# キーワード非含有ファイルを検索可能とするファイル間関連度を用いた検索手法

渡部徹太郎† 小林隆志‡ 横田治夫†

†東京工業大学 大学院情報理工学研究科　‡名古屋大学 大学院情報科学研究科

## 1 はじめに

近年，ファイルシステム内のファイル数が爆発的に増加しており，膨大な数のファイルを管理するために，適切なディレクトリ階層を構成することは難しくなっている．また，全てのファイルが適切にディレクトリ階層に格納されていたとしても，ファイル名が不適切であったり，ファイルが深い階層に格納されていた場合，ディレクトリ構造を辿るだけでは求めるファイルを探すのは困難である．

この問題に対しファイルにメタデータを付加することで求めるファイルを探索する研究も多くされていが，増え続けるファイル群に対して適切なメタデータの付加をユーザに要求するのは非現実的である．そのような背景から，ファイルシステムに対する全文検索を利用するファイル検索システムが用いられてきたが，キーワードを含まない画像ファイル，データファイル等を探し出すことはできないという問題があった．Google Image Search では HTML ファイル画像への参照情報を用いることで，画像に対するキーワード検索を提供しているが，一般のファイルでは基本的には関連ファイルへの参照情報が無い場合が多く適用できない．

このような問題に対し，我々は全文検索ができないファイルに対しても，キーワード検索を可能にする手法を提案してきた．さらに，提案手法の実装を用いて，被験者実験を行うことで，従来の手法と比べ検索結果の適合率と再現率が改善することを示してきた．

本稿では本研究の概要を紹介する．なお，ファイル間関連度の詳細な算出法は別報 [1] を，ファイル検索法，実装，評価実験の詳細は別報 [2] を参照されたい．

## 2 提案手法

本研究では全文検索できないファイル群に対してもキーワード検索を可能する手法を提案する．提案手法は大きく二つのステップからなる．

1. 事前に検索対象となるファイル群の全文インデックスを構築すると共に，ファイルアクセスログからファイル使用情報を抽出しファイル間関連度を算出する．
2. キーワード検索要求が来た場合は全文検索を行い，全文検索結果と関連している各ファイルに対して関連度を考慮した得点付を行い，得点の高い順に提示する（図 1 参照）．

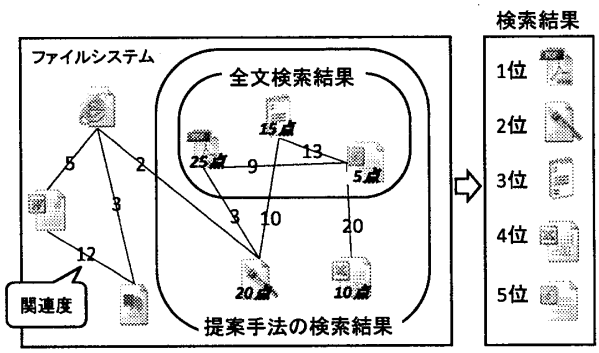


図 1: 提案するファイル検索法の例

### 2.1 ファイルアクセスログからの関連度の算出

本研究ではファイル間関連度として，頻繁に同時に使うファイル群に着目している．コンピュータ上である作業をする際，関連するファイルを同時に開いて参照しながら，あるいは必要な部分をカットアンドペーストしながら進めることが多い．このような場合，作業毎に使用するファイルは同じものが多いため，頻繁に同時に使われるファイル群は関連が高いと考えられる．

本研究ではこのようなファイル群を求めるために，ファイルサーバのアクセスログからユーザのファイル使用情報を抽出する．

#### 2.1.1 ファイル使用情報の抽出

本研究ではファイルサーバのアクセスログからユーザ毎にファイル毎の使用時間情報を抽出する．

対象を Windows に限定し，Samba を利用したファイルサーバのアクセスログを解析した結果，ファイル使用時間は，オープン時刻からクローズ時刻までとできない場合が多数存在することが分かった．

そのため本研究では，アクセスログから抽出したユーザの活動時間等を用いるログクリーニングの手法を提案し，ユーザーの実際の使用時間に近いファイル使用情報を抽出する．

#### 2.1.2 関連度の算出

本研究では関連度の高いファイル同士は長い期間，各作業の度に，同じタイミングで利用されるものと仮定

---

**A Method using Inter-file Relationship to Search Keyword-nonincluding Files**

Tetsutaro WATANABE†, Takashi KOBAYASHI‡, Haruo YOKOTA†

†Graduate School of Information Science and Engineering,Tokyo Institute of Technology. 152-8552, Tokyo, Japan.

‡Graduate School of Information Science, Nagoya University. 464-8603, Nagoya, Japan.

