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

2007－DBS－143（6）
2007／7／2

ユーザのスケジュール情報とPC内データ間の相関関係を利用したデスクトップ検索システムの開発

松原 靖子† 小林 一郎†

† お茶の水女子大学 大学院 人間文化創成科学研究科 理学専攻 小林研究室 〒112-8610 東京都文京区大塚 2-1-1
E-mail: †{yasuko-m,koba}@koba.is.ocha.ac.jp

あらまし コンピュータ内の蓄積データの量は，年々増加している．従来のデスクトップ環境は限界を迎えつつあり，新しいデータ管理手法の導入が強く求められている．本研究では，次世代デスクトップ検索の一手法として，コンピュータ内データとスケジュール帳に記入されたユーザの行動間の相関関係を用いたデスクトップ検索システムの開発を目指す．本システムの利用により，ユーザの記憶を利用した新たなデータ検索を実現できる．
キーワード 情報検索，デスクトップ検索システム

Development of a desktop search system using correlation between user's schedule and data in a computer

Yasuko MATSUBARA† and Ichiro KOBAYASHI†

† Graduate School of Humanities and Sciences, Advanced Sciences,, Ochanomizu University 2-1-1 Ohtsuka, Bunkyo-ku, Tokyo 112-8610 Japan
E-mail: †{yasuko-m,koba}@koba.is.ocha.ac.jp

Abstract Recently, as the development of computers, various kinds of tasks have become to be done by computers. In this situation, the amount of data stored in a computer has been increasing, and therefore it has emerged a new problematic issue that necessary data are often missing in a computer. This means that the conventional desktop information retrieval technology has reached to the limit and a new desktop information retrieval is required. In this paper, we propose a desktop search system that uses correlation between user's action history including his/her schedule and data in a computer. By using our system, we can retrieve information in a computer based on user's memory for his/her activity.
Key words information retrieval

1. 研究背景と目的

ソフトウェア技術の発展に伴い，多種多様の作業がコンピュータ上で行われるようになった．このような状況を受けて，コンピュータ内蓄積データの量は急激に増加し，重要なデータが後で取り出せなくなるという新たな問題が生じつつある．従来のデスクトップ環境は限界を迎えつつあり，新たなデータ管理手法の考案・導入が強く求められている．

現在の主なデータ参照手法は，保存場所からのファイル直接参照，あるいはファイル名・ファイルの種類等情報をクエリとした，ファイル検索の2つである．これらのデータ参照は，ファイルが所持する基本情報 (保存場所，ファイル名，作成日時情報等) のみを基に検索処理が行われている．

しかし，この基本情報は，ファイルに関するごく基本的な周辺情報しか保有しないため，大量のデータを検索する際利用する資源としては不十分である．このため，従来型のデータ参照法は，今後の更なるデータ増加には対応しきれないといえる．

これらの問題点を踏まえ，現在では新たなデータ参照方法の研究が幅広く行われている．中でも，各データ間の相関関係を利用したデータ検索は，様々な形で取り組まれているテーマのひとつである．

本研究では，これらの次世代型データ検索の一手法として，スケジュール帳に記入されたユーザの行動データと，コンピュータ内に蓄積されているファイルの間の時間的相関関係を用いた，新たなデスクトップ検索システムの開発を目指す．

本検索システムでは，従来の検索には無い要素である「ユーザの記憶」を用いての検索を行うことが可能となり，検索時のユーザの負担軽減が期待できる．

2. 相関関係を利用したデータ検索手法

本システムでは，各データ間の相関関係を利用して，検索処理を行っていく．

具体的には，「去年の旅行で撮影した写真」「このファイルを作成したときに閲覧していたWebサイト」のように，ユーザの行動とコンピュータ内データとの関係を辿りながら，必要なデータを探していく．

