

A NOVEL APPROACH FOR DATA COMPRESSION IN E- MAIL

Popuri Ramesh Babu¹, Gonuguntla Rama Swamy², Daruvuri Ravi Kiran³,
Devireddy Srinivasa Kumar⁴

¹Donbosco P.G. College, Pulladigunta, Guntur

²St.Mary's Engineering College, Deshmukhi, Hyderabad

³Narasaraopet Engineering College, Narasaraopet, Guntur

⁴St.Mary's Women's Engineering College, Budampadu, Guntur

ABSTRACT

Along with the Web, electronic mail is one of the most popular Internet applications. E-mail is asynchronous- people send and read messages when it is convenient for them, without having to coordinate with other peoples' schedules. E-mail requires fully reliable data transfer, that is, no data loss. E-mail can make use of as much or as little bandwidth as happens to be available. Development of data compression algorithms has come close on the heels of the popularity of e-mail application. Two compression algorithms, Huffman Coding and Lempel-Ziv-Welch (LZW) are used in this study to simulate e-mail application

KEYWORDS: *Compression, Huffman Coding, Link Utilization, channel and e- mail application.*

DATA COMPRESSION

One of the hottest imaging technologies is compression. In the early stages, compression was limited to JPEG, GIF and Group 3 fax. Compression technologies have multiplied. Now there are also Group 4, PNG, wavelet and fractal compression. Compression is applied to still images, audio and video. Three reasons to compress: save storage space, conserve bandwidth, and speed up application software.

Data compression has important application in the areas of data transmission and data storage. Many data processing applications require storage of large volumes of data, and the number of such applications is constantly increasing as the use of computers extends to new disciplines. At the same time, the proliferation of computer communication networks is resulting in massive transfer of data over communication links. Compressing data to be stored or transmitted reduces storage and/or communication costs. When the amount of data to be transmitted is reduced, the effect is that of increasing the capacity of communication channel. Similarly, compressing a file to half of its original size is equivalent to doubling the capacity of the storage medium. It may then become feasible to store the data at a higher, thus faster, level of the storage, hierarchy and reduce the load on the input/output channels of the computer system

HUFFMAN CODING

Huffman coding converts the pixel brightness values in the original image to a new variable-length codes, based on their frequency of occurrence in the image. Huffman coding is a statistical data-compression technique. This technique will reduce the average code length used to represent the symbols of data-brightness values. The Huffman coding ensures that the longest codes get assigned to the least frequent brightness and vice versa.

The brightness values are first listed in descending order of frequency of occurrence. The two at the bottom of the list (least frequent) are paired together into a node with a joint probability- the probabilities are combined. This new node is labeled 0 and 1. The next two lowest frequencies of occurrences are determined and paired. The new node gets the joint probability and is labeled 0 and 1. This continues until all brightness are paired.

Example:

A string of characters to be transmitted is 's₁ through s₈'. The relative frequency of each character is as follows:

$$s_1, s_2 = 0.25, \quad s_3, s_4 = 0.14, \quad s_5, s_6, s_7, s_8 = 0.055$$

According to Shannon's formula, the minimum average number of bits/character (entropy H [22]) is 8

$$H = - \sum_{i=1} p(s_i) \log_2 p(s_i) \text{ bits per codeword}$$

$$H = - (2 * 0.25 \log_2 0.25) + 2 * 0.14 \log_2 0.14 + 4 * 0.055 \log_2 0.055$$

$$= 1 + 0.794 + 0.921 = 2.715 \text{ bits/codeword}$$

Codeword Generation:

Huffman Code generation is presented in **Table -1**.

s₁ 0.25	s₁ 0.25	s₁ 0.25	s₁ 0.25	s₃₄ 0.28	s₂₅₆₇₈ 0.47	s₁₃₄ 0.53 (1)	Root s₁₂₃₄₅₆₇₈ 1.00
s₂ 0.25	s₂ 0.25	s₂ 0.25	s₂ 0.25	s₁ 0.25	s₃₄ 0.28 (1)	s₂₅₆₇₈ 0.47 (0)	
s₃ 0.14	s₃ 0.14	s₃ 0.14	s₅₆₇₈ 0.22	s₂ 0.25 (1)	s₁ 0.25 (0)		
s₄ 0.14	s₄ 0.14	s₄ 0.14	s₃ 0.14 (1)	s₅₆₇₈ 0.22 (0)			
s₅ 0.14	s₇₈ 0.11	s₇₈ 0.11 (1)	s₄ 0.14 (0)				
s₆ 0.055	s₅ 0.14 (1)	s₅₆ 0.11 (0)					
s₇ 0.055 (1)	s₆ 0.055 (0)						
s₈ 0.055 (0)							

