社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－ICS－151（7）
2008／3／6

# マルチユーザに対応したファイルネットワークの構築と時系列インターフェースへの適用

福井 秀徳† 森田 哲郎†† 岡野 真一†† 沼尾 正行††† 栗原 聡†††

† 大阪大学大学院 情報科学研究科 情報数理学専攻
†† 住友電気工業株式会社
††† 大阪大学 産業科学研究所 知能システム科学研究部門

**あらまし** 近年，情報技術の急速な発達に伴い，パーソナルコンピュータ (PC) の使用機会が増加している．一台のPCが保有するファイル数は数万程度にのぼり，組織で保有するファイル数となると更に膨大なものとなる．プロジェクトのスケジュール調整や，工程の確認を行うプロジェクト管理者は，組織内の主要な情報の所在や流れを常に把握することを求められるが，このような大量のファイルの中から個々のＰＣに分散するファイルの変化を把握することは困難である．本研究では，ファイルネットワークを構成することで，情報の流れを視覚化するシステムを構築した．本システムにより，データの分布状況が時系列インターフェースから容易に把握できることがわかった．

**キーワード** テキスト解析，アクセスログ，時系列インターフェース，デスクトップ検索，情報管理システム

# Construction of File Network for Multi-User and Application to Time-Based Interface.

Hidenori FUKUI†, Tetsuo MORITA††, Shinichi OKANO††, Masayuki NUMAO†††, and Satoshi KURIHARA†††

† Department of Information and Physical Science, Graduate School of Information Science and Technology, Osaka University
†† Sumitomo Electric Industries, Ltd.
††† Division of Intelligent Systems Science, The Institute of Scientific and Industrial Research, Osaka University

**Abstract** 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. For adjustments of the schedule and checks on the process, the project manager should understand whereabouts and transitions of main information in the organization, but it is difficult to understand transitions of files distributed to each PC in such a situation. 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.

**Key words** text analyze,access log,time-based interface,desktop search,information management system

## 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，従来にはなかった多くの便利なサービスの恩恵が受け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．組織の管理者はプロジェクトの進行，もしくはセキュリティ管理のために，所属メンバの活動や情報の流れを把握する必要があるが，大量のデータの中から，知りたい情報の

流れを抽出・把握するのは難しい．本研究では，ファイルのテキスト情報とユーザのファイルアクセス情報を用いて，ユーザにとって関連性の高いファイル同士が繋がったファイルネットワークを作成し，時系列インターフェースに適用する手法を開発した．本システムを用いることで，組織の中での情報の遷移と，各メンバの活動が容易に把握できることがわかった．以下 2. において，デスクトップ検索，プロジェクト管理ツールに関する既存のサービスと研究について紹介する．3. でテキスト情報とファイルアクセス情報の特性についての検証を行い，4. では 3. の結果を踏まえた上で，本研究のファイル相関抽出法を述べる．5. において，実際に我々が実装したシステムを使用し，有用性についての検討を行い，6. にて，まとめと今後の課題を述べる．

## 2. 関連研究

本研究において提案するシステムはファイルネットワークを用いて複数のコンピュータ間における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，プロジェクト管理システムとしての側面をもつ．本システムが描画するファイルネットワークは，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という二つの情報に基づいて作成している．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・サービスを紹介し，後半ではプロジェクト管理ツールという視点で見た場合の他研究とのアプローチの差異を述べる．

### 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，PC 内ファイルの各テキスト情報を高速に検索できるようにインデックスを作成し，ファイルアクセスを支援する．Windows Vista などの OS ではインデックスを用いた高速な検索機能があらかじめ用意されている．また，検索対象のファイルについて，ユーザが知っているあらゆる情報に対応するため，メタデータやタグ取り付け機能などを用いた，より高度な検索インターフェースの提案もされている [2]．

### 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．メール管理では，受信メールを時刻順に並べて，日時からメールを特定するといった方法が有効である．Ringel らは公的，および私的なイベントを時系列インター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 [3]．大澤らは，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時間軸インターフェースを実装した [4]．また，ファイル検索とは目的が異なるが，暦本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 [5]．

なお，本稿では，テキスト情報とアクセス情報に関して，どちらかの一方の情報を重視するのではなく，互いの特性を考慮し両者の長所を組み合わせた手法を提案する．テキスト情報とアクセス情報の特性，ならびに提案の具体的な内容については後の章で詳しく述べる．