図1は本システムのデータ検索方法のイメージ図である．

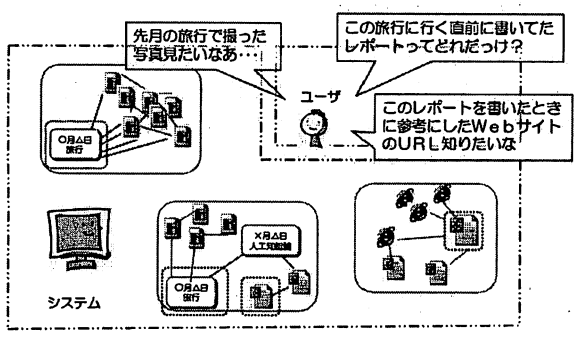

図1　検索システム操作イメージ

実際のシステムの動作の流れは，まずユーザによって，検索のきっかけになるような中心事象 (去年の旅行，今開いているファイル等) をひとつ指定してもらう．システムは，その中心事象に関する周辺情報を，クエリとして受け取り，そのクエリを元に，その中心事象と相関関係のあるデータを自動で収集し，ユーザへ提示するというものである．

本システムにおける各データ間の相関関係抽出には，3つの要素「時間的関係」「内容関係」「キーワード類似関係」を用いる．この3つの要素を利用することにより，システムは，各データ・行動の間の具体的な関係性について検証し，ユーザの要求するデータ群を発見することができる．

3. システムの構成

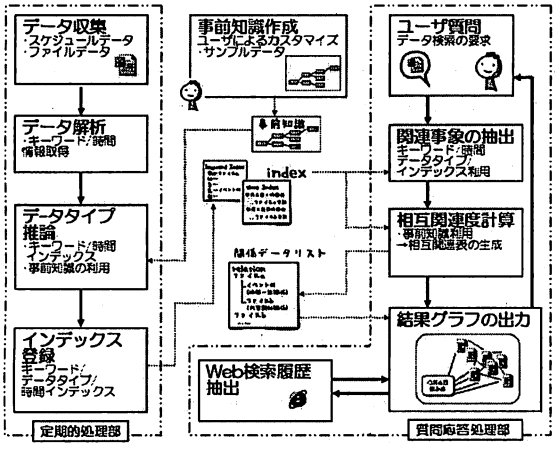

図2　システム全体の流れ

ここで，本システムにおける具体的な処理の流れについての説明を行う．

本システムは，各データの収集・解析を行い，インデックスを生成する「定期的処理部」(図2左)，検索時に実行される「ユーザ質問応答部」(図2右) の大きく2つの処理部に分かれる．

- 定期的処理部

定期的処理部では，各スケジュールデータ，ファイルデータにおける周辺情報を収集・解析する．このとき解析された各種情報は，インデックス (図2中央) に登録され，検索時に使用される．ただし，この処理は現時点ではスケジュールデータ・ファイルデータのみに対し行われ，Web閲覧履歴の厳密な解析およびインデックスへの登録は行わない．

- ユーザ質問応答部

ユーザ質問応答部では，ユーザからのクエリを受け取り，その情報を元に，インデックスを利用しながら，関連するデータ群を探し出す．探し出されたデータは，関連の強さを算出され，可視化されてユーザへ提示される．Web閲覧履歴については，別処理として，「この日に閲覧したWebサイトをみたい」という要求を受けたときのみ実行されるWeb検索履歴抽出処理機能 (図2中央下) を持つ．

さらに，本研究では，各ユーザの選好度に合わせた相関関係を抽出するための工夫として，ユーザの行動や生活に合わせて設定することができる事前知識を導入した．

「事前知識カスタマイズ部」(図2中央上) は，これらの知識データを準備するための事前処理部である．

事前知識カスタマイズ部では，相関関係を抽出する際に用いるパターン知識等作成処理を行う．具体的には，イベントデータのテンプレート作成，データ間関連確率パターン作成の2つの処理を行う．

イベントデータテンプレートとは，場所，時間，キーワード等の情報を元に，そのスケジュールがどのような性質のイベントであるかを自動判定するための枠組みのことを指す．データ間関連確率パターンとは，どのような性質をもつデータ同士が関連性を持ちやすいかを示す確率である．

