2H3-3

ファイルネットワークに基づいた情報の抽出と可視化

Extraction and Visualization of Information Based on the File Network

福井　秀徳[*1]　森田　哲郎[*2]　岡野　真一[*2]　沼尾　正行[*3]　栗原　聡[*3]
Hidenori Fukui　Tetsuo Morita　Shinichi Okano　Masayuki Numao　Satoshi Kurihara

[*1]大阪大学大学院　情報科学研究科　情報数理学専攻
Department of Information and Physical Science, Graduate School of Information Science and Technology, Osaka University

[*2]住友電気工業株式会社
Sumitomo Electric Industries, Ltd.

[*3]大阪大学　産業科学研究所　知能システム科学研究部門
Division of Intelligent Systems Science, The Institute of Scientific and Industrial Research, Osaka University

Recently, according to rapid development of information technology, the use of personal computer(PC) is increasing. One PC contains 10000 or more files, and the number of them become huger in an organization. In this research, we constructed the system that visualize transitions of information by constructing the file network. By using this system, distribution of data can be easily understood from the time-based interface.

1. はじめに

近年の情報技術の発達に伴い，それを取り巻く環境も変化に迫られている．我々の生活における多くの情報が電子化され，従来にはなかった数々の便利なサービスの恩恵が受られるようになった一方で，膨大な情報を人間の意識下で処理することができないという問題が生じてきている．ビジネス，プライベートの両面において，我々の生活に欠かせない存在となったパーソナルコンピュータ (PC) においても，この問題は顕著に現れている．一台の PC が保有するファイル数は万単位，多い場合では十万単位となり，組織が保有するファイル数は更に膨大なものとなる．そこで，膨大なファイルの管理に対応できる新しいアルゴリズムの開発が必要不可欠となっている．本研究ではファイル同士の繋がりを抽出し，時間軸インタフェースに適用することでファイルの変遷を視覚化するシステムを提案・実装し，有用性についての検証を行った．検証の結果，ファイル検索や情報遷移の把握において，本システムが有用に働くことがわかった．

2. 従来研究

本研究において提案するシステムは，ファイルが保有するテキスト情報とユーザのファイルアクセスログ情報という二つの手掛かりに基づいて作成したファイルネットワークを用いてファイルの変遷を視覚化するものである．リンク (関連性に基づいた繋がり) を辿ることで関連性の深いファイルを導き出すという意味では, ファイル検索ツールの側面をもち, 情報の流れを視覚化するという意味では, 情報管理システムとしての側面をもつ. そこで, 本章では, まずテキスト情報を用いたファイル検索とアクセス情報を用いたファイル検索についての従来研究・サービスを紹介する．次に情報管理システムという視点で見た場合の従来研究とのアプローチの差異について述べる．

連絡先: 栗原　聡，大阪大学　産業科学研究所　知能システム科学研究部門，大阪府茨木市美穂ヶ丘 8-1，06-6879-8426，06-6879-8428，kurihara@sanken.osaka-u.ac.jp

2.1 テキスト情報を用いたファイル検索

ファイル名, もしくはファイル自身が持つテキスト情報はファイルを特定する際の有用な手掛かりとなる．GoogleDesktop[1] 等のサービスでは, インデックスを利用した高速な検索を実現しており，Web ブラウザに検索文字列を入力すると，Web の検索結果と並んで PC 内の該当ファイルが提示される．Dumais ら [2] はファイルが保有するテキスト情報とメタデータによるファイルフィルタリングを実現している．Cutrell ら [3] は検索対象のファイルについての様々な情報に対応した高度な検索インタフェース (PHLAT) を実装した．

2.2 アクセス情報を用いたファイル検索

ユーザのファイルアクセス情報を用いて, 時間的な手掛かりからファイルを特定する方法も頻繁に用いられる．ユーザは全てのファイルに対して均一にアクセスするのではなく，アクセス対象となるファイルには偏りがある．一度アクセスしたファイルは，近い未来に再びアクセスされることが予想されるため，最近のファイルアクセス履歴を提示することは，ユーザの探しているファイルを発見する上で有用であるといえる．Ringel ら [4] は公的, および私的なイベントを時系列インタフェースに表示することで, ユーザがイベントとの相対的な時間感覚を頼りに効率よくメールを特定できることを示した．大澤ら [5] は, ユーザのデータ参照時間や回数などから算出した着目度に基づいたインタフェースを実装した．渡辺ら [6] はアクセス履歴から各ファイル同士の関連性を数値化し，検索結果に関連値の高いファイルを添えることでファイル検索の効率化を図った．また, ファイル検索とは目的が異なるが, 暦本 [7] はコンピュータの作業履歴を蓄積し, 時間移動によって過去の作業環境の再現を行うとともに, 時間に伴うＰＣ環境の遷移を視覚的に表した．

