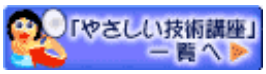
ご利用にあたっての注意
「ロスレス圧縮技術」は2006年～2008年当時の情報です。予告なしに更新、 あるいは掲載を終了することがあります。 あらかじめご了承ください。

ロスレス圧縮技術

情報を保ったままデータ量を元の1/5～1/2に減らすことができ、そして、元のデータに完全に復元する技術です。 どんな種類のデータにも
使えます。(ロスレス圧縮＝Lossless圧縮＝損失がない圧縮)

目次

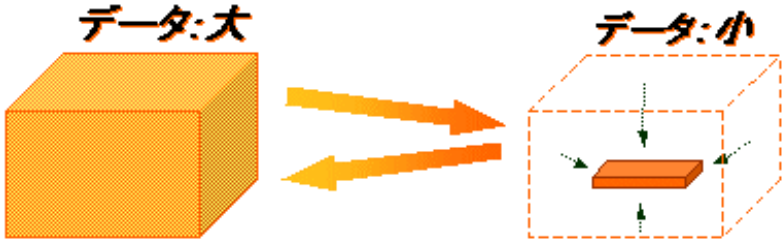


圧縮ってなんだろう

圧縮って何だろう

・圧縮：元のデータ量より、小さくすること
・復元：圧縮されたデータを元に戻すこと

元データを決まった方法で圧縮して、同じ方法で復元します。



圧縮技術には大きくわけて2種類あります

・Lossless（ロスレス）圧縮技術
「文書・プログラム」に使います。圧縮したデータを完全に元に戻せるからです。1字違うだけで、文書は意味が異な るので、圧縮率よりも完全な復元を最優先しています。
（圧縮率は元のデータ量の50パーセントから25パーセントです）

・Lossy（ロッシー）圧縮技術

「画像や音声」に使います。元のデータ量がとても多いので、圧縮率を最優先に考えた場合です。例えば、使う側に とって気づかない程度の復元（1画素の色階調が少しずれているだけでは、人は気づきません）で良い場合です。
（圧縮率は元のデータ量の0.01パーセントから0.001パーセントです）