ユーザは，これらの知識を自分の行動や生活に応じて事前に設定する．そして，ユーザによってカスタマイズされたこれらの知識は，定期的処理部及びユーザ質問応答部処理において利用される．

次に，本システムの大きな2つの処理の詳細について，処理の順に沿って説明を行う．

4. 定期的処理部

定期的処理部において重要な要素は，どのようなデータを扱うか，そしてそれらのデータをどのような形で管理・利用するかの点である．

4.1 データの収集

本システムでは，コンピュータ上で扱われる様々なデータに対し検索を行えるようにした．具体的には，以下の 3 種類のデータを扱った．

- ファイル利用履歴

ファイル名，保存場所，更新日時の 3 項目について，定期的に情報を収集する．

- ユーザのスケジュール

ユーザが定期的に記入しているスケジュールデータを解析し，題名，日時，場所，詳細メモを抽出・収集する．ユーザのスケジュールデータ収集には Google Calendar を利用する．

- Web 検索履歴

検索キーワード記入履歴，及び Web 閲覧履歴を解析し，ユーザへ提示する．ただし現時点では，定期的処理部において，高度な情報解析 (文書解析等) は行わない．Google Search History を利用する．

現時点では，これらの 3 種類のデータのみを管理対象としているが，他にもコンピュータ内に保管されているメール等のデータについても，同様の処理を行うことによって検索対象とすることが可能であると考える．

4.2 イベントの種類判定

本研究では，ユーザの行動を検索に取り入れるための工夫として，イベントの種類情報を利用した．ここでのイベントの種類とは，旅行や授業，会議，ゼミ等のことを指す．これらの情報を利用することにより，イベントの種類それぞれに対応した相関関係抽出処理が可能となり，より精度の高い検索が行えるものと考える．

イベントの種類情報は，本システムにおいて最も重要な要素のひとつである．しかし，カレンダー上に記入されたスケジュールデータに蓄積されているのは，時間，タイトル等の情報のみであり，イベントの種類に関する情報はない．そこで，本システムでは，蓄積されたスケジュールデータを元に，イベントの種類を自動で判断する処理を導入した．

この自動判定処理において，先述の事前知識のひとつ，イベントデータテンプレートを利用する．

本研究では，事前にユーザから与えられたイベントテンプレート情報をシステムに投入し，イベントの種類が何であるかを自動で判定する．以下は，実際の比較に用いた要素である．

時間要素	キーワード要素
日時	タイトル
所要時間	イベントの場所
時間帯	詳細メモ

図 3 は，イベント種類判定の様子を示す．システムは，スケジュールデータを解析する際，事前に与えられたテンプレートとの比較を行い，類似度（「類似度スコア」と呼ぶ）を算出する．そして最終的に，類似度スコアが最も高かったものを，実際のイベントの種類であると判定する．

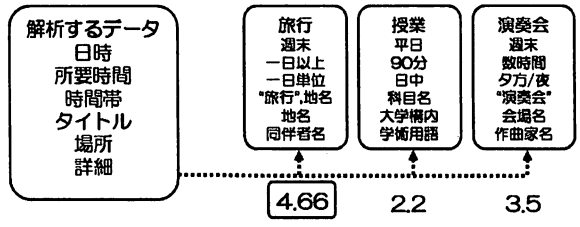

図 3 類似度スコア算出とイベント種類判定

4.3 インデックスの作成

システムによって定期的に収集されたデータは，順次解析され，重要事項を抜き出され，最終的にインデックスに登録される．これらのインデックスは，ユーザ質問応答部処理における関連データ抽出処理に利用する．

本研究では，time インデックス，data-type インデックス，keyword インデックスの 3 種類を作成した．

- time インデックス

各データの時間的情報を蓄積したインデックスである．time インデックスにより，時間的に相関関係のあるデータ群の発掘が実現できる．内容的に関係が無くても利用時期が一致していたファイル等を抽出できる．

