データ圧縮ソフトを用いた情報概念の学習

南 雲 秀 雄
新潟青陵大学福祉心理学科

Learning the Concept of Information Theory Using Data Compression Software

Hideo NAGUMO
NIIGATA SEIRYO UNIVERSITY
DEPARTMENT OF SOCIAL WELFARE AND PSYCHOLOGY

Abstract

We attempted to teach the concept of information theory to junior college students using data compression software. We conjectured that by using data compression software, we could teach the concept of information theory that says the more surprising the message, the more information it contains. Our goal is to teach students the basics of human communication through the concept of information theory, because information theory implies that we need to communicate differently according to the knowledge of the receivers of our messages. We applied our method to teach junior college students, and obtained a fair result in which 92% replied that they well understood or somehow understood the concept.

Key Words

data compression, information theory, learning

要 旨

本研究では，データ圧縮ソフトを使って，情報の概念を教えることを試みた。データ圧縮の原理を学習に応用すれば，メッセージを伝えるとき，それが相手の予測に反する内容を多く含んでいればいるほど，受け取る情報の量が多いという情報の概念が理解でき，ひいては，相手の持っている知識によってメッセージの伝え方を換えなければならないという，コミュニケーションの基本が理解できるのではないかと考えた。短期大学一年生に授業実践を行った結果，「よく理解できた」と「少し理解できた」を合わせて，92%の学生が一応の理解を示した。

キーワード

データ圧縮，情報理論，学習

1．はじめに

シャノンが提唱した情報理論は，さまざまな物理量と同じように，情報というものに測ることができる対象としての数学的な定義を与える[1]。この理論により我々は，あらゆる情報システムの設計に対する指針や，評価のための定量的尺度を得ることができるようになった[2]。情報理論に基づく情報の概念は，情報通信における効率の良いデータ転送や，コンピュータの記憶領域を節約するために重要であるが，また，人と人との効果的なコミュニケーションを考える上でも重要である。それは，情報とは伝達によって意味をもち，その伝達の手段はその情報を受け取る人に分かる言葉で語らなければ意味がないことを，情報理論が教えている[3]からである。

情報理論によると，相手に伝えるメッセージが，相手の予測に反する内容を多く含んでいればいるほど，受け取る情報の量が多いということになる。つまり，メッセージを受けたときの受け手の驚きが大きいほど，情報の量は多い[3]。例えば，相手が昨日の天気を知っているが，明日の天気予報を聞いていない場合，「昨日の天気は雨だった」という言葉より，「明日は雪が降るだろう」という言葉の方が，より多くの情報を含んでいる。

本研究では，データ圧縮ソフトを使って，情報の概念を教えることを試みた。データ圧縮は，情報理論を最も直接的に使う技術分野の一つである。その原理は，次に現れる文字や単語を予測し，予測に合っている場合には，それを短い符号で置き換え，反対に予測と掛け離れている場合には，長い符号で置き換えるというものである[4)5]。この原理を学習に応用すれば，メッセージを伝えるとき，それが相手の予測に反する内容を多く含んでいればいるほど，受け取る情報の量が多いという情報の概念が理解でき，ひいては，相手の持っている知識によってメッセージの伝え方を換えなければならないという，コミュニケーションの基本が理解できるのではないかと考えた。

2．情報理論とデータ圧縮

情報理論によると，情報は伝達することによって意味を持つ。情報の受け手は，アルファベットの一つを選ぶときのように，いくつかの選択肢を持っていて，その中のどれが当たっているか判断しようとする。ある1つの選択肢（ i ）の実現確率がPiであるとき，それが当たっている場合の情報の大きさは式（1）で与えられる[1]。このときの情報量の単位は「ビット」，すなわち2進数の何桁必要かという数である。これを噛み砕いた言い方で説明すると，「受け取ったメッセージの驚きが大きいほど，そのメッセージはより多くの情報を含んでいる」ということになる[4]。

$$log(1/Pi) \qquad (1)$$

一方，無損失のデータ圧縮ソフトを使うと，ファイルに含まれている情報を一切変更せずに，そのファイルのサイズを小さくすることができる。このことから，データ圧縮でサイズを小さくできるほど，そのファイルに含まれる情報の量が少ないと言える。そこで，驚きの量が多いほど，データ圧縮を行ってもファイルサイズが小さくならないことを示せば，驚きの量と，情報の量の関係が理解できるのではないかと考えた。

