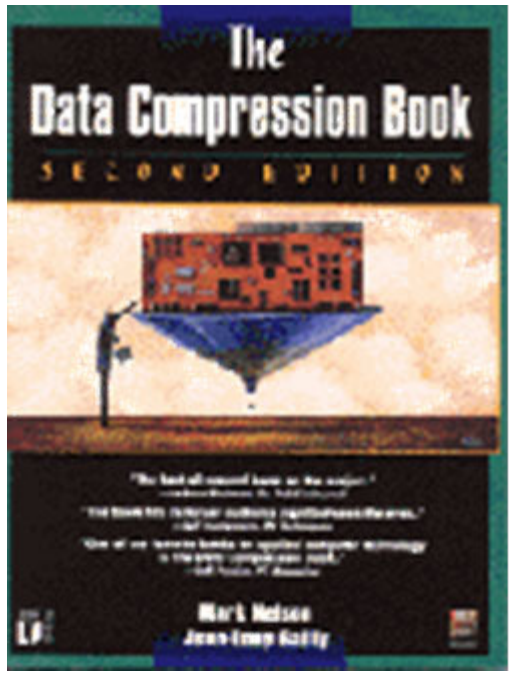


Información sobre el autor puede encontrarse en http://www.ecs.csun.edu/~dsalomon/

sobre el libro en http://www.ecs.csun.edu/~dsalomon/DC3advertis/DComp3Ad.html

The Data Compression Book
2nd edition

by Mark Nelson and Jean-loup Gailly,
M&T Books, New York, NY 1995 ISBN 1-55851-434-1 541 pages .



Información sobre los autores y el libro puede encontrarse en

http://www.marknelson.us/index.html
http://gailly.net

CURSOS

(La lista no es, en absoluto, exhaustiva. Realiza tus aportaciones)

Curso de compresión de datos impartido en Chalmers University of Technology (Suecia): material 2003-2004
http://www.s2.chalmers.se/undergraduate/courses0304/ess155/

Curso de compresión de datos impartido en Stony Brook University (NY, USA):
http://mnl.cs.stonybrook.edu/class/cse391/2003-spring/

REFERENCIA HISTORICA

Sin lugar a duda el trabajo de Shanon

C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal,* vol. 27, pp. 379-423 and 623-656, July and October, 1948.

debe estar presente en el material de la asignatura

http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html

Ver
http://cm.bell-labs.com/cm/ms/what/shannonday/work.html
sobre la importancia del trabajo de Shannon

MATERIAL DIDACTICO (La lista no es exhaustiva)

'**Interactive Data Compression Tutor**' es una ayuda para aprender compresión de datos basada en la web. Contiene información sobre los principios fundamentales y los mñetodos de la compresión de datos y algunos ejemplos de estos métodos.

Esta desarrollado por el departamento de ingeniería electrónica, eléctrica y de los computadores de la Universidad de Birmingham.
http://www.eee.bham.ac.uk/woolleysi/All7/body0.htm

Squeeze Page es una página diseñada para aprender algoritmos de compresión sin pérdida utilizando textos con gráficos y Java Applets.

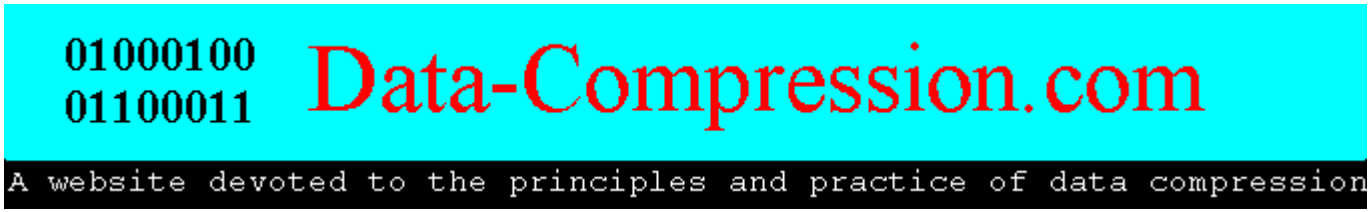
http://www.cs.sfu.ca/cs/CC/365/li/squeeze/

ha sido desarrollada en La Universidad Simon Fraser

http://www.cs.sfu.ca

SITIOS EN INTERNET

http://www.datacompression.info/



El portal de la asignatura:
contendrá los apuntes de teoría, las prácticas y el material complementario, así como toda la información sobre la asignatura.

http://www-etsi2.ugr.es/depar/ccia/ccd/