- data-type インデックス

これはデータの種類別に蓄積するインデックスである．ファイルデータの場合には拡張子の種類，スケジュールデータの場合には，イベントの種類 (旅行，授業，飲み会，等) ごとに登録する．これにより，類似データの抽出が可能となる．

- keyword インデックス

ファイル名や，スケジュールデータ内に含まれるキーワードを抽出し，蓄積する．

5. ユーザ質問応答部

ユーザ質問応答部での処理は，ユーザからの質問クエリを受け付ける質問処理，関連データ抽出計算を行う関係データ抽出処理，ユーザへ検索結果を提示する結果表示処理の 3 つから成る．さらに，ユーザが Web 閲覧履歴提示を要求した場合には，別処理として，Web 閲覧履歴検索処理が行われる．以下では，これらについて，各項目ごとに説明を行う．

5.1 ユーザ質問処理

本検索システムでは，すべての検索処理を，データ間の「関係」を利用しながら行っていく．

以下において，検索の処理の流れについて説明を行う．

まず初めにシステムは，データ参照のきっかけとなるようなクエリをユーザから受け取る．本システムにおけるクエリは，以下の 2 項目である．

- 検索の中心事象
  データを探すためのきっかけとなるような事象．
- 抽出したい関係の種類
  時間的関係，内容的関係，キーワード類似関係の 3 種類から指定．

本システムでは，すべての検索を「関係」によって行うため，検索されるデータは「検索の中心となるデータ」と「中心データと関連をもつデータ」に分かれる．検索をする際には，この「中心データ」をユーザによって指定してもらう必要がある．

さらに上記に示したように，本システムで扱う「関係」には 3 つの種類があり，それにより抽出対象となるデータも変化するため，これについてもユーザによって指定してもらう．

次に，システムは入力されたクエリ（検索中心データ情報）を元に関係するデータ群を抽出しユーザへ提示する．ユーザは提示されたデータ群を閲覧しながら，そこに提示されているデータの中から新たに再検索を行っていく．

ここでさらに具体例として「福島へ旅行へ行ったときの写真」を閲覧する際の検索の流れを示す．

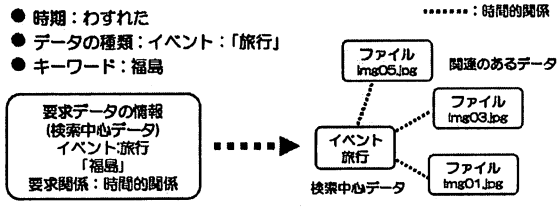

図 4　ユーザ質問のイメージ

ユーザが今，この旅行の時の写真（複数個）を閲覧したいとする．旅行に行った厳密な日時はわからないが，行き先である「福島」という地名のキーワード情報は判明していると仮定する．

撮影された写真は，旅行中に撮影したものであると自然に想像されることから，旅行イベントと時間的な関係を持つといえる．そこで，図 4 左下のような，中心事象の情報と要求関係をシステムへ与える．

システムは，与えられた内容にあわせて解析処理を行い，最終的に図 4 右のように，データ間の関係を可視化した状態でユーザへ提示する．ユーザは，提示されたデータの中から，閲覧したい写真を取り出すことができる．

検索中心データの指定には，ファイル名等，確実な手がかりとなる情報を指定することももちろん可能であるが，一方で，ユーザの曖昧な記憶をクエリとして提示することもできる．

例えば，曖昧な時期指定 (先月，先週等) や，イベントの種類情報 (旅行，授業等) などがクエリとして入力できる．

このような曖昧なデータの関連データを抽出する場合には，システムは，ダミー事象というものを一時的に作り出し，そのダミー事象を中心事象とおいて，相関関係のあるデータを抽出する．

このダミー事象生成処理を利用すれば，ユーザの記憶が曖昧で，誤りを含むクエリで検索をしてしまったとしても，(例えば，旅行に行ったのが，先月ではなく，先々月であったとしても) そのダミー事象と類似したデータとして，実在の事象を抽出することができる．