### 2.3 プロジェクト管理ツール

大平ら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 [6]．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，本研究が提案するシステムは関連の深いファイル同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．

## 3. テキスト情報とアクセス情報に関する検証

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．

### 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．各テキストから，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 (txt，html，doc，ppt，xls，pdf，tex 等) である．なお，キーワード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，ユーザ自身が，関連があると自覚していないファイル同士から共通のキーワード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，テキスト解析で得られる結果は一意的であり，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

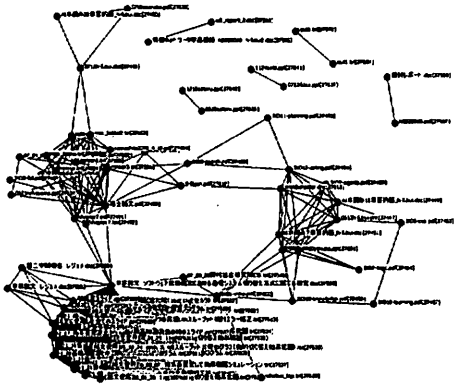

図 1 キーワード共起によるファイルネットワーク
Fig. 1 The file-network made by keyword co-occurrence.

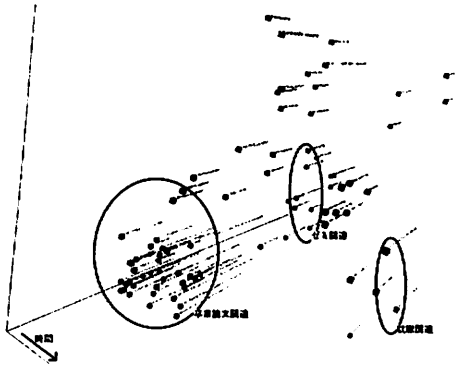

図 2 figrefp1 に時間軸を加えた 3 次元図
Fig. 2 The three-dimensional chart where time axis was added to Fig.1.

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ことが明らかになった．

## 4. ファイル相関ネットワーク視覚化システムの提案と実装

ここでは，提案手法と実装のための各ステップについて述べる．本システムの機能は主にファイル相関性の抽出と，ネットワークの描画に分けることができる．

### 4.1 ファイル相関性の抽出

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．そこで，本研究ではあるファイルがアクセスされた時刻の周辺で生じた，他のファイルへのアクセスに着目する (図 3)．アクセスされた二つのファイルについてテキスト解析を行うことで，情報がどの程度類似しているかを判断する．これらの情報を元にして，ファイルアクセスイベントをノード，情報の類似度をリンク強度としたファイルネットワークを作成する．

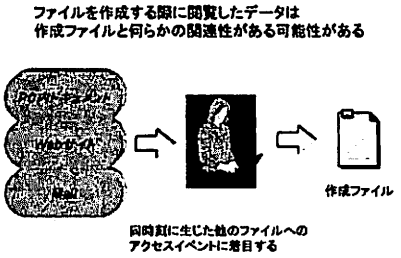


図 3 提案手法の概念
Fig. 3 The concept of proposal.

### 4.2 時系列インターフェースへの適用

作成したファイルネットワークを時系列インターフェースに描画する．時系列インターフェースを用いることで，ファイルが持つ情報の遷移が明らかになる．例えば，あるファイルに着目したとき，そのファイルが保有する情報が，他のファイルに分散していく様子が時系列インターフェースから取得できる．本システムが可能とする情報遷移の監視は，メンバ毎の作業フローの確認や，機密情報管理を行う上で，有益に働くと期待できる．

### 4.3 システムの実装

本システムの構成を図 4 に示す．ファイルネットワークを作成する際には，各ユーザのファイルアクセス情報と，アクセスされたファイルのテキスト情報が必要となる．そこで本システムでは現在，クライアント・サーバ型のネットワーク環境を構築し，必要な情報をサーバで集中管理している．

本システムでは図 4 のように，解析済みのデータをデータベースにて保存し，インターフェースの描画時に必要に応じて読み込んでいる．データベースはファイルアクセス情報を保有する AccessTimeTable，各ファイルの重要語を保有する KeyWordTable，リンク情報を保有する LinkTable の三つのテーブルによって構成されている．後に作成するネットワークにおいて，一つのファイルアクセスは一つのノードに対応する．従って，AccessTimeTable はノード情報を保有するテーブルであると言い換えることもできる．