†tetsu@de.cs.titech.ac.jp, ‡tkobaya@is.nagoya-u.ac.jp, †yokota@cs.titech.ac.jp

し，前節で抽出した各ファイルの使用情報の関係を基にファイル間の関連度を計算する．

各ファイルの使用情報の関係を表現するために，二つのファイルのファイル使用時間の重なりを共起と呼び，以下の四つ関連度要素を導入している．

$T$ 共起時間の累計が長いほど大きい．
$C$ 共起回数が多いほど高大きい．
$D$ 共起の間隔が離れているほど大きい．
$P$ ファイル使用開始パターンが類似しているほど大きい．

これらのの関連度要素を用いて関連度 $R$ を次のように算出する．$R = T^{\alpha} \cdot C^{\beta} \cdot D^{\gamma} \cdot P^{\delta}$

### 2.2 関連度を用いたファイル検索

前節で算出したファイル間関連度を用いたファイル検索方法を説明する．まず入力されたキーワードで全文検索を行う．次に，全文検索結果に含まれる各ファイルと関連しているファイルに対して，全文検索エンジンにより計算された TF-IDF による得点と，ファイル間関連度を考慮した得点付けを行う．最後に得点の高いファイル順に提示する．ここで，得点付けは TF-IDF の得点が高いファイルと関連の強いファイルに高得点がつくように計算する．

## 3 実装

本研究では提案手法を実装したファイル検索システム FRIDAL (File Retrieval by Inter-file relationship Derived from Access Log) を作成した．

FRIDAL は CIFS ファイルサーバとして広く利用されている Samba の debug level 2 のログファイルを解析し，ファイル間関連度を DB に蓄積する機能と，我々が提案する検索機能を有する Web アプリケーションとして実装されている．

Samba のログを利用するため，既存のファイルシステムに大きな変更を加えることなく容易に導入することが可能である．また，全文検索エンジンには Hyper Estraier[3] を用いた．Hyper Estraier は N グラムインデックス法を用いているため形態素解析に起因する検索漏れが生じないという利点がある．

## 4 評価実験

提案手法の有効性を評価するために，被験者実験を行った．関連度算出には 3 人の被験者の約 1 年間のファイルアクセスログを用いた．また関連度算出式の各乗数 $(\alpha, \beta, \gamma, \delta)$ は予備実験の結果より $(1, 1, 0.5, 0.5)$ が最適であると判断した．

実験方法は各被験者がキーワードに関連があると判断したファイル群を正解セットとし，提案手法を含む幾つかの方法での検索結果の 11 点平均適合率と上位 20 件の適合率と再現率の平均値をそれぞれ比較した．

比較する検索手法は FRIDAL，全文検索，ディレクトリ検索，Connections の 4 手法である．Connections は [4] で提案されている得点計算法を用いたファイル検索である．ディレクトリ探索は全文検索結果の含まれるディレクトリも探索する方法で，ユーザが全文検索で目的のファイルを見つけられない場合次に取る行動として自然であると考える．

11 点平均適合率は図 2 のようになり，上位 20 件の適合率と再現率の平均値は以下の表になった．

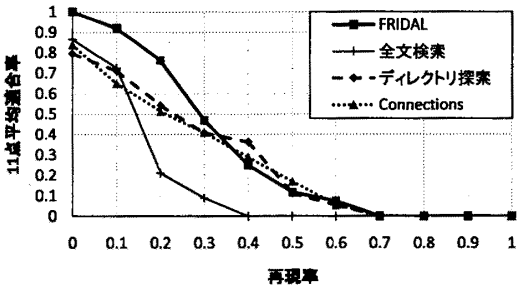


図 2: 各検索の 11 点適合率の平均値

| | 適合率の平均 | 再現率の平均 |
|---|---|---|
| FRIDAL | 72 % | 16 % |
| 全文検索 | 62 % | 12 % |
| ディレクトリ探索 | 62 % | 13 % |
| Connections | 48 % | 10 % |

図 2 では FRIDAL の値が他手法よりも高い値を示しており，上位 20 件の適合率，再現率の平均値共に最も高い値を示している．このことから，アクセスログから抽出したファイル間の関係を利用することで，全文検索のみや，検索結果を含むフォルダを探索する方式より効率よく関連ファイルを提示できることが確認できた．

## 5 まとめと今後の課題

既存の全文検索ではキーワードを含んでいないファイルは，そのキーワードと関連が在っても検索できないという問題点があった．

本稿では，全文検索できないファイルをキーワード検索可能とする手法の概要を示した．また，提案手法を実装した検索システムを紹介し，システムを利用した被験者実験によって提案手法の有効性を確認した．

今後の課題としてはファイルの複製，移動，改名処理への対処，関連の方向を考慮したファイル間関連度の改良，検索要求に応じた検索結果提示方法の改善などが挙げられる．

## 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 (19024028)，東京工業大学 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」および独立行政法人科学技術振興機構戦略的創造研究推進事業 CREST の助成により行なわれた．

## 参考文献

[1] 渡部徹太郎, 小林隆志, 横田治夫: ファイル検索に向けたアクセスログからのファイル間関連度の導出, 日本データベース学会論文誌 Vol.6, No.2, pp.65-68, 2007.

[2] 渡部徹太郎, 小林隆志, 横田治夫: キーワード非含有ファイルを検索可能とするファイル間関連度を用いた検索手法の評価第 19 回電子情報通信学会データ工学ワークショップ (DEWS2008) 論文集, 2008.

[3] Hyper Estraier, http://hyperestraier.sourceforge.net/

[4] Soules, C. A. N., and Ganger, G. R. Connections: using context to enhance file search. In Proceedings of the 20th ACM Symposium on Operating Systems Principles, pp.119-132. 2005.