図 5 は，曖昧な検索クエリからデータ検索を行う際の処理のイメージである．

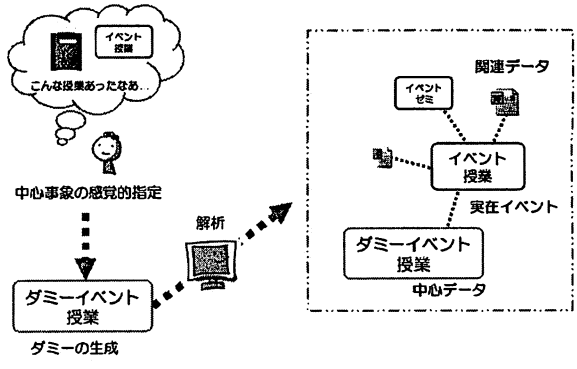

図 5　記憶にたよった曖昧検索のイメージ

5.2　相関関係データ抽出処理

ユーザ質問応答部処理において最も重要な課題は，各データの関係事象をどのように抽出していくかのということである．

本システムでは，時間的 (time) 関係，内容的 (data-type) 関係，出現単語類似 (keyword) 関係の 3 種類の関係性の観点から，情報を抽出する．

これらの関連性については，例えば，旅行イベントとそのとき撮影した写真ファイルは時間的関連性を持ち，定期的に行われる各ゼミイベントについては，内容が関連していると判断することができる．システムは，これらの関連性を手がかりとして，様々な関連データを同時に発見することができる．

関連の度合いについては，スコア算出処理を行い，関連の強さを数値で表す．このスコアが高く算出されたデータ間には，強い関連性があると判断される．複数の関連データのスコアを計算した後，強い関連性を持つと判断されたデータが，検索結果としてユーザへ提示される．

各関連性の発見には，上述した 3 種類のインデックスを用い

る．以下で，各データに関する関連度の具体的なスコア算出法について述べる．

- 時間的 (time) 関連

時間的関連スコアは，標準正規分布の式に当てはめ，時間が完全に一致したデータ同士は関連度が高く，逆に時間的に離れたデータ同士ほど低い関連度を算出するように設定した．

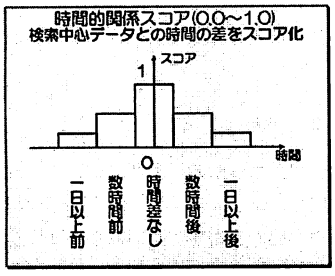

図 6　時間的関係性スコア計算グラフ

- 内容的 (data-type) 関連

内容的関連スコアは，各データ間の類似度をスコアとして算出した．ファイルデータの場合には，拡張子が一致したものは高スコアとし，まったく別の種類のデータはスコアを 0 とした．イベントデータの場合には，事前に判定されたイベントの種類を比較し，さらに，互いの性質 (時間的類似度，記述内容の類似度等) を比較し，類似度スコアを算出した．

- 出現単語類似 (keyword) 関連

単語類似関連については，データ間の単語一致数をスコアとして算出した．比較するデータとの共通単語が多ければ多いほど，単語類似関係スコアが高くなる．

以上の 3 種類のスコア算出を行い，最終的にスコアの高いものをユーザへ提示するのであるが，本研究ではさらにここでスコア補正処理を行い，より正確なデータ抽出を目指した．

スコア補正処理には，あらかじめカスタマイズ処理によって作成したデータ間関連確率パターン情報を用いる．このパターン情報は，あるデータが他のどのようなデータと強い関連性を持つかという情報である．