データ圧縮は，大まかに無損失のものと損失のあるものに分けられる。無損失のデータ圧縮は，ワープロソフトで作成したファイルのように，圧縮した後，それを復元したときに，元のファイルと完全に同じくならなければならない場合に使用する。これに対して，損失のある圧縮は，画像のように，復元したときに，人間の感覚で元と同じように見えたり聞こえたりすれば，厳密には同じファイルでなくても良い場合に使用する。本研究で用いたデータ圧縮ソフトは，「LZSS」と呼ばれる無損失のもので，ZivとLempelが1977年に開発した「LZ77」と呼ばれるソフトウェア[6]に，Bellが1987年に改良を加えて製作したものである[7]。これらのデータ圧縮ソフトは，辞書型データ圧縮プログラムに属し，すでに符号化が終わった部分のメッセージを辞書と

して参照しながら，未処理部分のデータを符号化するという手法を取っている。[4)]

3．授業実践

データ圧縮ソフトを用いて情報概念を教える授業を，平成10年６月に，短期大学一年生188人に対して行った。授業時間はおよそ１時間で，授業内容は以下の通りである。

(1) 四つの同じサイズの画像をデータ圧縮ソフトで圧縮して見せ，単純に見える画像ほどサイズを小さくできることを示す。
(2) それぞれ25個の平仮名が書かれている四つのテキストファイルの文字当てを行い，間違えた文字の数を記録させる。
(3) この四つのテキストファイルをデータ圧縮ソフトで圧縮してみせる。
(4) 文字当てで間違いが多かったファイルほど，圧縮しても小さくならないことをデータで示し，驚きの量と情報の量の関係を理解させる。

（１）において学生に見せたのは，図１に示す四つの画像である。図では白黒の画像に見えるが，実際はカラーの画像である。それぞれの画像は，一万画素（横100画素×縦100画素）のビットマップ画像で，圧縮前のファイル容量は，いずれも30,054バイトであった。画像１は，ピンク一色だけの単純な画像。画像２は，コンピュータの描画ソフトで描いた，一様な配色の幾何学的図形が描かれているもの。画像３は，やはり描画ソフトによるものであるが，所々にぼけたように見える部分のある画像。画像４は，人と風景をカメラで撮った画像である。

これらの画像の一つひとつを，学生にコンピュータ画面を見せながら圧縮した。圧縮に使用したLZSSソフトは，WindowsのDOSプロンプトにコマンドをタイプして使う種類のもので，使用するときには図２のような表示になる。圧縮が実行されると，DOSプロンプトのウィンドウには，圧縮前のファイルサイズ，圧縮後のファイルサイズおよび圧縮

