アクセスログ解析を用いて同一作業で使用されたファイル群を発見する手法の提案

小田切 健一† 渡辺 陽介‡ 横田 治夫†‡

†東京工業大学大学院情報理工学研究科計算工学専攻
‡東京工業大学学術国際情報センター

1 はじめに

近年のストレージ技術の進歩により、古いデータを削除することなく全て保存しておくことが容易になった。しかし、保存したデータを再利用するために大量のファイルの中から目的のファイル群を発見するのは困難である。デスクトップ検索が普及しているが、キーワードを含まないファイルも多く存在する。

我々の研究グループでは、複数ディレクトリに分散した同一作業で使用されたファイル群を発見し、仮想ディレクトリとしてユーザに提示する研究を行っている。例えば、書類作成に使用したファイルが文書・画像・データディレクトリに分散して置かれていて、ユーザがその配置を忘れてしまった場合、そのファイルは発見できなくなってしまう。我々が提案する仮想ディレクトリを用いると、作業に使用したファイル群が集約されているため、作業に用いたファイルを容易に発見することが出来る。また、我々の手法はファイルサーバーのアクセスログを用いるため、テキストを含まないファイルであっても発見することが出来る。

我々はこれまでに CO 法 [2] と FI 法 [3] という二つの手法を提案してきた。CO 法は使用時間が重複するファイルを発見する手法だが、実際の作業ではファイルの使用が必ず重複するとは限らないという問題があった。FI 法ではログファイルをトランザクションという単位に区切り、一定期間内にアクセスがあったファイルを見ることで使用時間が前後するファイルも発見可能にした。しかし、発見したファイル集合を結合していく際に要素の重複があるファイル集合同士を結合していたため、1 作業が要素の重複がない複数の集合として発見されるとそれらを結合することが出来ないという問題があった。本論文ではこの問題を系決するために、ファイル集合内のファイルの使用時間に着目したファイル集合結合処理を行う手法を提案する。

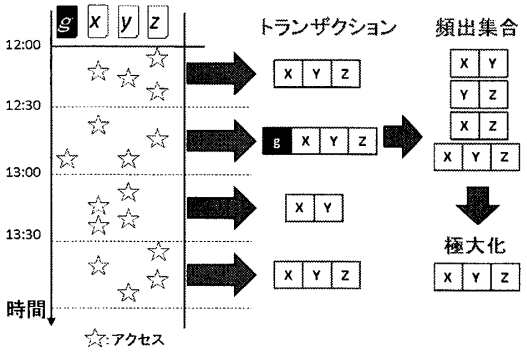

図 1: ファイル集合の発見

2 提案手法

本稿で提案する COFI 法 (Clustering using Overlap of use-time for Frequent Itemsets) は、「ログクリーニング」「トランザクションへの分割」「ファイル集合の発見」「クラスタリング」という手順でなっている。

2.1 ログクリーニング

ログファイルにはユーザの意図的なアクセスのほかに、バックアップ・自動プレビュー・ファイル検索などのユーザが意図しないアクセスも記録されている。ユーザが意図しないアクセスが含まれると、誤ったファイル同士が関係すると判定されてしまう。そこで、1 秒間に 4 ファイル以上の高頻度アクセスを機械的なアクセスとして除去を行う。

2.2 ファイル集合の発見

本手法ではアクセスログを一定期間ごとに分割し、各期間内でアクセスがあったファイル群をトランザクションとする（図 1）。これにより、多少使用時間がずれているファイルも同じトランザクションに入る。しかし、図中の g のような偶然使用したファイルもトランザクションに入ってしまう。しかし、偶然使用したファイルを含む集合は低頻度のため、頻出集合発見を

A Method for Discovering Files Used in Same Task by Analyzing Access Log

Kenichi OTAGIRI†, Yousuke WATANABE‡and Heruo YOKOTA†‡
†Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
152-8552, Tokyo, Japan
‡Global Scientific Information and Computing Center, Tokyo Intitute of Technology
152-8552, Tokyo, Japan
{otagiri, watanabe}@de.cs.titech.ac.jp
yokota@cs.titech.ac.jp

