4D-8

タイムスタンプ蓄積情報に基づくファイル関連性推定システムの開発

幸 嘉平太 †
† 大分県産業科学技術センター

1 はじめに

情報機器の記憶装置は、大容量化が急速に進んでいる。特に、パソコンのハードディスクドライブは、数10GB から数 100GB サイズの装置が主流となっている。このため、動画などの特殊な巨大ファイルを除くと、一般のワープロや表計算の文書、メールファイルなどは、ほぼ無尽蔵に格納できるようになった。過去のデータを破棄・整理する必要性が低くなり、無数のファイルが格納されているユーザも多い。

2 現状と問題点

ファイルの格納法として、フォルダの階層構造を用いるディレクトリ方式が一般的である。フォルダを小分類・中分類・大分類と階層化し、それらの中に、適当なファイルを格納する方式である。

この方式では、ファイルやフォルダが非常に多数になると、個人による管理が困難になる。ファイル管理が面倒なため、デスクトップや一時フォルダに、ファイルを乱雑に格納している利用者は多い。

OS にはファイル検索ツールが付属するが、機能的に十分ではない。デスクトップ検索など、インデックスを用いてドライブ全体を網羅的にサーチするツールも登場している。これらのツールは、基本的にキーワード型の検索である。汎用的なワードを用いると雑音の多い結果しか得られない。そもそも、適当なワードを思い出さなければ、検索ができない。

3 本技術の概要

3.1 検索方式の分類

本稿では、利用者が行うファイル検索を、大きく 2 つのタイプに分類した。キーワード型と関連型である。

- キーワード型：「ファイル名や文中に、＊＊＊が含まれているファイルを探したい」などのタイプ
- 関連型：「このファイルに関連性があるファイルを探したい」などのタイプ

キーワード型に対応するツールは多く開発されているが、関連型に重点を置いたものは少ない。しかし、あるファイルに対して、そのファイルと関連のあるファイルを探したい場面は多い。「このファイルを編集する際に、参考としたファイル」や、「このファイルと同時に用いていたファイル」を見つけたい場合である。

3.2 アプローチ

多数のファイルに対して、それら相互の関連性を手動で入力することは、現実的ではない。そこで、ファイルに付随するタイムスタンプに着目した。一般に、OS はファイルに対して、3 種類のタイムスタンプ、作成日時・アクセス日時・更新日時を自動的に付与している。この情報を基本として、ファイル相互の関連性を抽出する技術の開発を行った。

パソコンを用いてファイルを編集する際、短時間に複数のファイルにアクセスすることは多い。同時に開いたり、連続的に開いたりする。そのような環境の場合、短い時間間隔で操作を行ったファイル群は、同一の目的で編集されたファイル群と考えることができる。ごく短い時間内に、利用者が意図や目的が異なるファイルを操作することは、考えにくいからである。

例えば、ファイル A とファイル B の利用時刻が 10 秒しか離れていない場合、同一目的の作業で使用した可能性が高い。1 秒であれば、その可能性はさらに高まる。逆に、数 10 日以上も離れている場合は、それらファイル間の関連性は低いと判断してもよい。

このタイムスタンプを利用して分析するのであるが、1 ファイルに 3 個のタイムスタンプだけでは、関連性算出の根拠としては乏しい。そのため、本技術では、以下の工夫を行った。

3.3 タイムスタンプの蓄積

1 つのファイルに対して、作成日時は作成した時点の 1 つしかない。しかし、アクセス日時と更新日時は、それらの操作を行った回数分が存在する。一般に、OS は最新のタイムスタンプしか保持しない。過去の更新・アクセスによるタイムスタンプは、失われてしまう。そこで、アクセスや更新の操作を監視し、発生するたびにタイムスタンプをデータベースとして蓄積する処理を導入した。この処理により、1 つのファイルに対して、多くのタイムスタンプを得ることができる。

図 1 に、1 つのファイルに対して、複数のタイムスタンプが蓄積されたイメージ例を示す。蓄積されたタイムスタンプが多いほど、分析対象となる時間距離の組み合わせ数が密になる。単純には、Combination 数

File relevancy estimation system using with time stamps

† YUKI Kaheita (ka-yuki@oita-ri.go.jp)
Oita Industrial Research Institute (†)