2.3 情報管理システム

大平ら [8] はソフトウェア開発データを自動収集・解析するプロジェクト管理ツール Empirical Project Monitor (EPM) を作成した．EPM では, プロジェクトにおける様々な統計データの時間的な推移が取得でき, ソフトウェア開発プロセスを定量的な視点で確認することができる．EPM はこのようにプロジェクトの全容を把握するという点で優れたツールである．一方で, 本稿において提案するシステムは関連の深いファイル同

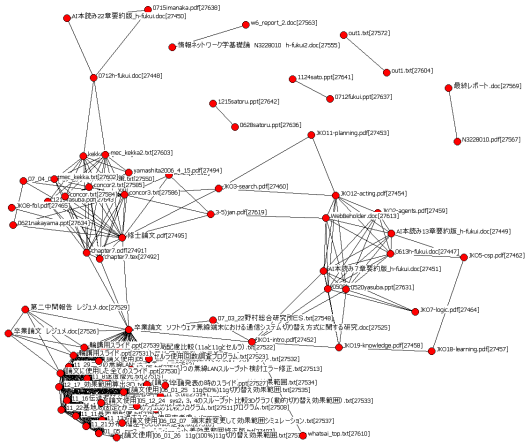

図 1: 重要語の共起に基づいたファイルネットワーク

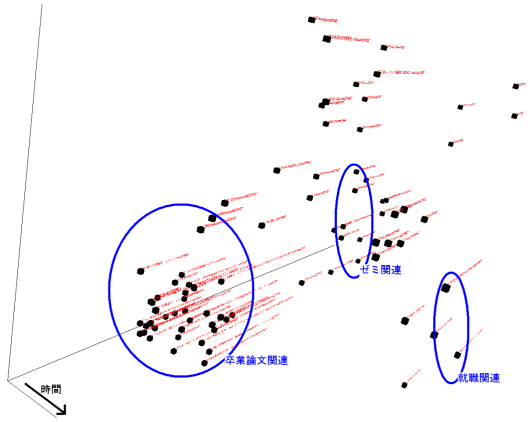

図 2: 重要語共起ネットワークに時間軸を加えた 3 次元図

士を繋いだネットワークから, グループ内の情報の流れとユーザの振舞いを詳細に把握できる点において強みがあるといえる. 例えば, 本システムを適用することにより, プロジェクトに何らかの問題が生じたとき, プロジェクト全体の詳細な流れを追っていく中で, どの過程に問題があるかを見出せる可能性がある.

本研究と類似するアプローチとして，バージョン管理システムが挙げられる [9][10]. バージョン管理システムでは，ファイル内容の更新を監視し，各更新時のファイルの状況を管理・再現する. 通常，同一ファイルの内容が更新された場合，以前の情報は消去される. ソフトウェアのコードや原稿管理などで過去のバージョンのファイル情報が必要になることがあるが，こういった際にはバージョン管理システムの利用が有効である. ファイル内の情報の変化を抽出するバージョン管理システムに対して，本システムでは，異なるファイル間における情報のやりとりについても対象としている.

3. ファイルのテキスト情報とアクセス時刻情報に関する検証

ここでは, テキスト情報とアクセス情報について, 簡単な検証を交えた上で両者の特性を挙げ, 組み合わせることの有用性について述べる. 本章で行う予備検証は，第 4 章で述べる提案手法の有用性の裏付けとなるものである.

3.1 テキスト情報から得られるファイル相関性の検証

図 1 は筆者が PC 内で作業を行う際に利用するディレクトリ内のデータに対して, テキスト情報を用いて作成したファイルネットワークである. 各ドキュメントから, 重要語を抽出し, 重要語が一つ以上共起しているファイル同士にリンクを張った. テキスト解析の対象としたのは, 一般的なユーザが使用する機会の多いテキスト情報を含むフォーマット (プレーン・テキスト・ファイル，Microsoft Office Word ドキュメント，Microsoft Office Excel ファイル，Microsoft Office PowerPoint ファイル，PDF ドキュメント，など) である. なお, 重要語が共起しなかったノード (エッジが存在しないノード) の表示はここでは省略している.

