1G3-2

時間軸を主体にしたデータ間相関関係抽出とデスクトップ検索への応用

Extraction of Time-based Data Correlation and Its Application for Desktop Information Retrieval

松原靖子*1　　小林一郎*2
Yasuko Matsubara　　Ichiro Kobayashi

*1*2 お茶の水女子大学　理学部　情報科学科
Dept. of Information Sciences, Faculty of Science, Ochanomizu University

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 information retrieval method that uses correlation among data and develop a system based on the method. By using our system, we can retrieve information, based on the correlation between user's action and stored information in a computer.

1. 研究背景と目的

コンピュータの発展に伴い，多種多様な作業がコンピュータ上で行われるようになった．このような状況下において，コンピュータ内に蓄積されるデータの量は急激に増加し，データの埋没という新たな問題が出現している．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発見・導入が強く求められている．

現在主流となっているデータ参照手法は，保存場所からのファイルの直接参照，あるいはファイル名やファイル種類等の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基本的に，ファイルが所持する基本情報 (保存場所，ファイル名，作成日時等情報) のみによって行われている．しかし，この基本情報というものは，情報量が非常に少ないため，大量のデータを管理する際に使用するものとしては適さない．そのため，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，これらの次世代データ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，従来の検索には存在しなかった新しい要素，ユーザの記憶を用いて検索を行う．本システムの利用により，ユーザは従来の検索のような負担を感じることなくデータ参照を行うことができると考えている．

2. システムの概要

図 1 に各部の構成と処理の流れを示す．

本システムは，各データの収集・解析を行う「定期的処理部」(図 1 左)，検索時に実行される「ユーザ質問応答部」(図 1 右) の大きく 2 つの処理部に分けられる．

連絡先: 松原靖子，お茶の水女子大学　理学部　情報科学科
小林研究室，東京都文京区大塚 2-1-1，03-5978-5709，
yasuko_m@koba.is.ocha.ac.jp

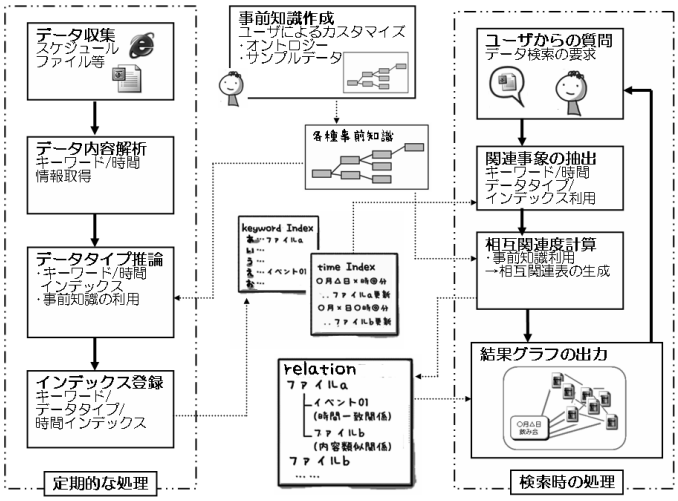

図 1: システム全体の流れ

- 定期的処理部

  定期的処理部では，各データ情報の収集・解析を行う．このとき解析された各種情報は，インデックス (図 1 中央) に登録され，検索時に使用される．

- ユーザ質問応答部

  ユーザ質問応答部では，システムが解析・蓄積したインデックスを用いて，要求されたデータを探し出す．探し出されたデータは，関連の強さを算出され，可視化された状態でユーザへ提示される．

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，事前知識を導入した．そこで，これらの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」(図 1 中央上) も作成した．

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．

3. 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．

以下で，それらについて示す．

3.1 データの収集

本システムでは，以下の 2 種類のデータを管理対象とした．

- ファイル利用履歴

  [ファイル名，保存場所，更新日時]

- ユーザのスケジュール (Google Calendar を使用)

  [題名，日時，場所，詳細メモ]

現時点では，これらの 2 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 Web 履歴やメール等のデータも同様の処理を行うことによって管理することが可能であると考えている．

3.2 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，より柔軟な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．表 1 は，実際の比較に用いた要素である．

表 1: イベント種類判定要素

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

システムは，スケジュールデータを解析する際に，事前に与えられたテンプレートとの比較を行い，類似度スコアというものを算出する．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．

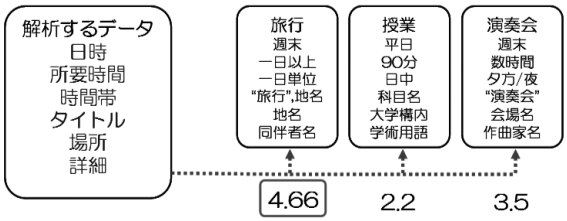

図 2: 類似度スコア算出とイベント種類判定

3.3 各種インデックス

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記入される．これらのインデックスは，ユーザ質問応答部処理において，クエリユーザ要求データ抽出処理及び，関連データ抽出処理に利用する．

本研究では，keyword インデックス，time インデックス，data-type インデックスの 3 種類を用意した．

- keyword インデックス

  ファイル名，スケジュール内に含まれるキーワードを抽出し，蓄積する．本システムではこれらの各単語の意味関係をオントロジー形式で事前に登録することにより，類似した単語の発掘も可能にする．