Weight order = 0.055 0.055 0.055 0.055 0.11 0.11 0.14 0.14 0.22 0.25 0.28 0.47 0.53

Huffman Code:

Symbol	s ₁	s ₂	s ₃	s ₄	s ₅	s ₆	s ₇	s ₈
p(s _i)	0.25	0.25	0.14	0.14	0.055	0.055	0.055	0.055
Codewords	10	01	111	110	0001	0000	0011	0010
#bits	2	2	3	3	4	4	4	4

Average number of bits/codeword using Huffman coding is:

$$2(2 * 0.25) + 2(3 * 0.14) + 4(4 * 0.055) = 2.72 \text{ bits/codeword that is, } 99.8\% \text{ of Shannon's value.}$$

DATA COMPRESSION APPLICATIONS TO NETWORKS

Two most common applications: **(1) Data storage:** A body of data is compressed before it is stored on some digital storage device, for example, a computer disk or tape. This process allows more data to be placed on a given device. When data is retrieved from the device, it is decompressed; and **(2) Data communications:** Communication lines that are commonly used to transmit digital data include cables between a computer and storage devices, phone lines, and satellite channels. A sender can compress data before transmitting it and the receiver can decompress the data after receiving it.

Case Study: E-mail

Typical compression rates used for simulation for e mail applications are shown in **Table** below.

Algorithms	Huffman Coding	LZW
Applications		
E-mail	Compression rate: 40%	Compression rate: 60%

The compression rates for Huffman coding and LZW are 40% and 60%, respectively. In each case, the message size is 100 bytes, warm-up length (specifies the time in the simulation when the application will begin to collect data) is 10 seconds and replication length (the number of seconds of simulated time during which statistics are collected for each report) is 50 seconds. The bandwidth used for simulation run is 64kbps for each compression algorithm. Telecommunication traffic is often described as a Poisson process. The number of messages in successive time intervals has been observed. The distribution of the number of observations in an interval is Poisson distributed. The parameter entered for the Poisson distribution is the interarrival time (IAT, the time from the start of one message to the start of the next). The total number of simulation runs for both algorithms are 20. Each simulation ran for approximately 2 minutes. The simulation run statistics for compressed and uncompressed data for the first algorithm, Huffman Coding are shown in **Tables 1 and 2**, respectively.

Examination of the data in **Tables 1 and 2** show that as the bandwidth increases:

1. IAT decreases
2. Packets/Second increase (see **Figures1 and 2**)
3. Bits/Second increase
4. Simulation run time is the same for bandwidth runs and the message size used for E-mail is small.
5. The rate of transmission for compressed data at each bandwidth is less than for uncompressed data.

As the tables show, when IAT decreases, the number of packets/second increases linearly in both data types. This is depicted in **Figures 1& 2**.

Bandwidth: 64kbps, Replication Length: 50s					
Message Size: 100 bytes = 800 bits					
Huffman Code (40%)	Bandwidth Used for Simulation				
	50%	60%	70%	80%	90%
Inter-Arrival Time (IAT) Poisson Distribution (poi)	0.0425	0.0354	0.0303	0.0263	0.0238
Packets/Second(pps)	24	28	33	38	43
Bits/Second(bps)	19200	22400	26400	30400	34400
Simulation Time (min)	2	2	2	2	2