本研究ではクライアント間の情報の遷移を抽出するため，メールを解析する．メールはメッセージの伝達手段として用いられ，その中にはユーザのスケジュールと関わりの深い情報も多い．また，外部から送られた添付ファイルの第一保管場所としても利用される点においても，情報の発着点として着目すべきポイントとなる．次に，システムを実装する上での具体的な処理について説明する．

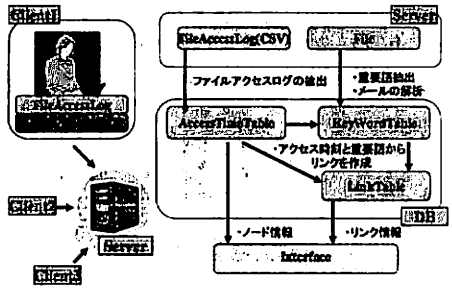


図 4　システム構成

Fig. 4　The system architecture.

### 4.3.1　ファイルアクセスイベントの監視

本システムは ManagementCore [7] を用いて各クライアント PC 内のプロセスを常に監視する．ManagementCore は PC 内のプロセスログを取得し，サーバマシンに送信する IT 資産管理ソフトである．ManagementCore のログから，ファイルアクセスに関する情報 (クライアント PC 名，ファイルアクセス時刻，アクセスファイルパス，アクセスイベント種別など) が抽出され，AccessTimeTable に登録される．また，テキスト解析を行うため，アクセスが生じたファイルのデータもあわせてサーバに送信する．

### 4.3.2　重要語抽出

アクセスが生じたドキュメントファイルに関して，重要語の抽出を行う．Office ファイルや PDF ファイルのような特有のフォーマットファイルに対しては Namazu [8] のフィルタ機能を適用し，プレーンテキストへと変換した上で解析を行った．キーワードを抽出する際には，茶筌 [9] と TermExtract [10] をそれぞれ用いた．茶筌は形態素解析システムである．茶筌によって切り分けられた単語情報を元に，TermExtract では単名詞の連結に基づいた重要語抽出が行われる (詳しくは [10] を参照)．得られた重要語は KeyWordTable に登録する．

アクセスがあったファイルがメールであった場合，メールのヘッダ情報から宛先，差出人，メールタイトル，送信 (受信) 時刻，添付ファイル名などを取得する．メールの本文テキストからは，通常のドキュメントファイルと同様に重要語を抽出し，KeyWordTable に登録する．

### 4.3.3　リンク生成

ここでは提案手法に基づいたファイルアクセス間リンクを生成する．本システムではアクセスの時間差について，パラメータ T を設ける．二つのファイルアクセスの時間差が T 秒以内であれば，二つのアクセスイベントは共起しているとみなす．アクセスイベントが共起している場合，二つのファイルアクセスは共通の目的のために生じた関連アクセスであると仮定できる．ファイルアクセスが生じたとき，その時刻から T 秒前までの他のファイルアクセスイベントとのリンク強度を算出する．時刻 $t_A$ にファイル a に対してアクセスイベント A が生じ，時刻 $t_B$ にファイル b に対してアクセスイベント B が生じた場合，リンク強度の計算方法は以下の通りである．

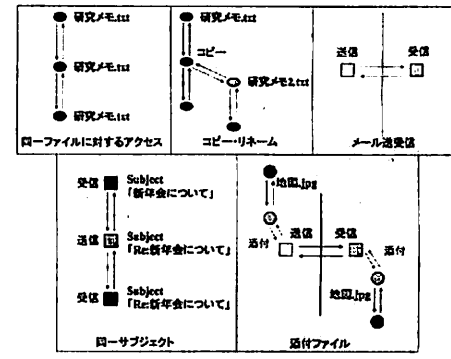


図 5　リンク生成の例

Fig. 5　Examples of link generation.

$$L_{A\to B} = \frac{KeyWord_a(t_A) \cap KeyWord_b(t_B)}{KeyWord_a(t_A)} \qquad (1)$$

$L_{A\to B}$ はファイルアクセス B に対してファイルアクセス A が持つ関係値である．$KeyWord_a(t_A)$ は時刻 $t_A$ の時点でファイル a が持つ重要語数となる．$KeyWord_a(t_A) \cap KeyWord_b(t_B)$ は，時刻 $t_A$ におけるファイル a と時刻 $t_B$ におけるファイル b に，共通して見られる重要語の数である．

