D-033

番組ファイルの差分検出手法の検討

Study of Method for Detecting Differences Between Program Files

南 浩樹† 金子 豊† 竹内 真也† 藤沢 寛† 和泉 吉則†
Hiroki Minami Yutaka Kaneko Shinya Takeuchi Hiroshi Fujisawa Yoshinori Izumi

１．研究の背景

放送局では、作業フローの効率化に向け、現行のＶＴＲを用いたシステムから、映像・音声ファイルを用いた番組制作・送出システムへの移行が進められている。さらなる効率化を目指し、我々はさらにネットワークを利用したシステムの研究を行っている[1]。図１にネットワークを利用した番組制作・送出システムを示す。カメラで収録された映像や音声はファイルの形で編集サーバに取り込まれ、編集機により１つの番組ファイルに編集される。完成した番組ファイルは、ネットワークを経由して登録サーバに転送されて登録され、送出に備える。

登録された番組は、誤りの修正や番組をより良くする目的で一部分を手直しすることがある。このとき、手直ししたファイルを転送して差替える必要がある。特に、放送直前では差替時間の短縮化は重要な課題である。しかし、放送局で扱う番組ファイルは編集による画質劣化を考慮して高画質にしているため、サイズが非常に大きく、差替に時間がかかる。例えば100Mbpsで圧縮されたHDTVの１時間番組は約45GBであり、このファイルを1Gbpsのネットワークで転送し直すと、計算上最速でも６分以上かかる。通常、手直しにより修正された部分は番組全体より十分小さいので、修正箇所だけを転送して差替えることができれば差替時間を短縮できる。このためには、登録サーバ上にある手直し前の番組ファイルと編集サーバ上にある手直し後の番組ファイルを比較して修正箇所を知る必要がある。

本稿では、ネットワーク接続された別々のサーバにあるサイズの大きいファイルを比較して相違箇所を短時間で特定する手法を検討し、性能を評価したので報告する。

２．差分特定手法の提案

差分（相違箇所）の特定とデータ転送時間の合計が、番組ファイル全体を転送するよりも短くなければ、差分のみを転送する意味がなくなる。従って、短時間で差分を特定することが要求される。

２．１．比較データ量の削減

別々のサーバにあるファイルを比較するために、どちらかのファイルを比較対象のファイルがあるサーバに転送すると、これだけでファイル全体を転送して差替えるのと同等の時間がかかる。従って、ファイルの実データを用いずに比較する必要がある。

番組ファイルのフォーマットはMXF[2]を想定している。MXFは、記録メディアへの入出力を意識して、すべてのフレームの先頭を記録メディアの管理ブロックの先頭に合わせることができる。すなわち、すべてのフレームサイズをブロックサイズの整数倍にできるため、ファイルの修正により修正箇所のサイズが変わる場合でも、増減するサイズは常にブロックサイズの整数倍となる。このため、ブロック単位よりも細かい単位で比較する必要はない。

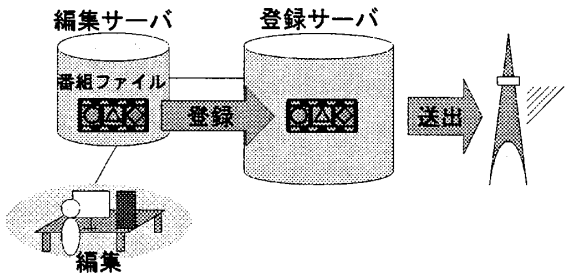

図１：ネットワークを利用した制作・送出システム

そこで、ファイル比較のデータ量削減のため、番組ファイル保存時に管理ブロックサイズごとにハッシュ値を計算し、ハッシュ値列も保存することとした。ファイルの実データの代わりにハッシュ値を用いることにより、比較時間や比較に必要なデータ転送時間を十分に短くできる。

２．２．大容量データの比較手法

ハッシュ値列を用いることで、実データより十分に小さくできるが、それでもサイズは非常に大きい。手直しによる修正箇所のサイズは一般的には小さいが、差分のサイズが大きい場合にも対応する必要がある。既存の文書比較アルゴリズムの多くは、変更履歴の検出や差分出力量を少なくすることを目的としているため、正確な差分の出力に注力される。そのため、差分のサイズが小さい場合は高速だが、差分のサイズが大きくなると時間がかかる。一方、このシステムでは、短い共通部分は見逃しても短時間で差分出力を得ることが求められる。例えば、大部分が異なるファイル同士の比較の場合は、わずかな共通部分を検出するために多大な時間を要するよりは、全く異なるファイルであるという結果を短時間で得られることが求められる。

そこで、サイズの大きいデータの高速比較手法を提案する。共通部分でない部分が差分なので、共通部分を高速検出できれば良い。図２に示すような１箇所だけ連続した共通部分をもつファイルＡとファイルＢを用いて説明する。ファイルＡとファイルＢの全ブロック数をそれぞれ N_A, N_B とし、共通部分のブロック数を L とする。以下の説明でブロックの比較とは、ブロックのデータから計算したハッシュ値で比較することを意味する。

全ブロックを総当りで比較すると比較回数は非常に多くなるため、比較対象のブロックを限定し、その他のブロッ

† ＮＨＫ放送技術研究所（システム）
Broadcasting Systems, Science & Technical Research Laboratories, NHK