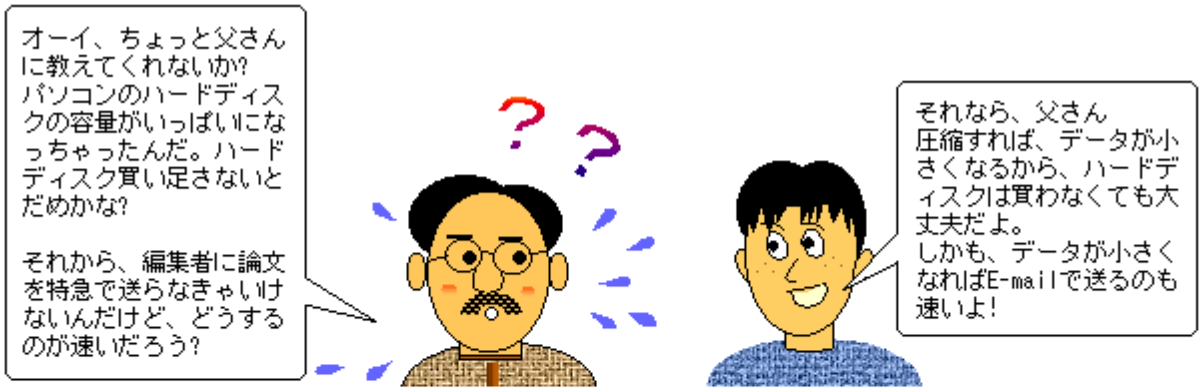
（注）ロッシー圧縮については「画像圧縮技術」をご覧下さい。

復元性と圧縮率により、LossLess圧縮かLossy圧縮かを使い分けます。
この講座では、ロスレス圧縮技術についてご紹介します。

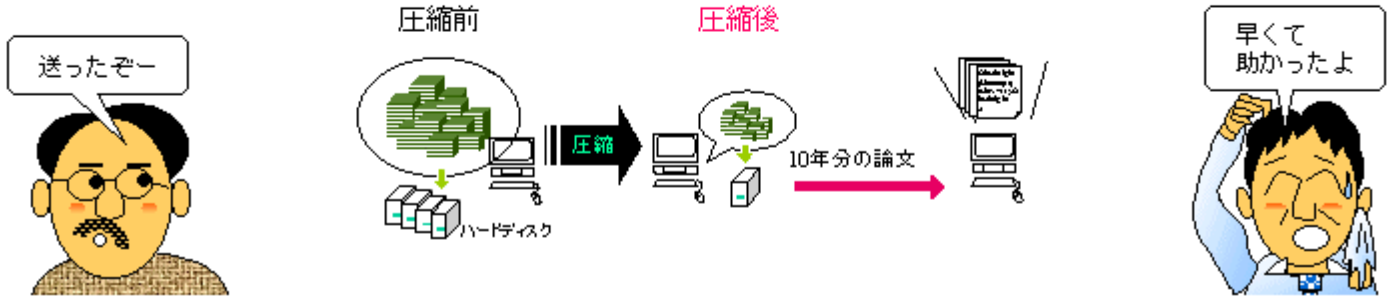
「やさしい技術講座」一覧へ
この講座の最初のページへ

ロスレス圧縮の特徴

困っている父と圧縮を良く知っている息子の会話



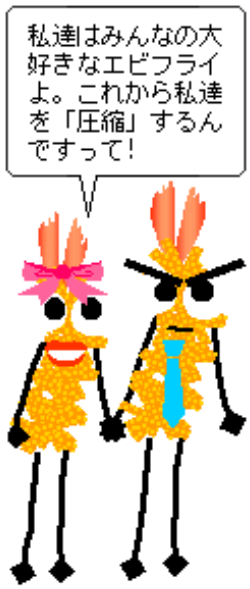
データ量が多い文書、プログラムなどを圧縮して保存しておけば、まだまだ記録できます。E-mailで送付する時にも、データ量が少なくなるので、早く送付できます。



「やさしい技術講座」一覧へ
この講座の最初のページへ

原理

データは、データ情報冗長性（詳しくは「情報と冗長性」へ）という構成で成り立っています。ロスレス圧縮は、 このうちの「冗長性」という部分を、限りなく小さく圧縮し、完全に復元するという原理です。では、圧縮する データをエビフライに例えて見てみましょう。



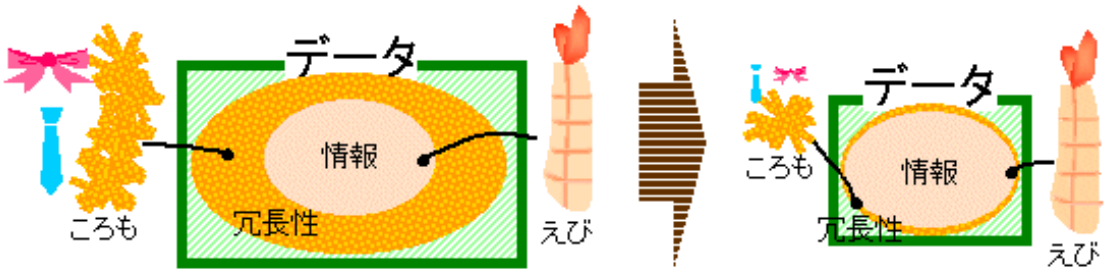
圧縮へ突入。

おや。
エビとコロモにわかれてしまいました。

わかれるってどういうことだろう。
わかれるとどうなるんだろう。

エビフライは主役のえびとそれを引き立てるころもにわけることができます。データも肝心な情報とそれを助ける冗長性にわけることができます。

- 情報:えび
- 冗長性:ころも、リボン・ネクタイ



エビフライは、主役であるえびが欠かせないのはもちろん、えびを引き立てるためにころもも必要です。
データも、肝心な情報は欠かせませんが、やはり冗長性(情報を助けるもの)も必要なのです。