Table 1: E-mail (Huffman) Simulation Run Statistics for Compressed Data

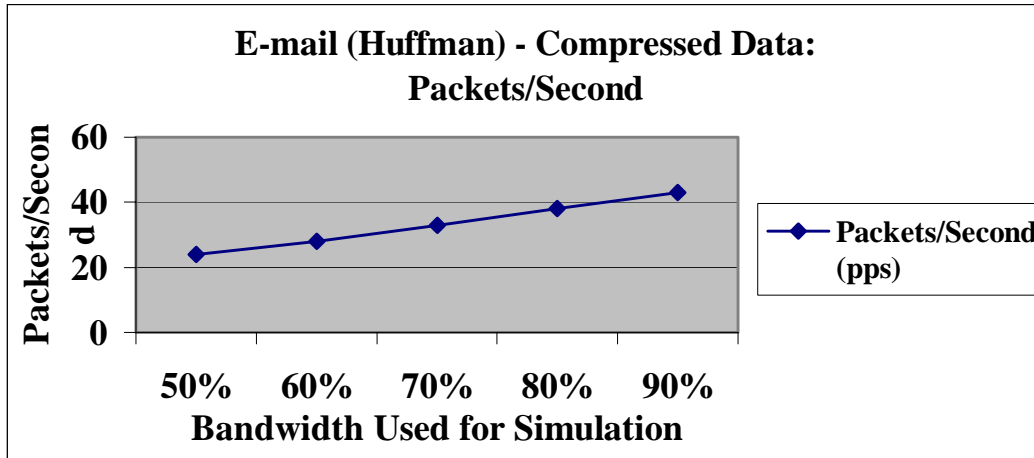


Figure 1: E-mail (Huffman) - Compressed Data: Packets/Second

Bandwidth: 64kbps, Replication Length: 50s					
Message Size: 100 bytes = 800 bits					
Huffman Code (40%)	Bandwidth Used for Simulation				
	50%	60%	70%	80%	90%
InterArrival Time (IAT) Poisson Distribution (poi)	0.025	0.021	0.018	0.016	0.014
Packets/Second(pps)	40	48	56	64	72
Bits/Second(bps)	32000	38400	44800	51200	57600
Simulation Time (min)	2	2	2	2	2

Table 2: E-mail (Huffman) Simulation Run Statistics for Uncompressed Data

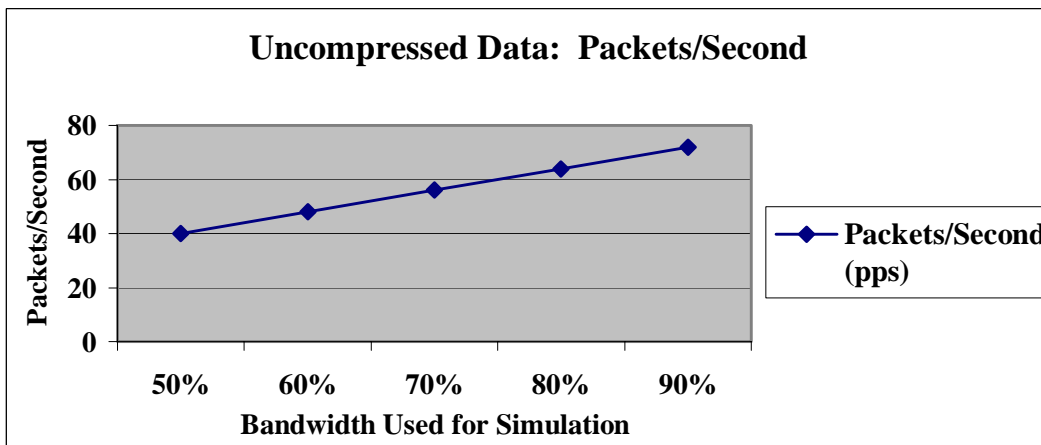


Figure 2: E-mail (Huffman) - Uncompressed Data: Packets/Second

HUFFMAN CODING

Huffman Coding is chosen for E-mail compression because text compression seems natural for it. In text, we have a discrete alphabet that has relatively stationary probabilities. For example, the probability model for a particular message will not differ significantly from that of another message.

Reports are produced automatically at the end of each simulation run. COMNET III generates a statistical report of Link Utilization for various processing nodes as shown in **Tables 3 and 4** for compressed and uncompressed data, respectively. The number of frames delivered is equal to the number of packets delivered from each link for E-mail application.