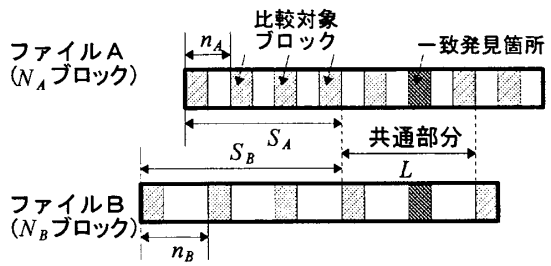

図２：ファイルの共通部分の検出方法

クは比較対象としない。ここでは、$n_An_B \leq L$ かつ最大公約数が１となる自然数 n_A と n_B を選び、先頭ブロックから、ファイルAは n_A ブロック間隔、ファイルBは n_B ブロック間隔で比較対象ブロックを設定する。$n_An_B \leq L$ を満たす共通部分の中には、同じ値の比較対象ブロックが必ず存在するため、$L_{min}=n_An_B$ とすると、L_{min} は検出可能な共通部分の最少ブロック数である。値が一致した比較対象ブロックを起点に前後を１ブロックずつスキャンすることで、共通部分の範囲が分かる。この方法により、総当りよりも n_An_B 倍のオーダーで速く比較できる。さらに比較回数を減らすために、ファイルAのすべての比較対象ブロックをB-Treeのアルゴリズムで Tree を構成し、見かけ上ハッシュ値の小さい順にソートする。ファイルAの比較ブロック１個のソートおよびファイルBの比較ブロック１個の比較に要するハッシュ値の比較回数の期待値を c（定数）に近似すると、ハッシュ値の比較回数は次式となる。

$$\left(\frac{N_A}{n_A}\right)c+\left(\frac{N_B}{n_B}\right)c \qquad \left(c \text{ は } o\left(\log_2\frac{N_A}{n_A}\right)\right)$$

$n_An_B=L_{min}$（定数）の条件下では、この式を最小にする n_A と n_B の値は、次のようになる。

$$n_A=\sqrt{\frac{N_A}{N_B}L_{min}},\quad n_B=\sqrt{\frac{N_B}{N_A}L_{min}}$$

従って、比較回数を少なくするためには、n_A と n_B は、最大公約数が１となる条件下で上式に近い値を選択する。

３．差分出力時間の評価

提案した手法を用いて、実験システム上のサーバで差分出力時間を測定した。使用したサーバは、CPU は Pentium D (3.0GHz)、メインメモリは 1GB である。

ブロックサイズは、実験システムで使用しているファイルシステムの管理単位である 4kB とした。ハッシュ値のサイズは 128bit とした。このとき、ハッシュ値列のサイズはファイルの実データの 1/256 であり、１時間番組の場合は約 180MB（約 1100 万ブロック）である。１時間番組ファイル同士の比較であれば、メインメモリ上だけで計算できる情報量である。

３．１．提案する手法での差分出力時間

差分出力時間の評価のため、図３に示すブロック数 L の共通部分を持つブロック数 N のファイルAとファイルBを作成した（$N_A=N_B=N$）。ファイルAの共通部分の中心とファイルAの中心を一致させ、ファイルBの共通部分の中心はファイルBの中心より ΔN ブロック後方に設定した。N は約１時間の番組ファイルを想定し 10^7 とした。L

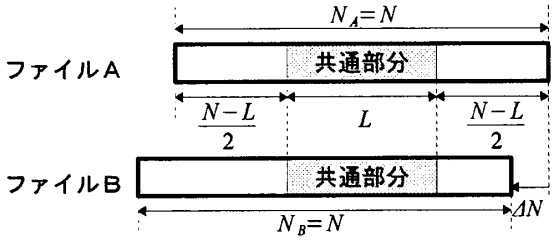

図３：比較したファイルの共通部分の位置関係

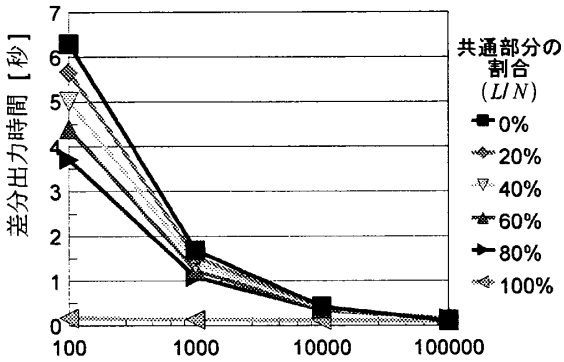

図４：検出可能な共通部分長と差分出力時間の関係

は N の 0%から 100%まで 20%間隔とした。L_{min} は 100, 1000, 10000, 100000 とした。例えば、L_{min} =10000 (N /1000)のときは約３秒以上、L_{min} =100 のときは約１フレーム（1/30 秒）以上の共通部分が必ず検出できる。

L と L_{min} を固定し、ΔN を０を中心に N/10 間隔で変化させて差分出力時間を測定した結果、ΔN に関係なくほぼ同じ時間であった。次に、ΔN を０に固定し L_{min} を変化させて差分出力時間を測定した。その結果を図４に示す。差分出力時間は L_{min} に反比例する。また、共通部分の長さ L が長いほど短時間で出力できる。共通部分のない２つのファイル同士でも、例えば１フレーム以上の共通部分がないことを７秒以内という短時間で判定できる。

３．２．既存手法との比較

上記と同様に図３の２つのファイルを用いて、既存の代表的なバイナリ差分出力ツールである xdelta[3]で差分出力時間を計測した。ΔN =0 に固定した場合は、差分出力時間は $N-L$ にほぼ比例し、L =0 のときに約 50 秒であった。差分のサイズが大きいほど提案方式は優位性がある。

４．まとめ

サイズの非常に大きいファイルを比較して短時間で差分を特定できる手法を提案し、実験により効果を確認した。

参考文献

[1] 金子,“ネットワーク利用コンテンツ制作・送出システム”，映像情報メディア学会誌, Vol.60, No.5, pp.697-701, (2006)

[2] SMPTE Standard 377M: “Material Exchange Format (MXF)”

[3] http://xdelta.org/