ここで述べた，テキスト情報とアクセス情報の共起によるリンク生成以外にも，アクセスイベント同士に関連があると判断できる場合には，強度 1 のリンクを張ることがある．以下にあげたようなケースでは，互いのアクセスイベントには関連があるものとみなす．

**同一ファイルに対するアクセス**　同一ファイルに対する前後のアクセスに対してリンクを張る．

**コピー，リネーム**　コピー，もしくはリネームイベントと転送先ファイルアクセスに対してリンクを張る．転送先ファイルへのアクセスが複数ある場合は，イベントが生じた時刻以降の直近のアクセスとリンクを張る．

**メール送受信**　送信したメールと受信したメールが同一であれば，送信イベントと受信イベントにリンクを張る．このリンクはクライアント間で結ばれるリンクとなる．

**同一サブジェクト**　送信，もしくは受信メールのサブジェクトが同一であれば，前後のメールイベントにリンクを張る．

**添付ファイル**　送信，受信したメールに添付ファイルがあった場合，メールイベントと添付ファイルへのアクセスに対してリンクを張る．

**・送信メール**　イベントが生じた時刻以前の直近の添付ファイルへのアクセスとリンクを張る．

**・受信メール**　イベントが生じた時刻以降の直近の添付ファイルへのアクセスとリンクを張る．

各項目についてリンク生成の例を図 5 に示した．生成されたリンクは LinkTable に登録される．

### 4.3.4　インターフェースにおける描画手法

AccessTimeTable のノード情報と LinkTable のリンク情報

表 1　パラメータと実験環境
Table 1　Parameters and experimental environment.

| | |
|---|---|
| アクセス共起の時間幅 T | 3600 |
| 最小関係値 S | 0.001 |
| クライアント数 | 3 |
| 検証期間 | 2ヶ月 |

に基づいてネットワークを作成する．インターフェースでは一つのファイルアクセスがトリガとして与えられると，このファイルアクセスを第 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．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 S 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，最大ものを採用する．

本システムでは，クライアント毎 (PC 毎) に描画エリアを作成する．ファイルアクセスを示すノードはアクセスが生じたクライアント PC 毎に配置する．このとき X 座標は，できるだけノードが密着しないように与える(注1)．

インターフェースにおける縦軸は時間軸を表すため，ノードの Y 座標はファイルアクセスの時刻によって決定する．現状では下部に示されたノードほど時刻が新しいファイルアクセスを示している．

## 5. 動作実験

本システムを実際に適用した．パラメータと実験環境(注2)を表 1 に示す．

### 5.1 結果 1:グループ内の情報遷移

本システムにおいて，実際にファイルネットワークを描画した際のスクリーンショットを図 6 に示す．画面左側の選択リストから第 1 ノードとなるアクセスイベントを指定すると，右側の描画ウィンドウにファイルネットワークが描画される．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．縦軸は時間を示しており，新しいアクセスが生じた場合，ノードは画面の下部に追加される．丸いアイコンで示されたノードはファイルへのアクセスを，四角いアイコンで示されたノードはメールの送受信イベントを示している．

図 6 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．このように，ファイル同士の情報の引継ぎや，クライアント間での情報伝達といった情報の遷移を時系列インターフェースから容易に把握できる．

### 5.2 結果 2:メンバの PC 内活動

図 7 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，ファイルネットワークから，各メンバの PC 内における活動を確認することができる．

## 6. まとめと今後の課題

本研究では複数のクライアントマシンについてのファイルネットワークを作成し，それを時系列インターフェースに適用することで，組織内の情報の流れや，メンバの活動などを容易に把握できるシステムを実装した．今後の予定として，クライアント数やデータ数を増やし，より詳細な検討を行っていく．また，インターフェースに関しては，キーワード検索などの既存のアルゴリズムと融合させることで，ユーザが求める情報を素早く描画する手法などを開発し，操作性の向上を目指す．

本研究を通して課題点も見つかった．ファイルアクセスログという人間の行動に基づいた情報とテキスト情報を融合する点が，本研究における提案手法の特徴の一つとして挙げられるが，現在の PC のアクセスログの中にはユーザ自身が自覚していないものも存在する．例えば，今回の解析では，ウィルス対策ソフトによるファイルアクセスは，ユーザの意識外でのアクセスとして除外している．しかしながら，こういった機械的なファイルアクセスはソフトウェアに依存しているため，全てのプロセスをあらかじめ指定し，除外するのは難しい．本システムが更に高いパフォーマンスを得るためには，現状の PC アクセスログから人のアクセスログを抽出するステップが事前に必要であると考える．