ANALYSIS

- **Compressed Data**

The number of packets delivered from Mumbai to Trivandrum is 1223 at 50% bandwidth, but at 60% bandwidth, it is 1191, a decrease of 3 percent. As shown in **Table 3**, the link utilization rate went down as well by 0.03% at 60%, from 30.58% to 29.78%. The number of packets delivered at 70% increased to 1567, up by 31.5% from 60% bandwidth, but decreased to 1224 at 80% and again increased to 1613 at 90%. From Mumbai to Bangalore link, notice a gradual increase from 50% to 60% bandwidth, but a decrease at 70% and an increase at 80% and 90%. From Hyderabad to Bangalore and Trivandrum to Hyderabad, however, the number of packets delivered increases as bandwidth increases. The link utilization also goes up for both links.

Application: E-mail (64kbps), Huffman Code (40%)					
Message Size = 100 bytes, Replication Length = 50s,					
Simulation Time = 2min					
Compressed Data	Bandwidth Used for Simulation				
Link	50% 19.2kbps	60% 22.4kbps	70% 26.4kbps	80% 30.4kbps	90% 34.4kbps
From Mumbai-Trivandrum					
Frames/Packets Delivered	1223	1191	1567	1224	1613
Packets/second	24.460	23.820	31.340	24.480	32.260
Bytes Delivered	122300	119100	156700	122400	161300
KBPS Delivered	19.568	19.056	25.072	19.584	25.808
Utilization (%)	30.58	29.78	39.18	30.60	40.33
From Mumbai-Bangalore					
Frames/Packets Delivered	1592	1629	1593	1707	1867
Packets/second	31.840	32.580	31.860	34.140	37.340
Bytes Delivered	159200	162900	159300	170700	186700
KBPS Delivered	25.472	26.064	25.488	27.312	29.872
Utilization (%)	39.83	40.73	39.85	42.70	46.68
From Hyderabad-Bangalore					
Frames/Packets Delivered	1347	1365	1545	2257	2463
Packets/second	26.940	27.300	30.900	45.140	49.260
Bytes Delivered	134700	136500	154500	225700	246300

KBPS Delivered	21.552	21.840	24.720	36.112	39.408
Utilization (%)	33.68	34.13	38.63	56.43	61.60
From Trivandrum-Hyderabad					
Frames/Packets Delivered	1308	1433	1611	1889	2321
Packets/second	26.160	28.660	32.220	37.780	46.420
Bytes Delivered	130800	143300	161100	188900	232100
KBPS Delivered	20.928	22.928	25.776	30.224	37.136
Utilization (%)	32.70	35.83	40.28	47.23	58.03

Table 3: E-mail (Huffman) - Links: Channel Utilization/Utilization by Application for Compressed Data

UNCOMPRESSED DATA

There is an increase in the number of packets delivered from Mumbai-Trivandrum from 1515 at 50% bandwidth to 3275 at 70% bandwidth, as shown in **Table 4**. The number of packets delivered at 80% (3049) and 90% (2863) bandwidths decreases by 7% and 6%, respectively. The utilization of Mumbai-Trivandrum link increases from 50% to 70% and decreases at 80% and 90% bandwidths. From Mumbai-Bangalore link notice a gradual increase in the number of packets delivered from 1747 (at 50%), to 2738 (at 60%). However, the number of packets delivered at 70% is 2441, a decrease of 11%. At 80% (3032) and 90% (3364) bandwidths the number of packets delivered increases again by 24% and 11%, respectively. Similarly, an increase in the number of packets delivered from Hyderabad-Bangalore link is observed, 2272 at 50%, 2664 at 60%, 2118 at 70%, 3014 at 80% and 3655 at 90%. The utilization rate of both Mumbai-Bangalore and Hyderabad-Bangalore links decreases at 70%. From Trivandrum-Hyderabad link the number of packets delivered and utilization rate increase linearly from 50% bandwidth to 90% bandwidth.

