氏　　名	奥村晴彦
学位（専攻分野）	博士(学術)
学位記番号	総研大乙第62号
学位授与の日付	平成１１年３月２４日
学位授与の要件	学位規則第４条第２項該当
学位論文題目	高速・可逆な実時間データ圧縮アルゴリズムの開発研究
論文審査委員	主　査　教　授　本島 修
	教　授　須藤 滋
	助教授　山口 作太郎
	教　授　渡邉　國彦（核融合科学研究所）
	教　授　寺町　康昌（職業能力開発大学校）

論 文 内 容 の 要 旨

核融合科学研究所の制御データ処理装置は、著者達が平成５年から共同研究として進めてきた「ワークステーションを用いたデータ収集・解析・制御システムの研究」の成果として開発されたものである。これは、大規模システムに対応した構成を取り、容易にCHの拡大を計る事が出来るシステムであり、アナログデータを増幅した後、アナログ・デジタル変換器(ADC) でサンプリング及びデジタイズを行い、多チャンネルの時系列データとして計算機に取り込む。これをバイナリーファイルとして保存する、データベース管理ソフトに登録すると同時にネットワークに実時間配送し、WWWブラウザーによってユーザーがデータ取り込むことが出来るシステムであり、今後の大型システムでのデータ処理システムの一つの大きな方向を示す形となった。

さて、このようなシステムを考える場合、データ量が膨大となり、データの保存・転送のコストが無視できない事が予測される。そのために、圧縮によってデータ量が数分の１になれば、データ保存用のハードディスクなどの費用は数分の１になると同時にネットワークの輻輳（混雑）も数分の１に留めることが出来る。この理由によりデータ圧縮の研究を開始した。

データ圧縮の技術は、情報量の損失のない可逆圧縮と、情報量の損失のある非可逆圧縮の２種類があり、後者は近年マルチメディアへの応用のために盛んに研究が行われているが、計測データに利用することは出来ない。更に、処理形態としては、実時間処理とバッチ処理があり、後者は従来から研究が広く行われてきたが、当然ながら実時間でネットワークを介して監視するデータは高速の実時間圧縮が必要になる。以上の要請のために、今回新たに高速・実時間・可逆なデータ圧縮アルゴリズムを開発し、コーディングに成功し、核融合科学研究所のデータに適用した結果を以下に述べる。

データ圧縮の研究は1948年のShannonの情報理論に端を発し、その後、彼の学生であったHuffmanによって符号化による圧縮方法が考案された。現在のすべての符号化による圧縮はHuffmanにその基礎を置く。その後、1976年にRissanen他によって算術符号化が開発されるが、理解することに多大な努力を要することもあり、それほど一般化されなかった。一方、その後パーソナルコンピュータの発達もあり、Zip-LempelによってLX77, LZ78と言われる圧縮ソフトが1977年及び78年に開発され、一般に利用されるようになった。その後、WelchによってLZ78の改良が行われ、LZWと言われるコードが作られた。一方、著者等は、1988年にLZ77をベースに圧縮ソフトの開発を始め、その年に著者はLZARIを発表し、吉崎はLZHUF, LZarcを発表し、広く日本でも利用されるようになった。そして、1989年に著者は現在広く利用されているLHAと言われるソフトのアルゴリズムを発表し、1990年にC言語で書かれた圧縮ツールを発表した。そして、吉崎がLHAとして完成し、広く利用されるようになった。これとほぼ同じアルゴリズムのソフトとしては、フランスのJean-loup Gaillyによるgzipがある。

さて、圧縮アルゴリズムとして上記に書いた要請があるため、以上の今までの圧

縮アルゴリズムを検討し、２つの方法を採用することにした。一つはHuffmanの符号化でり、もう一つはデータの時間的変化を直前の数点のデータから予測する方法を組み合わせた方法である。従って、この方法は予測誤差符号化と言うことが出来る。具体的には、直前の数点のデータから次のデータを予測し、それからの誤差の分布を正規分布及びLaplace分布を用いて、予測誤差の分布グラフを作る。実際のデータとの比較を行うと、誤差分布は正規分布とLaplace分布の間位になっている。これは、時折、急激に変化するデータがあり、そのため予測誤差の２乗和の平均分散を推定すると、分布の中央部の度数分布から推定した分散より大きめの値を得るからである。そして、そのようにして得た分布データを長さ制限のある符号に変換する。ここでは８ビットに制限し、符号化を行った。このような、長さ制限のある符号化を行ったのはは最初にLarmore and Hirschbergであり、比較的高速でメモリーを余り使わない効率の良い圧縮方法として知られている。しかし、そのような方法でも大変手間のかかる方法であり、実際に利用される符号は一部に限られるので、より単純で高速な方法を考案した。それは、場合分けを８ビットに制限した時には、406通りしかないので、予めその表を作っておき符号化手順を高速に行うことである。このような方法をとっても１シンボル当たり0.047ビットの損しかなく、これは最悪の場合であるため、実際にはほとんど影響は出ないからである。