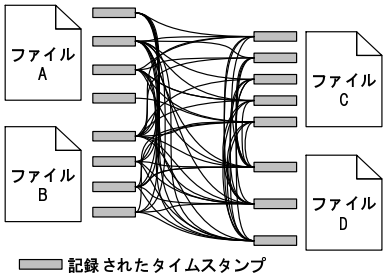

図 1: 時間距離の組み合わせイメージ

個の時間距離を得ることができる。

参考として、著者が使用するパソコン内の文書ファイルの分析例を図 2〜3 に示す。120GB のハードディスクドライブの中に、Office 系の文書ファイルが約 3,000 個格納されていた。図は、通番 406 のファイルに対して、他のファイルとの時間距離をプロットしたものである。縦軸が時間距離、横軸がファイル通番を示す。通番が離れるほど、フォルダ階層としての距離は遠くなる。

1,000 日間のスケールで見た場合、他のファイルとの時間距離は、ある程度一様に分布しているように見える。しかし、1 日間のスケールで拡大してみた場合、時間距離が短いファイルは、特定のファイル群に偏っている様子が分かる。なお、これらの図は、基礎調査として、時間距離の単純な加重平均をプロットしたものである。

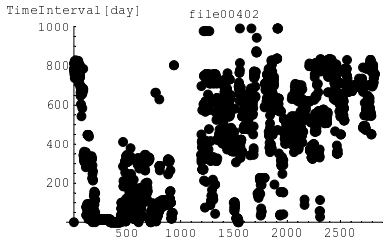

図 2: 時間距離の分布 (1000 日間幅)

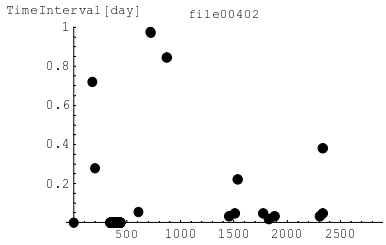

図 3: 時間距離の分布 (1 日間幅)

3.4 利用特性の反映

利用者によっては、「更新作業が中心のため、更新日時を重視して判断したい」などの作業特性があるかもしれない。この作業特性を統計処理の際、重み付けとして反映できる処理も組み込んでいる。

3.5 コピー・ペースト情報の併用

ファイル編集の作業において、ファイル間のコピー・ペースト操作もよく行われる。この操作の場合、コピー元とコピー先には、当然、高い関連性があると考えられる。このコピー・ペーストの操作情報もリアルタイムに収集し、蓄積を行う処理も加えた。そのデータとタイムスタンプ分析を併用することにより、関連性の判断がより妥当になると考えている。

4 おわりに

本技術は原理的に単純であり、処理自体も軽い。ファイルの整理や探索に苦しむ利用者には、一定の効果があると考えている。

ブックマーク管理や Web 閲覧履歴の管理などへの適用も検討している。ブラウザに無数のブックマークが登録され、見通しが悪くなっている利用者は多い。短時間内にアクセスしたサイトは、同一目的にて閲覧した可能性が高い。同様のアプローチが有効であると考えている。

本技術を開発するにあたり、主に特許情報を中心に先行技術の調査を行った。類似する出願が大手メーカーより数件見つかったが [1]〜[4]、タイムスタンプの蓄積、複数距離の統計処理、コピー・ペースト操作情報の併用など、技術的な新規性・進歩性があると考えている。逆に、学会論文などの学術的な調査は不十分な面がある。参考になる類似研究があればご指摘いただけると幸いである。

本技術は特許出願し、審査請求中である。県内企業と実施契約を結び、Windows 用ソフトウエアとして実用化する予定である。

参考文献

[1] 株式会社東芝, “ データ表示装置およびデータ表示方法 ” 特開 2000-76109

[2] 株式会社キャディックス, “ 電子ファイル管理方法及び電子ファイルを管理するプログラムを格納したコンピュータ読み取り可能な記録媒体 ” 特開 2000-339206

[3] 株式会社リコー, “ 文書閲覧システム ” 特開 2002-312402

[4] 富士通株式会社, “ 電子文書管理装置および管理方法 ” 特開 2005-25550