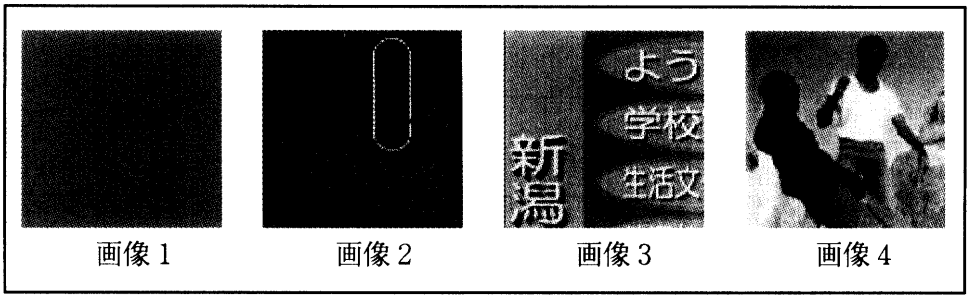

図１　学生に示した四つの画像

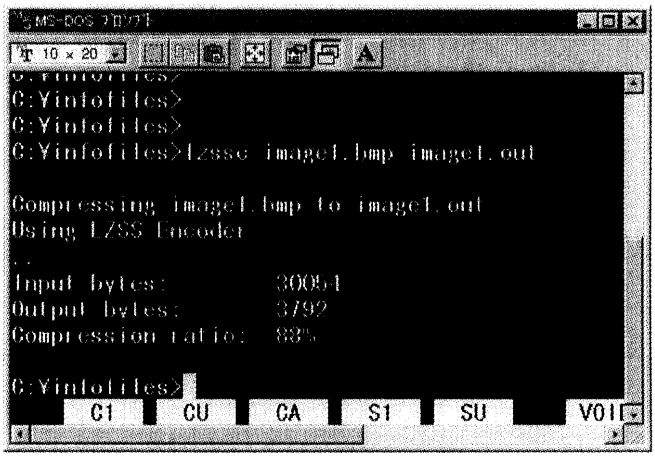

図２　ＬＺＳＳデータ圧縮ソフト実行画面

率が表示される。

圧縮前後のファイルサイズをまとめると表1のようになる。この結果を見ると，複雑に見える画像のほうが圧縮するのが難しく，また，幾何学的な模様の画像より自然な風景の画像のほうが圧縮するのが難しいということが分かる。

（2）で学生に文字当てを行わせた四つのファイルに入っている文字列を表2に示す。それぞれのファイルには，25個のひらがなが入っている。すべて全角の2バイトコードであるので，圧縮前のファイルのサイズはどれも50バイトである。学生には，すべてのファイルに，ひらがなが25個入っていることを告げ，「よ」と「ょ」のような大文字と小文字は区別するとして，文字当てを行わせた。

この文字当てを行うために，学生には予想する文字と正解の文字を書き込むための用紙を与えた。その用紙の一部を図3に示す。学生には，一つひとつの文字を予想させた後，すぐに正解を示した。それにより学生は，それまでの正解の文字列を見ながら次に来る文字を予測できる。勿論，それぞれのファイルで一番目に来る文字については，予測するための基準がないので，まったく根拠のない予測をしなければならない。

ファイル1は，「ら」だけが25個並んでいるファイルである。このようなファイルの場合，学生は最初の2，3文字は間違えるかもしれないが，「ら」がいくつか続いていることを見ると，次の文字も「ら」であるだろうことは，容易に予測ができる。LZSSを使用してこのファイルを圧縮するときも，初めの一文字はその文字のデータをそのまま符号化するしかないが，2つ目の文字からは，「この文字から始まる24文字は，一つ前の文字から始まる24文字とまったく同じ」という情報として符号化できるので，効率良く圧縮する

表1　画像ファイルの圧縮前後のサイズ（バイト）

画　像	画像1	画像2	画像3	画像4
圧縮前	30,054	30,054	30,054	30,054
圧縮後	3,792	3,898	15,439	27,778

表2　文字当てファイルの中の文字列

文字番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
ファイル1	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら	ら
ファイル2	い	い	い	い	い	た	た	た	た	た	た	た	た	た	た	ら	ら	ら	ら	ら	に	に	に	に	に
ファイル3	み	た	よ	み	た	よ	か	え	る	を	み	た	よ	み	た	よ	み	た	よ	う	み	を	み	た	よ
ファイル4	じ	ょ	う	ほ	う	の	な	か	に	あ	る	と	く	に	じ	ゅ	う	よ	う	な	じ	ょ	う	ほ	う