Application: E-mail (64kbps), Huffman Code (40%)					
Message Size = 100 bytes, Replication Length = 50s, Simulation Time = 2min					
Uncompressed Data	Bandwidth Used for Simulation				
Link	50%	60%	70%	80%	90%
	32kbps	38.4kbps	44.8kbps	51.2kbps	57.6kbps
Mumbai-Trivandrum					
Frames/Packets Delivered	1515	2299	3275	3049	2863
Packets/second	30.300	45.980	65.500	60.980	57.260
Bytes Delivered	151500	229900	327500	304900	286300
KBPS Delivered	24.240	36.784	52.400	48.874	45.808
Utilization (%)	37.88	57.50	81.88	76.25	71.60
Mumbai-Bangalore					
Frames/Packets Delivered	1747	2738	2441	3032	3364
Packets/second	34.940	54.760	48.820	60.640	67.280
Bytes Delivered	174700	273800	244100	303200	336400
KBPS Delivered	27.952	43.808	39.056	48.512	53.824
Utilization (%)	43.68	68.45	61.03	75.83	84.13
Hyderabad-Bangalore					
Frames/Packets Delivered	2272	2664	2118	3014	3655
Packets/second	45.440	53.280	42.360	60.280	73.100
Bytes Delivered	227200	266400	211800	301400	365500
KBPS Delivered	36.352	42.624	33.888	48.224	58.480

Utilization (%)	56.80	66.60	52.95	75.35	91.40
Hyderabad-Trivandrum					
Frames/Packets Delivered	1933	2294	2641	2806	3523
Packets/second	38.660	45.880	52.820	56.120	70.460
Bytes Delivered	193300	229400	264100	280600	352300
KBPS Delivered	30.928	36.704	42.256	44.896	56.368
Utilization (%)	48.33	57.33	66.05	70.15	88.08

Table 4: E-mail (Huffman) - Links: Channel Utilization/Utilization by Application for Uncompressed Data

Thus First row in Tables 3 and 4 represents packets delivered during the simulation. They fluctuate for some links and increase as the bandwidth increases for others. This fluctuation could be a function of network delay and/or link/node failure.

Fluctuations in:

- Packets/Second (Packets Delivered/Replication Length, 50s);
- Bytes Delivered (Packets Delivered * Message Size);
- KBPS (Packets/Second * Message Size * 8bits); and
- Link Utilization

can be seen as reflections of the fluctuations in packets delivered. Thus the utilization rate for compressed data is consistently lower than for uncompressed data. This implies that compression permits transmission of larger volume of data in shorter intervals.

E-mail (Huffman) Link	Link Utilization (%)	
	Compressed Data	Uncompressed Data
From Mumbai=Trivandrum		
50%	30.58	37.88
60%	29.78	57.50
70%	39.18	81.88
80%	30.60	76.25
90%	40.33	71.60
From Mumbai-Bangalore		
50%	39.83	43.68
60%	40.73	68.45
70%	39.85	61.03
80%	42.70	75.83
90%	46.68	84.13
From Hyderabad-Bangalore		
50%	33.68	56.80
60%	34.13	66.60
70%	38.63	52.95
80%	56.43	75.35
90%	61.60	91.40
From Trivandrum-Hyderabad		
50%	32.70	48.33
60%	35.83	57.33
70%	40.28	66.05
80%	47.23	70.15
90%	58.03	88.08

Table 5: E-mail (Huffman) - Comparison of Utilization (%) for Compressed & Uncompressed Data

		Ranges	
Application	Compression Algorithm	Compressed Data	Uncompressed Data
E-mail	Huffman		
Packets Delivered		330 – 750	225 – 610
Packets Dropped		129 – 347	161 – 435
Drop Rate (%)		30 – 75	41 – 142

CONCLUSION

Simulation runs ranged over applications like e-mail. Links and Message and Response Sources are used in simulating the results. Five bandwidth percentages are used to observe the variation in utilization rate (Links), drop rate and packets-dropped/packets-delivered ratio (Message and Response Sources). Channel utilization rate, in the case of links, is an accurate reflection of the difference in efficiency between transmission of compressed and uncompressed data.