つまりロスレス圧縮は、この肝心な「情報=えび」はそのままに、情報を助けている「冗長性=ころも、リボン・ネクタイ」を、ある計算方法で少なくする(=ころもを薄くする)という原理なのです。

そして、圧縮したデータを復元すると…。

立派なエビフライに戻りました。

「やさしい技術講座」一覧へ

この講座の最初のページへ

ホーム | サイトマップ

富士通ホーム | 富士通のアクセシビリティ

情報と冗長性

圧縮には、「情報」や「冗長性」という言葉がでてきます。特に「冗長性」というのは聞きなれませんが、それは一体どういうものでしょうか。 【例文：青い海白い空】で見てみましょう。

「い」の繰り返し

青い海　　広い空

形容詞+名詞というパターン

冗長性（規則性）

- 繰り返し：「い」という文字の繰り返し
- 文法的な規則性（よくあるパターン）：名詞（「海」や「空」）の前には形容詞（「青い」や「広い」）が配置されるなど

情報（必要最低限の肝心な内容・エッセンス）

本質的なデータ：冗長性以外の「青」「広」「海」「空」

このように、この文には二つの冗長性があります。

（注）冗長…くどくどしくて長いこと。（国語辞典より）

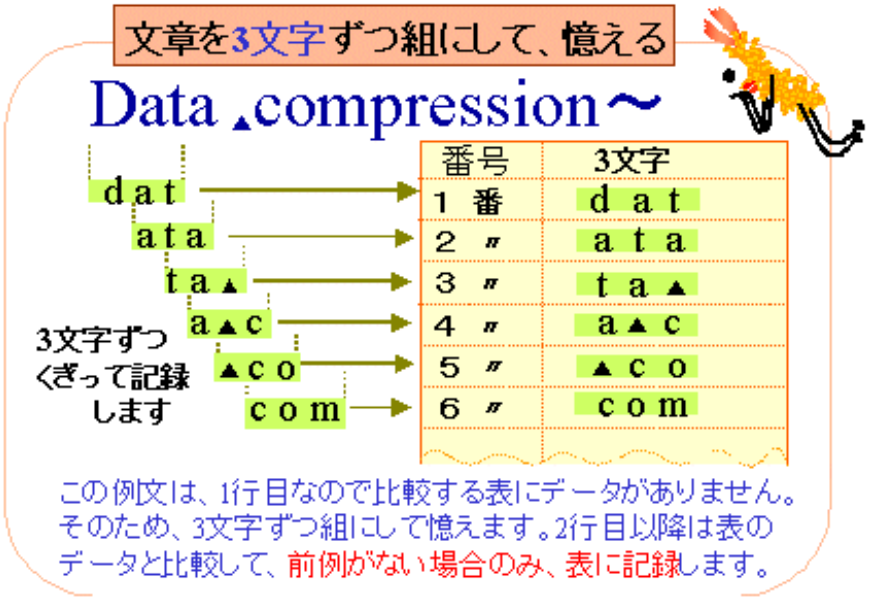
「やさしい技術講座」一覧へ
この講座の最初のページへ

SLCAの圧縮方法

データの圧縮技術は色々な方式が開発されています。当社では、お客様が安心してお使いいただける、 高性能な圧縮技術を開発しました。例文を使いながら、その技術を説明しましょう。 （少し難しいので、右側にイメージ図を付けます）

・SLCAの圧縮方法

↓



(注)番号は記録した3文字が、文章中の何番目に表示されていたかわかるようにつけています。

・イメージ図

エビフライ達が人材派遣会社に登録しにやってきました。

↓

構成を確認して、記録します。

↓

前の文章で同じ3文字があったら、同じ単語か比較する

例文1. 2行目にcompressionがきた時、3文字にすると:com。
憶えたcomは、表をみると前の文章中の6番目に出てきた
ことがわかります。その前の文章中のcomの後ご続く文字を比
較します