グラフの左側中央, 右側中央, 左側下部の三箇所において, 類似性の高いファイルの集まりが見られる. これらのファイルは比較的内容が近いファイルの集合となっていることから, テキスト情報のみで, 大まかな分類が可能であることがわかった. しかしながら, グラフの中には, 筆者にとって直感的に分かり辛いファイルの繋がりも存在する. これは, 利用者自身が, 関連があると自覚していないファイル同士から共通の重要語が抽出されてしまうことが原因である. このように, テキスト解析は高い精度でファイルの相関性を抽出するが, ユーザが意図しないファイル同士が関連のあるファイルとして抽出される恐れがあることが分かった. また, ここで得られた結果のように, ユーザの個性や特徴に応じた結果は期待できない.

3.2 アクセス情報から得られるファイル相関性の検証

図 1 の平面グラフに対して新たに第 3 の軸としてファイルの最終更新時刻を加え,3 次元空間上にノードをプロットしたのが図 2 である. 新たに時間情報が与えられたことで, 就職活動関連ファイルのような時間に強い依存性を持ったファイル群が新たなまとまりとして抽出されている. ユーザはなんらかの目的のためにファイル操作を繰り返しており, ここで見られた時間的に近い関係にあるファイルは共通の目的のために使用されたファイルであると考えられる. ユーザは自身のファイル操作をイベントの前後関係によって記憶に留めることが多いため, 近い時刻にアクセスされたファイル同士は, ユーザの記憶の想起を促す効果も期待できる. 時間情報のみを用いた解析では高い精度の関係抽出は難しいが, テキスト解析の結果と組み合わせることで, ユーザに応じたファイル相関性の抽出が可能になる可能性がある.

4. 提案手法

我々はあるドキュメントを作成する際に，他のドキュメントを参照することが多い. 例えば，論文を書く場合について考えてみても，過去の論文を参照する，Web で調べ物をする，メールで締め切り日時を確認する，といったように他のドキュメントへのアクセスが頻繁に行われている. ユーザはなんらかの目的のためにファイル操作を繰り返していることから，近い時刻にアクセスされたファイルは共通の目的のために使用されたファイルであると考えることができる.

4.1 リンク強度の算出

テキスト同士の関係の判断やファイルネットワークの構築には，ファイルアクセスイベント同士の繋がりの強さを示すリンク強度を利用する. アクセス時刻が共起しているファイルアクセスイベント A(以下，イベント A) とファイルアクセスイベント B(以下，イベント B) が存在すると仮定する. イベント A の被アクセスファイルの重要語数を $KeyWord_A$，イベント B

の被アクセスファイルの重要語数を $KeyWord_B$，これら二つのファイルに共通して見られる重要語数を $KeyWord_{A\cap B}$ としたとき，イベント A から見た，イベント n に対するリンク強度（以下,「$L(A \to B)$」）を次の式から算出する*1.

$$L(A \to B) = KeyWord_{A\cap B}/KeyWord_A \qquad (1)$$

ただし，両者が「明らかに関連があるイベント」であると判断された場合のリンク強度は 1 とする．明らかに関連があるイベントとは以下のようなものである．

- ファイルの複製元と複製先に対するアクセスイベント
- ファイルの移動元と移動先に対するアクセスイベント
- ファイル名変更元と変更先に対するアクセスイベント
- 同一ファイルに対するアクセスイベント
- 同一電子メールの送信イベントと受信イベント
- 同一サブジェクトのメールイベント
- 添付ファイルに対するアクセスイベントと添付元のメールイベント

4.2 関係抽出

一方が読み込みイベントであり，かつ，他方が書き込みイベントであった場合，作業の中で，情報遷移が生じた可能性がある．そこで，表 1 に示した条件にて，二つのファイルアクセスイベントを以下の 4 つの関係のいずれかに特定する．

表 1: リンクが示す関係 (R が読み込みイベント，W が書き込みイベントとする)

$L(R \to W)$	$L(W \to R)$	関係
2/3 以上	2/3 以上	複製
2/3 以上	2/3 未満	取り込み
2/3 未満	2/3 以上	部分引用
2/3 未満	2/3 未満	部分共有

4.3 ネットワーク描画アルゴリズム

ファイル同士の繋がりをよりマクロな視点で確認するためには，ネットワークとして視覚化するのが有効である．本システムではトリガとなるイベントが与えられると, このファイルアクセスイベントを第 1 ノードとして, リンクを辿りながら次々と別のノードを探索していくことで，ネットワークを描画する．各ノードは第 1 ノードとの関係値を持っており, この関係値が最小関係値 (可変パラメータ) を下回る場合, 無効ノードとなる．探索中に無効ノードが見つかった場合, 無効ノードのリンク先を探索することはしない．第 1 ノードからのホップ数が i のノードの関係値 W_i は以下のように計算される．

$$W_i = W_{i-1} \times L_{i-1\to i} \quad (\text{ただし } W_0 = 1) \qquad (2)$$

W_{i-1} は W_i に対してリンクを張るノードである．W_i の関係値が複数存在する (第 1 ノードまでのルートが一つでない) 場合は, 最大のものを採用する．