図 7 は関連確率パターンの例である．

		旅行	人と会	演奏	授業	ゼミ	講演	締め	jpg	txt	doc	pdf	ppt
		1	1	2	12	3	2	2	10	7	4	5	6
旅行	1	0	0	0	0	0	0	0	1	0	0	0	0
人と会	1	0	0	0	0	0	0	0	0	0	0	0	0
演奏	2	0	0	0	0	0	0	0	0.5	0	0	0	0
授業	12	0	0	0	1	0	0	0.08	0	0.33	0.08	0.25	0
ゼミ	3	0	0	0	0	1	0	0	0	0	0	0	0.67
講演	2	0	0	0	0	0	1	0	0	0	0	0	0
締め	2	0	0	0	1	0	0	0	0	0.5	1	0.5	0
jpg	10	8	0	2	0	0	0	0	1	0	0	0	0
txt	7	0	0	0	4	0	0	1	0	0.29	0.14	0	0
doc	4	0	0	0	2	0	0	4	0	2	1	0.5	0
pdf	5	0	0	0	5	0	0	2	0	0	2	1	0
ppt	6	0	0	0	0	6	0	0	0	0	0	0	1

図 7　データ間関連確率パターン表

この表から，イベント「授業」に関係する可能性が高いファイルが doc，pdf ファイルであること等がわかる．

このスコア補正処理を行うことにより，より頑健な関連データ抽出が実現できる．

5.3　各データ間相関関係の可視化表示

本システムでは，データ間の相関関係をグラフ構造によって可視化し，ユーザへ提示する．

各データをアイコンで表示し，それらと関係しているデータが線で結ばれる．強い関係であればあるほど，太い線で結ばれ，互いに引き寄せられる．

この可視化処理によって，システムは時間的な関係と内容の関係をそれぞれの相関関係を扱いやすい形で提示し，ユーザはその検索結果グラフから，関係する複数個のデータを視覚的に発見することができる．アイコンをたどって検索を繰り返すことで，データ間を渡り歩いたネットサーフィンのような感覚のデータ抽出・参照も行うことが可能となる．

5.4　Web 閲覧履歴検索処理

本システムでは，Web 閲覧履歴の検索機能も持つ．現時点では，複雑な文書解析等処理は行わず，スケジュールデータやファイルデータのように，インデックスへの登録も行っていない．Web 閲覧履歴検索処理では時間的関係についてのみ扱っている．この時間的関係とは，具体的には「この日に閲覧した Web サイト」といったような，特定の日の履歴を抽出してきてユーザへ提示するというものである．

図 8 は，Web 閲覧履歴検索処理のイメージである．

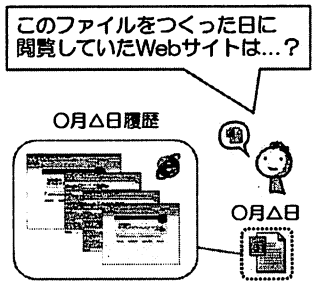

図 8　Web 閲覧履歴検索処理

この例では，ある特定のファイルを作成した日に，ユーザがどんなサイトを閲覧していたかを提示している．この Web 閲覧履歴検索を行えば，具体的な日時がわからなくても，曖昧な記憶のみ (ファイルを作った日，ゼミがあった日等) で，Web 閲覧履歴を取り出すことができる．

6.　システム実行例

本システムを用いて実際の検索を行ったときの動作について考察する．ここでは，「旅行で撮影した写真」の参照を例に挙げる．図 9，図 10 は，検索クエリ記入画面である．

まず，図 9 で，検索中心データの手がかり (イベント：旅行) 情報を記入する．次に，図 10 で，探したい関係の種類，関係の強さ等情報を記入する．

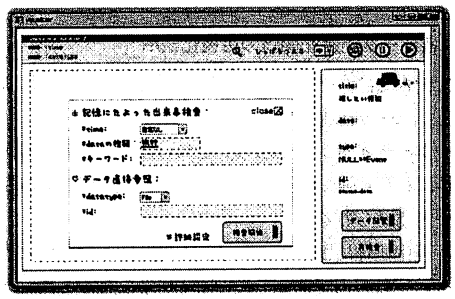

図 9　クエリ記入例:1(中心事象指定)