1行目の Data compression～
2行目の compression ＝com+8文字が一致
8文字

2行目にでてきた
compressionは
"Com8"
と憶えます
つまり4文字でOK

例文2. 2行目にcommunicateがきた時、3文字にすると:com。
憶えたcomはcompressionなので、comの後ご続く文字
で同じものはありません。

1行目の Data compression～　Pとmが不一致
2行目の communicate＝ com+0文字(一致している文字が無い)

communicateは
そのまま憶えます
つまり11文字必要

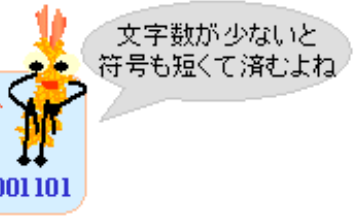
(注)実際は、"municate"以下も同様に、3文字ずつにしてすでに記録してある表に一致しているものがあるか検査します。

↓

前の人と同じ物がないかを確認。

↓

1文字ずつ符号化



例文1 4文字分の符号でOK
010010111

例文2 11文字分の符号が必要
01001011100100100 1101

たくさん使われている文字ほど、短いビット数の符号で表します。
圧縮率が高くなるんだね。

↓

いつも同じように書かなくても、前に出てきた言葉なら、マークを書いておけばラクだね。

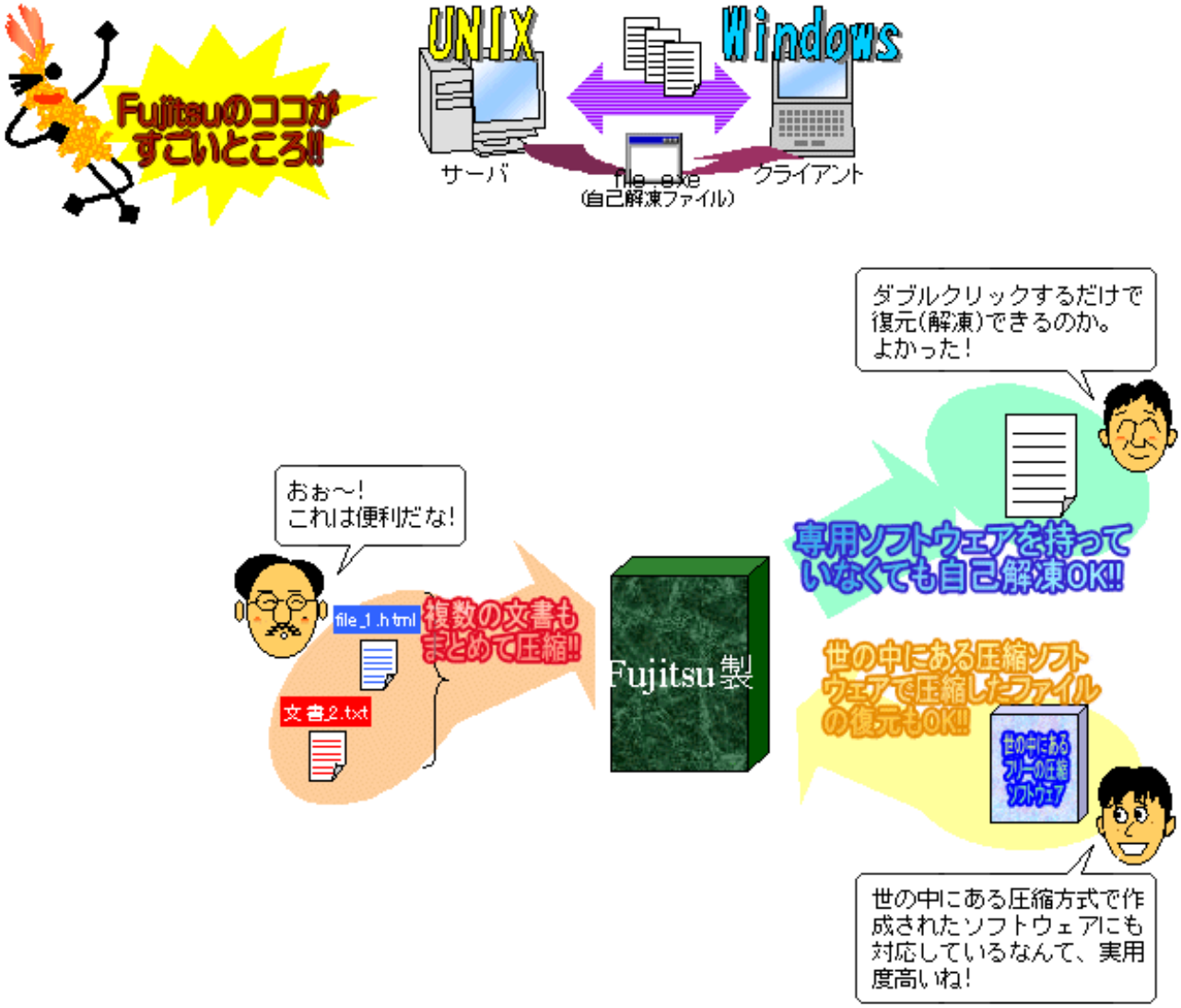
「やさしい技術講座」一覧へ

この講座の最初のページへ

ホーム | サイトマップ
富士通ホーム | 富士通のアクセシビリティ

SLCAの特徴

当社のソフトウェアは異なるOS間のやり取りも可能です。

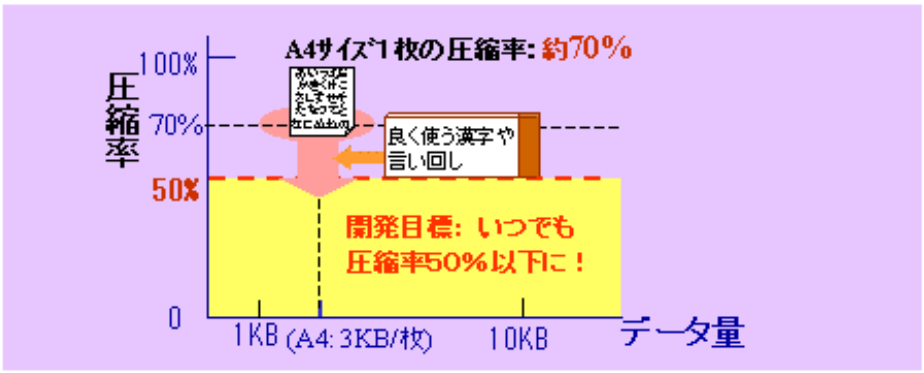


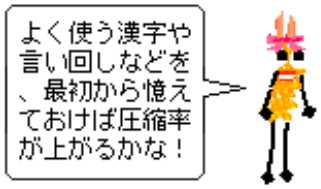
「やさしい技術講座」一覧へ
この講座の最初のページへ

課題～高圧縮率の実現～

データ量が少ない場合でも、高い圧縮率を実現したい

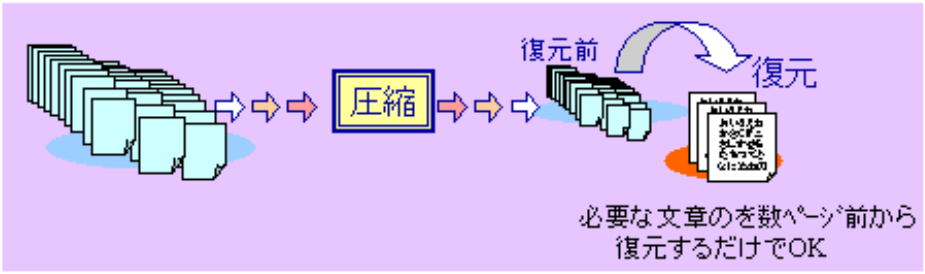
データ量が多くなると圧縮率が高くなる原理のため、データ量が少ない場合には、圧縮率50%を達成 できません。データ量の大小にかかわらず、圧縮率を50%以下を達成することが課題です。