- time インデックス

  ファイル・スケジュール各データの時間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．

4. ユーザ質問応答部

ユーザ質問応答部内の処理は，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．以下では，これら 3 つについて，順に説明を行う．

4.1 ユーザ質問処理

本システムでは，データ間の「関係」を利用してデータを参照していく．以下では，具体的にどのような作業を行ってデータ参照を行うかについて述べる．

まずはじめに，ユーザはデータ参照のきっかけとなるようなクエリをシステムに提示する．ここでいうクエリとは，検索の

中心となるデータの情報である．イベントならイベントの情報（イベントの時間や内容の情報），ファイルならファイルの情報（ファイル名，ファイルの保存場所）などがクエリとなる．

システムはそのクエリ（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．

図 3 は，データ参照法の流れの一部である．

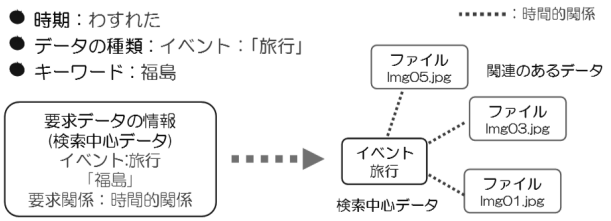

図 3: ユーザ質問のイメージ

ここでは,「福島へ旅行へ行ったときの写真」を閲覧する例を用いる．

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する．

撮影された写真というのは，もちろん旅行中に撮影したものであるから，旅行イベントとの時間的な関係がある．そこで，図 3 左下のような，中心事象の情報と要求関係をシステムへ提示する．

システムは，提示された内容にあわせて解析処理を行い，最終的に図 3 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．

4.2 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，各データの関連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) 関連，内容的 (data-type) 関連，出現単語類似 (keyword) 関連の 3 種類の関連性の観点から，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．

以下に，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，図 4 のような分布になるように設定した．

  これらの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，遠いデータ同士ほど低い関連度を算出するように設定した．

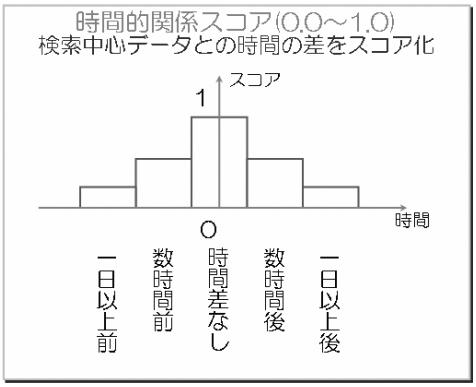

図 4: 時間的関係性スコア計算グラフ

- 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，あるいは性質が似ているもの (文書ファイル同士，画像ファイル同士等) 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のことである．

図 5 はそれら事前知識の一部の例である．この表から，イベント授業に関係する可能性の高いファイルが doc，pdf ファイル等であることがわかる．

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．

これらのスコアは各ユーザによって異なるものであるため，各ユーザがカスタマイズを行うと，これらの数値が自動修正される様な仕組みが必要となる．

本システムでは，1ヶ月間の実際のスケジュールとファイル間の関連パターン (例えば，jpg ファイル計 10 個の中で，8 個

		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 5: 相関関係スコア表

は旅行イベントと関係していた等の情報) をシステムに投入し，図 5 のような関連確率を自動計算する処理を導入した．

4.3 検索結果の可視化表示

ユーザが直感的に利用できるデータ検索システムを実現するためには，結果の表示にも工夫が必要である．

本システムでは，データ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，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参照も行うことが可能となる．

5. システムの実行例

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．まず，ユーザの把握している情報をクエリとして記入する．

図 6 は，本システムの検索クエリ記入画面である．ここでは「旅行イベントと時間的に関係しているデータ」をクエリとして記入した．

図 6: クエリ記入例

図 7 は，図 6 の検索結果である．

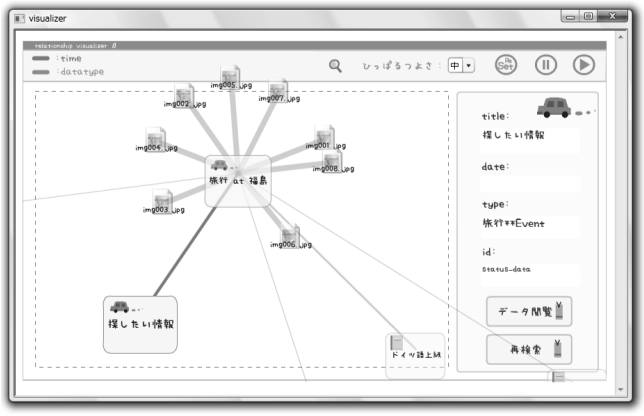

図 7: システムによる結果提示例

図 7 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．

ユーザは，これらの情報を元に，ファイルを開いたり，予定を確認したり，さらなる検索を行いながら，必要なデータを取り出すことができる．

6. 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参照法を提案した．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観察されたが，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．

今後は，関連事象抽出処理の更なる精度の向上と，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を行うつもりである．

参考文献

[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] Lifestreams,D.Gelernter,Eric Freeman,Yale University, http://www.cs.yale.edu/homes/freeman/lifestreams.html