REFERENCES

- [1]Ahamed, Syed V. and Lawrence, Victor B (1997). **Intelligent Broadband Multimedia Networks**. Norwell MA, Kluwer Academic Pub.
- [2]Ahamed, Syed V. and Lawrence, Victor B (1997). **Design and Engineering of Intelligent Communication Systems**. Norwell, MA, Kluwer Academic Pub
- [3] Aggarwal, J.K., Duda, Richard O. and Rosenfeld, Azriel (1977). **Computer Methods in Image Analysis**. New York, IEEE Press
- [4] Andrews, H.C. and Hunt, B.R. (1977). **Digital Image Restoration**. Englewood Cliffs, NJ,Prentice-Hall
- [5] Anttalainen, Tarmo (1999). **Introduction to Telecommunications Network Engineering**. Norwood, MA, Artech House, Inc
- [6]Black, Uyles D. (1989). **Data Networks: Concepts, Theory, and Practice**. Englewood Cliffs, NJ,Prentice-Hall.
- [7]Boisseau, Marc, Demange, M & Munier, J.-M. (1994). **High-Speed Networks**. Chichester,England, John Wiley & Sons Ltd.
- [8]Bose, Kausik (1994). **Information Networks in India: Problems and Prospects**. New Delhi, Ess Pub
- [9]Castleman, Kenneth R. (1979). **Digital Image Processing**. Englewood Cliffs, NJ, Prentice-Hall.
- [10]Chen, W.H. & Pratt, W.K. (1984). **Seene Adaptive Coder**, IEEE Transactions on
- [11]Chowdary, Th Communications, COM-32:225-232, March 1984. (1988). **Telephone in Rural Areas: An Indian Experience**. Telematics and Informatics, March, 1988, pp.29-37.
- [12]Comer, Douglas E. (1999). **Computer Networks and Internets**. Upper Saddle River, NJ,Prentice-Hall
- [13]Cooperman, Michael, Paige, A., and Sieber Richard W. (1989). **“Broadband Video Switching.”**
- [14]Cutler, C.C. **Differential Quantization for Television Signals**. U.S. Patent 2.605.361 July 29, 1952.
- [15]Davies, D.E.N., Hilsum, C. and Rudge, A.W., ed. (1993). **Communications After AD2000**. London, Chapman & Hall.
- [16]Deasington, R.J. (1984). **A Practical Guide to Computer Communications and Networking**, 2nd ed. Chichester, England, Ellis Horwood
- [17]Deniz, Dervis Z. (1994). **ISDN and Its Application to LAN Interconnection**. New York, McGraw-Hill.
- [18]Ekstrom, Michael P (1984). **Digital Image Processing Techniques**. New York, Academic Press.
- [19]Falkner, Matt (1984). **Modeling ATM Networks with COMNET III**. Version, 1.0, August 30,1996, CACI
- [20]Garzia (1999). **Network Modeling Simulation & Analysis**. Electrical Engineering &Electronics Series, vol.61
- [21]Gonzalez, Rafael C. and Woods, Richard E. (1993). **Digital Image Processing**. Reading, MA,Reading, MA Addison-Wesley

- [22]Graham, Deryn and Barrett, Anthony (1997). **Knowledge-Based Image Processing Systems**. London, Springer-Verlag
- [23]Green, William B. (1983). **Digital Image Processing**. New York, Van Nostrand Reinhold Co.
- [24]Gupta, V.S. (1999). **Communication Technology, Media Policy and National Development**. New Delhi, India, Concept Pub. Co.
- [25]Hall, Ernest L. (1979). **Computer Image Processing and Recognition**. New York, Academic Press.
- [26]Miller, Michael J. and Ahamed, Syed V. (1988). **Digital Transmission Systems and Networks, vol. II: Applications**. Rockville, MD, Computer Science Press.
- [27]Murphy, R.J. "Bert". (1987). **Telecommunications Networks: a Technical Introduction**. Indianapolis, Indiana, Howard W. Sams & Co.Press.
- [28]Watson, Andrew B., ed. (1993) **Digital Images and Human Vision**. Cambridge, MA, The MIT
- [29]Wayner, Peter (2000) **Compression Algorithms for Real Programmers**. San Diego, CA, Academic Press.
- [30]Ziv, J. & Lempel, A. (1977). **A Universal Algorithm for Data Compression**, IEEE Transactions on Information Theory, IT-23(3): 337-343, May. 1977.
- [31]Ziv, J. & Lempel, A. (1978). **Compression of Individual Sequences via Variable Rate Coding**, IEEE Transactions on Information Theory, IT-24(5): 530-536, Sept. 1978.