以上のアルゴリズムをコーディングし、実際の核融合科学研究所の大型ヘリカル装置のデータの圧縮を行った。実際のデータは熱電対で測定した温度、歪ゲージで測定した歪、マイクロ波の出力、磁場、コイル電流及び電圧、プラズマからの輻射などでからなるデータ郡である。急激に変化するデータはやはり圧縮比はそれほど大きくなく、16ビットの生データに対して、8－10ビット程度になるが、変化の少ないデータでは2ビット以下に圧縮された。そして、それら全体を通じて、元データ（複数の性質の違ったデータをすべて含む）が8.74MBに対して、通常バッチファイルの圧縮に利用されているZip,LHAが4.8MB程度に圧縮し、今回開発したソフト（NIFSqと言う）は、2.00MBに圧縮することができた。従って、従来のソフトより圧縮比で２倍程度性能が良い。更に、圧縮時間であるが、CPUのクロックが400MHzのPentium-II, のパーソナルコンピュータ（オペレーティングシステムはLinux）ではZip, LHAに比べて3分の1から4分の1以下の時間で行うことが出来た。従って、従来のソフトに比べてかなり高速なったことが分かった。そして、適応型のソフトのために、現状のデータ処理システムではほとんど無制限なCHまで圧縮可能のソフトとなった。

以上より、当初の要請は満足されたと考えるべきである。今後の課題はアルゴリズムの最適化及びユーザーインターフェースを改良した上でMacintoshなども含めた計算機環境での利用を可能にし、C++やJavaに書き直すことなどである。

論文の審査結果の要旨

本論文は数々の研究活動に欠かすことが出来ない大量の実験データのデータ圧縮法の独創的な開発研究を主題としている。具体的には、共同利用施設である核融合科学研究所の大型ヘリカル装置（LHD）の制御データ処理装置を対象としており、本研究の成果により、時系列データについてはデータサイズは無制限であり、現状のパソコンで180GB/day程度の処理を行うことが出来る。今後の各種実験で使用することのできる一般性の高いデータ圧縮法を開発し、実地にLHDでの実験データに応用した結果、その有効性を実証するに至っている。

奥村氏は昭和63年以来、我国で開発され世界で広く利用されているデータ圧縮ソフトウェア「LHA」のアルゴリズム開発を行っており、平成5年以降は核融合科学研究所で共同研究「ワークステーションを用いたデータ収集・解析・制御システムの研究」の研究代表者を努め、その成果として本データ圧縮技術を開発するなど、当該分野の第一人者の一人として実力を広く認められている研究者である。

LHDプラズマ実験のごとく大型装置を用いる研究では計測データは今後ますます膨大な量になり，計測データ圧縮技術の開発の必要性は近年とみに高まっている。データ量を数分の1にできるならば、データの取り扱い量を飛躍的に向上させ、研究者の解析能力を更に高めることが出来る。さらには、高速可逆性のあるデータ圧縮によりデータ保存およびネットワーク配信のためのコストも数分の1にできるなど、貴重な技術であると言える。

データ圧縮の技術は，情報量の損失のないロスレス（lossless）な圧縮と，損失のあるロッシー（lossy）な圧縮とに大別できる。特にロッシーな圧縮は近年マルチメディアへの応用のため盛んに研究が行われているが、計測データに用いることができるのはロスレスな圧縮であり、そのために特別なアルゴリズムの開発が求められてきた。

その処理形態であるが、一般には実時間圧縮とバッチ圧縮とに分けることができる。連続的に収集するデータの場合には、実時間での圧縮が必要となる。これに対して，短時間に多量のデータを収集する場合は，ファイルに落とした後にバッチ処理的に圧縮することが可能であるが、後者の場合でも，ディスクへの書き込みの手前で実時間で圧縮することにより，ディスクの書き込み速度を超えた速度でのデータ収集が可能になる。

汎用のロスレス圧縮ツールとしては，バッチ方式の圧縮に限れば、過去に奥村氏がアルゴリズムの基本部分を開発したLHAを始めとして，多数のものがある。しかしこれらのツールは計測データを圧縮するために開発されたものではなく、圧縮速度やリアルタイム特性上問題があった。

本論文の主要なテーマである新たな圧縮法はNIFSqと名付けられ、データの予測を行い、予測値からの誤差のバラツキを正規分布又はラプラス分布を仮定した論理的実時間アルゴリズムを用いている。これによって、典型的な計測データを約1/4に圧縮でき、しかもスループットは単一CPUで約2Mサンプル/秒であり、LHAを始め比較したデータ圧縮ツールのどれをも上回る性能が得られている。しかもメモリ

使用量は、各チャンネルあたり予測に必要な数点前のデータの処理能力をもたせるだけで良いので、チャンネル当たり約100バイトで済む。しかも、同時に圧縮できるチャンネル数は事実上無制限という構成である。この研究成果は研究論文としてプラズマ核融合学会誌に1997年末に発表されている。

以上述べたように本論文はロスレスデータ圧縮についての有用なプログラムの提供とその包括的な検討が含まれており，この分野の最先端を切り開く研究成果として評価される。よって，本審査委員会は，本論文が博士学位論文として十分な水準にあり，本専攻にふさわしい内容を持つものであるとの結論に達した。

審査委員全員参加によって、口頭試問を実施し，論文内容に関する知識等について試験を行った。その結果、情報理論，符号理論，コンピュータネットワーク技術，プログラミング技術等、本研究に必要な多岐にわたる知識と技術について精通していると認められた。奥村氏を筆頭著者とする論文は英文3編を含め合計7編および著書6件が出版されており、語学力については，論文要旨の英語版，添付された英文の参考論文等の審査により，十分な能力を有すると認定した。