**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．

(注1)：X 座標の間隔が最も大きいノード間の中点に配置していく．ある描画エリアに n 個目のノードを描画する場合，そのノードの X 座標は下の式により求める．

$$Width(\frac{2n+1}{2^a} - 1) \qquad (3)$$

(ただし $a$ は $2^{a-1} \leq n < 2^a$ を満たす整数)

$Width$ は描画エリアの横幅を示している．仮に，$Width = 400$ であった場合，ノードの X 座標は一つ目のノードから (200, 100, 300, 50, 150, 250, 350, 25,... ) となる．

(注2)：解析対象のクライアント PC の内の 2 台は学生所有の研究用デスクトップ型 PC であり，残りの 1 台は教員所有のモバイル型 PC である．

**文　献**

[1] : Google Desktop, http://desktop.google.com/en/.
[2] Cutrell, E., Robbins, D., Dumais, S. and Sarin, R.: Fast,

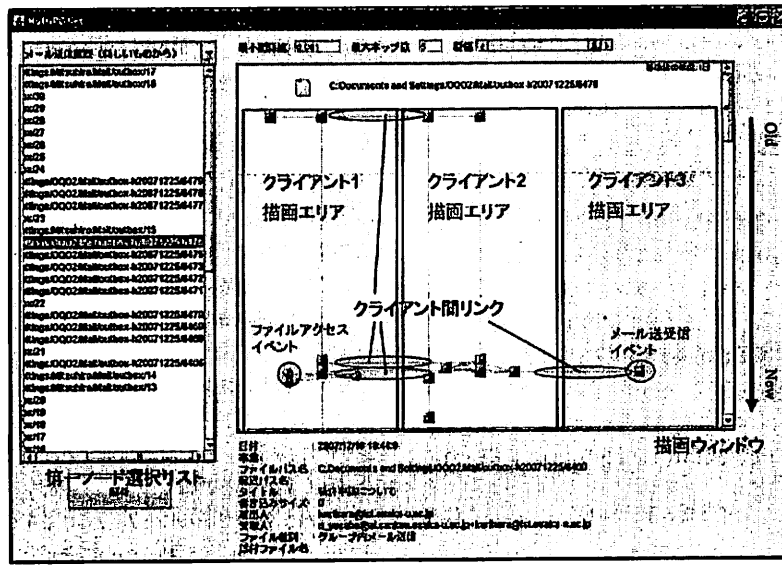


図 6 システムインターフェース 1

Fig. 6 System interface 1.

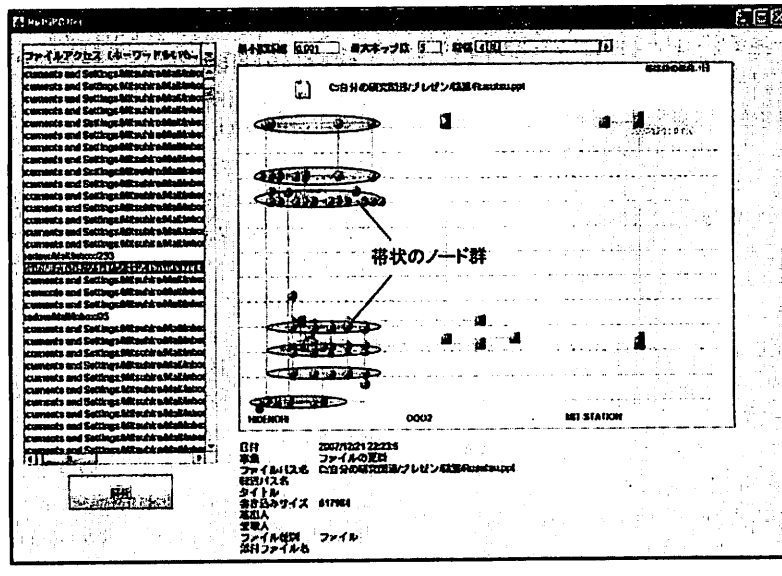


図 7 システムインターフェース 2

Fig. 7 System interface 2.

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).

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] : IT 資産管理システム ManagementCore, http://mcore.jp/.

[8] : 全文検索システム Namazu, http://www.namazu.org/index.html.ja.

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

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