File 1

番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
予想																									
正解																									

間違った文字の数 _____

図3　文字当て回答用紙の一部

ことができる。

ファイル２では，やはり同じ文字が続くが，全部で４種類の文字が使われている。同じ文字が続いている部分では，次に来る文字を予測するのは容易であるが，文字が変わるところでは，その文字を言い当てるのは非常に難しい。LZSSによる圧縮においても，文字が変わるところの最初の文字は辞書に無いので，辞書を参照する符号化ができない。そのため，ファイル２はファイル１に比べると圧縮率が低い。

ファイル３では，「みたよ」というパターンが何度か繰り返される。この繰り返しがあるために，学生はある程度この「みたよ」の文字列を言い当てることができる。同じように，LZSSも，二度目以降の「みたよ」の文字列は，例えば「10文字前に現れた３文字の文字列と同じ」という情報で符号化することができる。

ファイル４では，「じょうほう」というパターンが２度現れるだけなので，LZSSでは，この繰り返しだけが，圧縮に寄与できる部分であるが，人間の場合には，文脈から予測できる文字もいくつかある。

これら四つのファイルを圧縮するときの，圧縮前後のサイズを表３に示す。

これら，四つのファイルの文字当てをした後，学生はそれぞれのファイルで正解できなかった文字の数を用紙に書き込む。その後で，LZSSを使用してこれらのファイルを圧縮し，学生に圧縮後のファイルサイズを示した。

４．結　果

学生188人に文字ファイルの文字当てを行わせ，指示どおりの回答を行わなかった24人の調査票を除外した164の調査票をもとにして集計を行った。四つのファイルの文字当てにおける間違いの数の平均と標準偏差を，表４に示す。

また，それぞれのファイルに対する間違った文字の数の平均と，圧縮後のファイルサイズをグラフにしたものを図４に示す。この図を見ると，圧縮後のファイルサイズと，間違った文字の数の平均は，ともに式（２）を満たしている。しかし，ファイル３とファイル４とで間違った文字数の平均値が近く，またその標準偏差が大きいため，間違いの数が純粋に式（２）の順になったのは，164サンプル中71サンプルだけであった。