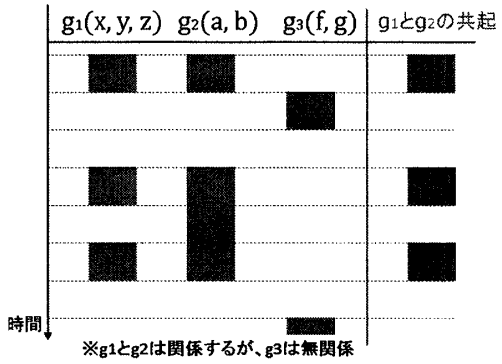

図 2: 集合内ファイルの使用時間の重複

適用することで除去することができる。頻出集合のうち他の集合の部分集合になっているものは除去する（極大化）。

2.3 クラスタリング

同時に使用されたファイル集合を発見すると、「書類 X の図表作成のファイル群」「書類 X のグラフ作成のファイル群」といった小さな単位であることが多く、「書類 X 作成作業」というような大きな作業単位になっていないことが多い。そこで、2.2 節で発見されたファイル集合のうち同じ作業に属すると思われるファイル集合同士を結合する必要がある。この際、FI 法では要素が類似するファイル集合同士を結合していた。しかし、「図表作成のファイル群」と「グラフ作成のファイル群」では使用しているファイルが全く異なり重複がなく、この集合同士を結合することが出来なかった。

COFI 法では、見つかったファイル集合内の各ファイルの使用時間に着目する。ここではファイルの使用時間はどのトランザクションに入っているかに対応する。例を図 2 に示す。g_1, g_2, g_3 は前ステップで発見されたファイル集合を示し、x, y, z, a, b, f, g は集合内のファイル要素を示す。各ファイル集合内のファイルが入っているトランザクションを赤で示している。g_1 と g_2 に着目すると、互いにファイル要素の重複はないが、互いに同じトランザクションで出現したファイルを含んでおり、関係があると分かる。この関係の強さを同一作業指数と呼ぶ。g_1, g_2 を集合、集合 X 内のファイルが含まれるトランザクション集合を $TranIDs(X)$ とすると、同一作業指数は以下で定義される。

$$\text{同一作業指数}\ (g_1, g_2)$$
$$= dice(TranIDs(g_1), TranIDs(g_2))$$
$$= \frac{2|TranIDs(g_1) \cap TranIDs(g_2)|}{|TranIDs(g_1)| + |TranIDs(g_2)|}$$

上記の同一作業指数を用いてクラスタリングを適用し、同じ作業に属していたファイル集合を結合する。結合されたファイル集合には作業のファイルが集約されているので、それらを仮想ディレクトリとして出力する。

表 1: 実験結果：全正解セット平均

手法	Precision	Recall	F-measure
CO	0.589	0.390	0.345
FI	0.756	0.401	0.474
COFI	0.679	0.436	0.503

3 比較実験

4 ユーザの約 1 年間半ほどの Samba[1] のアクセスログを用いて、これまでに提案した CO 法,FI 法と比較実験を行った。正解セットには 8 セットの複数ディレクトリにファイルが分散している作業を用いた。仮想ディレクトリに手法の意図したとおりに複数ディレクトリに分散したファイルが置かれているかどうかを評価した。各々の正解セットが含まれる仮想ディレクトリについて正解ファイルや無関係ファイルの数を調査した。各手法の平均 Precision,Recall,F 値を表 1 に示す。提案手法が分散したファイルを仮想ディレクトリに集約することが出来ており、これまでに提案した手法に対し Recall が上昇し F 値でも良い結果を出していることが分かる。比較実験の詳細は文献 [4] を参照されたい。

4 まとめと今後の課題

本論文では、アクセスログを用いて複数ディレクトリに分散した同一作業に属するファイル群を発見する新しい手法として、ファイル集合内のファイルの使用時間に着目したクラスタリングを用いる手法を提案した。

今後の課題として、より多様な正解セットによる評価、メモリ・計算量削減のための頻出集合発見の近似アルゴリズムの採用等があげられる。

謝辞

本研究の一部は文部科学省科学研究費補助金特定領域研究 (#21013017) の助成により行われた。

参考文献

[1] *Samba.* http://us3.samba.org/samba/.

[2] 小田切, 渡辺, 横田. アクセス履歴を用いたユーザの作業に対応する仮想ディレクトリの生成, 2009. 第 1 回データ工学と情報マネジメントに関するフォーラム（DEIM2009）.

[3] 小田切, 渡辺, 横田. ユーザ作業を反映する仮想ディレクトリ生成のためのアクセス履歴解析手法, 2009. 第 148 回 データベースシステム・第 95 回 情報学基礎 合同研究発表会.

[4] 小田切, 渡辺, 横田. 頻出ファイル集合のアクセス時間を考慮した仮想ディレクトリ生成手法, 2010. 第 2 回データ工学と情報マネジメントに関するフォーラム（DEIM2010）.