*1 重要語抽出に関しては既存のアルゴリズムを用いる．具体的には茶筌 [11] を用いて形態素解析を行い，その結果から TermExtract[12] によって重要語を抽出する．TermExtract では単名詞の連結に基づいた重要語の抽出がおこなわれる．

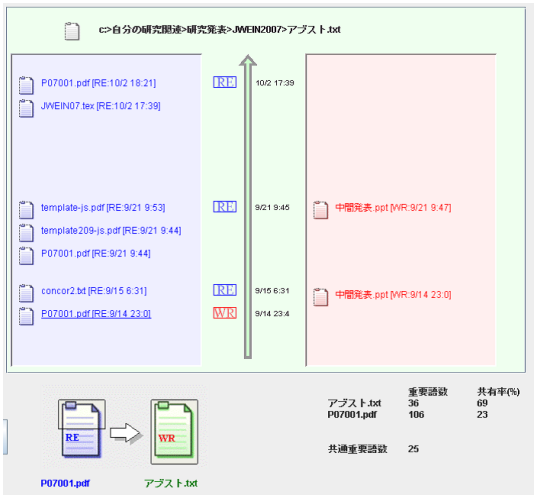

図 3: ファイル変遷の描画

5. 動作検証

5.1 ファイル検索システム

図 3 は実際に本システムを筆者の PC に適用した結果である．中央の四角いアイコンは，クエリファイルに対するアクセスを示しており，その左右には同時刻にアクセスされた他のファイルが示される．関連ファイルにマウスカーソルを合わせると，クエリファイルとの関係を示すアイコンが下部に表示される．ユーザはこれらの情報を頼りに，目的のファイルや，新たなクエリを発見することができる．

実際に本システムを検索システムとして用いたときの有用性について，被験者実験を通して検証した．4 人の被験者に無作為に選択した 12 のドキュメントに対して編集作業を依頼し，後日，作業時に扱った参照ファイルや被験者自身が作成したファイルを検索してもらった．編集内容は要約，文字列のコピー・ペーストといった，一般的なものをこちらで指定した．いずれの場合も被験者はあるファイルを参照しながら，新たなファイルを作成する．本システムにおける検索の際には，これらの二つのファイルの一方をクエリファイルとし，もう一方のファイルを検索する．ファイルをクエリとして用いた検索が可能である本手法と，ファイルの保存場所を頼りにディレクトリを辿る一般的な検索手法とを比較した．尚，両手法において，検索中にテキストエディタを用いてファイルの内容を閲覧することも可能とした．

表 2 によると，ディレクトリ構造の複雑さやファイル数に影響を受け易い一般的なファイル検索手法に比べて，本手法はファイルの検索に要した時間やクリック数において安定して優れた結果が得られた．また，ファイル同士の関係を提示する本システムでは，ファイルの中身を改めて確認する手間を省く効果も見られた．

表 2: 被験者実験の結果．() 内は標準偏差

	階層構造に基づいた検索	本システムによる検索
平均検索時間 [秒]	65.1(41.5)	20.2(17.3)
平均クリック数 [回]	6.88(6.15)	2.42(1.28)
平均ファイル閲覧数 [回]	0.917(1.02)	0.333(0.816)