ファイル１ ＜ ファイル２
＜ ファイル３ ＜ ファイル４　（２）

表３　文字ファイルの圧縮前後のサイズ（バイト）

文字ファイル	1	2	3	4
圧縮前	50	50	50	50
圧縮後	11	18	28	45

表４　文字当ての結果

文字ファイル	1	2	3	4
間違った文字数の平均	3.8	5.7	11.3	12.9
標準偏差	1.6	1.5	2.3	2.3

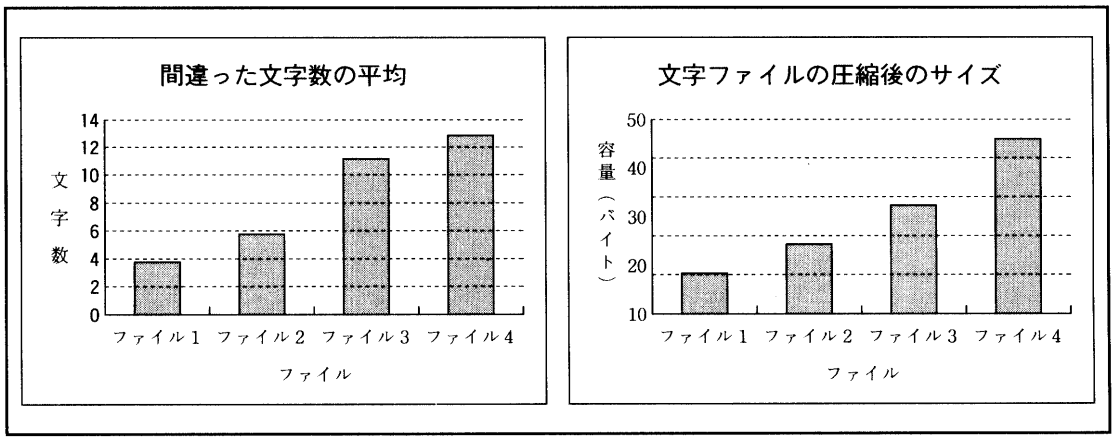

図４　間違った文字数と圧縮後のサイズ

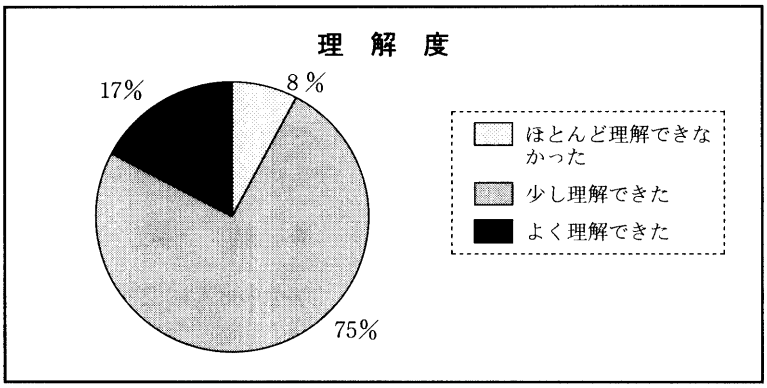

図5　学生の理解度

それぞれのファイルの間違った文字の数と，圧縮した後のファイルサイズを比較させた後，「メッセージの中の驚き（間違い）の量が多いほど情報の量も多い」という説明をして，そのことへの理解度を調査した。理解度の調査は，次の3つの選択肢から1つを選ばせることで行い，その結果は図5のようになった。

(1)　ほとんど理解できなかった
(2)　すこし理解できた
(3)　よく理解できた

結果として，「よく理解できた」は17%に留まり，おおかた（75%）の学生が少し理解できたと答えたことは，やや残念な結果であった。しかし，一般には理解しづらい概念の学習であるので，「よく理解できた」と「少し理解できた」を合わせて，92%の学生が一応の理解を示したことは成果であった。

5．まとめ

テキストファイルの文字当てを行い，文字を当てるのが難しいファイルほど，データ圧縮ソフトで圧縮したときのファイルのサイズが大きいことを示して，「メッセージの中の驚き（間違い）の量が多いほど情報の量も多い」という情報の概念を教える試みを行った。短期大学一年生に授業実践を行った結果，情報概念の理解について，「よく理解できた」は17%に留まったが，「よく理解できた」と「少し理解できた」を合わせて，92%の学生が一応の理解を示した。

この情報概念の理解を一歩進めて，人にメッセージを伝えるとき，相手の持っている知識を確認し，相手がどのような知識を持っているかによってメッセージの伝え方（話し方）を換えることが重要であるという，コミュニケーションの基本を理解させることが今後の課題である。

参考文献

1）Shannon, C.E. and Weaver, W., The mathematical theory of communication, University of Illinois Press, Urbana, IL, 1949
2）電子情報通信学会編「電子情報通信ハンドブック」オーム社，1998，434－445頁
3）太田正光，大芝猛，田坂修二「情報科学概論」講談社，1996，41－42頁
4）Bell, T.C., Cleary, J.G., and Witten, I.H., Text Compression, Prentice Hall, 1990, pp28－33 and pp206－243
5）Sayood, K., Introduction to Data Compression, Morgan Kaufman Publishers, 1995
6）Ziv, J. and Lempel, A., A Universal Algorithm for a Sequential Data Compression, IEEE Trans. Information Theory, IT－23 (3), 1977, pp337－343
7）Bell, T.C., A Unifying Theory and Improvements for Existing Approaches to Text Compression, Ph.D. Thesis, Department of Computer Science, University of Canterbury, New Zealand, 1987