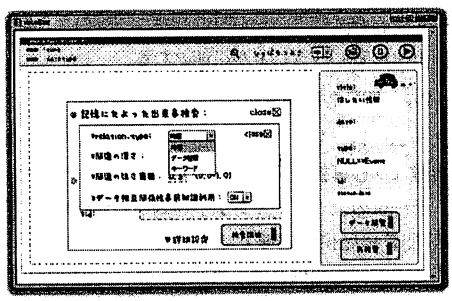

図 10　クエリ記入例:2(関係の種類指定)

以上のようなクエリを記入すると，図 11 のような検索結果が提示される．

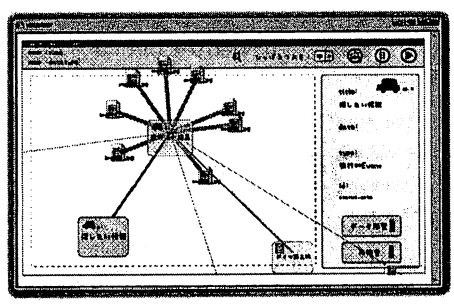

図 11　システムによる結果提示例

図 11 によって，旅行イベントと画像ファイルの間に強い関連性があり，互いが引き寄せられている状態が確認できる．

さらにここでシステムに対し，「この日に閲覧した Web サイトをみたい」という要求を出すと，システムは，図 12 のような検索結果を提示する．

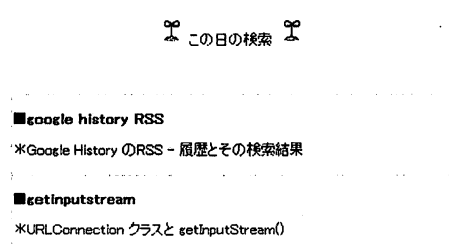

図 12　Web 検索履歴結果提示例

この Web 検索履歴結果表示画面では，その日に行った検索のクエリキーワード，そして閲覧した Web サイトの URL を提示する．

ユーザは，このような一連の検索操作を通して，関連する様々な情報を抽出することができる．

本システムを利用しながら，ファイルを探したり，予定を確認したり，そしてさらなる検索を行いながら，必要なデータをストレスを感じることなく取り出すことができる．

7. ま と め

本研究において，ユーザの行動とコンピュータ内データ間の相関関係を利用した新しいデータ検索手法を提案した．さらに，これらのデータ参照を実現するための手法として，複数のインデックス，イベントの種類の自動判定，事前知識による関連度指数の補正処理を導入した．

本研究では，ある程度満足のできるデータ検索が実現できたことが確認できたが，イベントの種類が増えた場合や，扱うデータの種類が増加した場合に対応できるか等の問題については未解決である．本システムを実際に使用する際の使い易さについても，考慮すべき重要な問題といえる．

さらに，現時点では Web 閲覧履歴データに対して，他のスケジュールおよびファイルデータのような，事前のデータ解析を行っていない．しかし Web 履歴についても，内容解析等処理を行えば，さらに柔軟な情報検索を行うことが可能になると考えている．

今後は，これらの課題を解決すると同時に，関連事象抽出処理の更なる精度向上や，ユーザが直感的に使用できるデータ提示インタフェースの改良等についても行っていく予定である．

文　献

[1] 超整理法, 野口悠紀夫, 中公新書 (1993)
[2] 形態素解析システム茶筌, 松本研究室, http://chasen.naist.jp/hiki/ChaSen/
[3] オントロジー工学, 溝口理一郎, 人工知能学会 (2005)
[4] オントロジーエディタ HOZO,(溝口研究室) http://www.hozo.jp/
[5] Google Calendar http://www.google.com/calendar/
[6] Google Search History http://www.google.com/searchhistory/
[7] Aduna AutoFocus http://www.aduna-software.com/products/autofocus/
[8] Blog Keyword Visualizer http://bkv.so-net.ne.jp/
[9] 俺デスク　大澤 亮 http://oredesk.net/
[10] "Time-Machine Computing: A Time-centric Approach for the Information Environment",Jun Rekimoto
[11] Lifestreams,D.Gelernter,Eric Freeman,Yale University, http://www.cs.yale.edu/homes/freeman/lifestreams.html