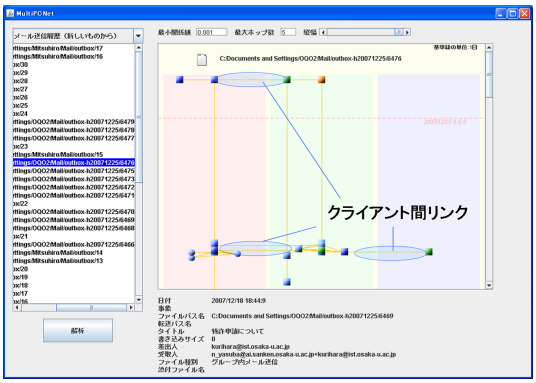

図 4: 複数端末を対象にしたファイルネットワーク 1

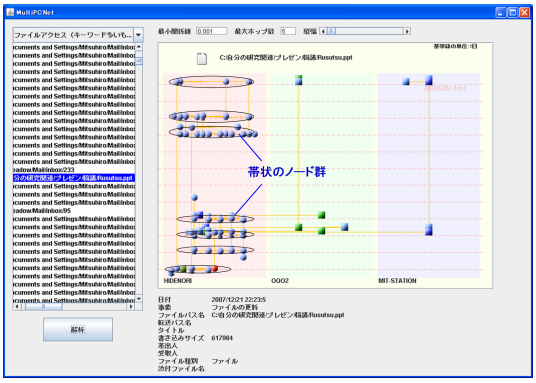

図 5: 複数端末を対象にしたファイルネットワーク 2

5.2 情報管理システム

組織内に分布する関連ファイルの所在を把握するために，本手法を応用することができる．例えば，ある機密ファイルをクエリとしたとき，そのファイルから派生した二次的な機密ファイルの存在が明らかになれば，セキュリティ管理の点において有用である．図 4 と図 5 は本手法によって得られたファイル同士の繋がりを，ネットワーク化し，時間軸インタフェースに描画している．描画ウィンドウは，左右中央の 3 つのエリアに分かれており，それぞれのエリアは各クライアントマシンに対応する．円形のノードはファイルへのアクセス，角形のノードはメールへのアクセスを示している．

図 4 は各クライアントがメールにより情報を伝達する様子を描画している．複数のエリアを横断する形で張られた水平のリンクが複数見られるが, これらはクライアント間でのメールの送受信を示している．またそれらのメールに付随する形で関係の深いファイルの存在が明らかになっている．本システムでは，グループ全般に跨る相関ネットワークから，端末間での情報の遷移を時系列インタフェースから容易に把握できる．

図 5 において, 左側に示されたクライアントに着目すると, 定期的に帯状のノード群を見ることができる．時間的, 内容的な共起が見られるこれらのファイル群は一連の作業の中でアクセスされたものであることから, ここで形成された帯状のネットワークはユーザの活動そのものを示しているとみなすことができる．このように本システムではネットワークを手掛かりにして，それに関係する人の活動の履歴を追跡，分析することも可能となる．

6. おわりに

本研究ではテキスト情報とファイルアクセス情報を組み合わせることでファイルの相関性を抽出し，時間軸インタフェースに適用した．本システムをファイル検索システムとして用いた場合の有用性を被験者実験による検証を交えて示した．さらに，提案手法を複数のＰＣに適用したところ，ファイルネットワークから組織内の情報の流れが容易に把握できることが分かった．

謝辞

本論文をまとめるにあたり, 住友電工システムソリューション株式会社の吉江信夫氏をはじめ, 住友電気工業株式会社の研究員の方々には多大なる御協力および貴重な御討論をいただきました．この場をお借りして, 謹んで感謝の意を示させて頂きます．

参考文献

[1] ：Google Desktop, http://desktop.google.com/en/.

[2] Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R. and Robbins, D.: Stuff I've Seen: A system for personal information retrieval and re-use, *Proceedings of SIGIR 2003*, pp. 72–79 (2003).

[3] Cutrell, E., Robbins, D., Dumais, S. and Sarin, R.: Fast, flexible filtering with PHLAT—Personal search and organization made easy, *In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM, pp. 261–270 (2006).

[4] Ringel, M., Cutrell, E., Dumais, S. and Horvitz, E.: Milestones in time: The value of landmarks in retrieving information from personal stores, *Proceedings of Interact 2003*, pp. 184–191 (2003).

[5] 大澤　亮，高汐一紀，徳田英幸：俺デスク:ユーザ操作履歴に基づく情報想起支援ツール，情報処理学会第 47 回プログラミング・シンポジウム (2005).

[6] 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，*DBSJ Letters*, Vol. 6, No. 2, pp. 65–68 (2007).

[7] Rekimoto, J.: Time-Machine Computing：A Time-centric Approach for the Information Environment, *In UIST '99:Proceedings of the ACM Symposium on User Interface Software and Technology*, ACM, pp. 45–54 (1999).

[8] 大平雅雄，横森励士，阪井　誠，岩村　聡，小野英治，新海　平，横川智教：ソフトウェア開発プロジェクトのリアルタイム管理を目的とした支援システム，電子情報通信学会論文誌，No. 2, pp. 228–239 (2005).

[9] Rochkind, M.: The source code control system, *IEEE Trans Software Eng SE-1*, pp. 364–370 (1975).

[10] Tichy., W. F.: RCS - a system for version control, *Software Practice and Experience*, Vol. 15, No. 7, pp. 637–654 (1985).

[11] ：形態素解析システム茶筌, http://chasen.naist.jp/hiki/ChaSen/.

[12] ：TermExtract, http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html.