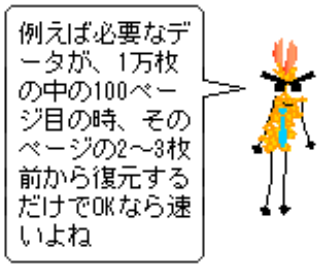




大量に圧縮したものから必要な部分だけを復元したい

大量のデータをひとまとめにして圧縮しているため、必要な文章が途中にある場合でも、復元時は 最初から作業を行うので時間がかかります。そこで、必要な文章だけを復元をできることが課題です。





「やさしい技術講座」一覧へ
この講座の最初のページへ

小話～英語・中国語・日本語について～

右の表は、全く同じ内容の社長の挨拶を英語・日本語・中国語に表し、それぞれの圧縮率を比較してみました。すると、圧縮率に差がでています。それは英語に比べて中国語・日本語は、圧縮する部分が少ないためです（英単語の後に必ず空白が入ります。これは元のデータ量は多くなりますが、冗長性の一つなので圧縮できる部分になります）

言語	元のデータ	圧縮後	圧縮率
英語	5.0KB	2.3KB	46%
日本語	4.2KB	2.3KB	56%
中国語	3.0KB	2.1KB	71%

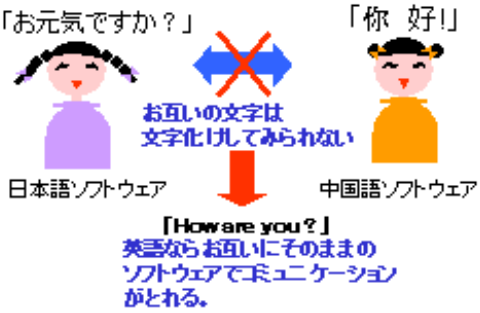
この圧縮率は、元データ量を100%とした場合、どれくらい小さくなるかを表しています。
（使用した元データ：当社社長 年頭の挨拶）

データ圧縮は、言語コードに関係なく圧縮することができます。圧縮率はその言語の冗長性に左右されます。

データ圧縮は圧縮する元のデータを問いませんが、ブラウザは言語によって専用の変換方式をインストールしなければなりません。少しブラウザの話をしましょう。

中国語のブラウザのお話

中国語は、中国語文字コードを用いているので、日本語文字コードのブラウザを使って原文のまま読むことができません。中国語文字コードのインストールが必要です。
実は、本場中国の人達は、中国語でちょっと大変な思いをしています。それは、沢山の民族が集まっている中国では、言葉にも色々な種類があります。公用語としては、北京語ですが、香港や南のエリアは、広東語を使っています。その他に上海語や四川語等沢山の地方語があります。



ソフトウェアも中国語版と言っても種類があり、更に、北京語だけでも大陸版と台湾版があります。例えば、「謝」と入力したい場合、日本語と同じ様にローマ字入力するとしても大陸版では「xie」と入力しますが、台湾版では「shung」と入力します。
しかも台湾版では、日本で言う旧漢字の様な古い文字しか変換の文字が出ませんが、大陸版になると文字を省略して使っている新漢字も辞書に入っています。
広東語に至っては、北京語と全く発音が違うので、同じ中国人同士でも全く言葉が分かりません。例えば、香港では広東語が公用語なので上海の人が行くと外国に行ったのと同じ様に言葉が通じないのです。同じ中国人同士で英語で会話する事になります。不思議な感じです。

日本語環境のパソコンのブラウザで文字化けを解消するためには、例えばマイクロソフト社からフリーで、globalIMEなどを提供しています。この辞書を入れると読む事はできますが、こちらから書く事は出来ないので、やりとりをしようと思うと中国語ソフトウェアを入れないと駄目なようです。

（注）ちょっと宣伝ですが、富士通研究所は、富士通研究開発中心有限公司（北京）において、中国語ソフトウェアの研究を積極的に進めています。

「やさしい技術講座」一覧へ
この講座の最初のページへ