時間軸を主体とした相互関係抽出とデータ管理への応用

松原 靖子　(指導教員：小林 一郎)

1　研究背景と目的

近年，コンピュータの発展に伴い記憶容量の増大による情報爆発という新たな問題が生じるようになった．コンピュータ内部のデータが増え続けた結果，従来から採用されていた，手作業によるフォルダ・ファイルの管理は限界に達している．今後の更なる情報量増大に備えるためにも，新しいデータ管理手法が必要である．

本研究では，次世代データ管理の一つの手法として，各データ間の相関関係を用いたデータ管理法を提案し，その上でこの手法を利用した，新たなデスクトップ上データ検索のシステムの開発を目指す．このデータ検索システムを利用することにより，ユーザは大量に保存されているデータの中から，データ間相関関係を利用して必要な情報を取得することが可能となる．

2　システムの概要

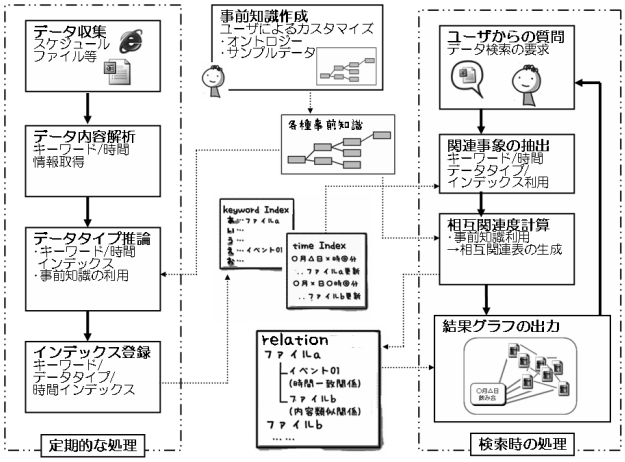

図 1: システム全体の流れ

システムは以下の３部から構成される．

- 定期的処理部
- ユーザ質問応答部
- 事前知識カスタマイズ部

データを定期的に収集・解析する定期的処理部，ユーザの質問によって発動するユーザ質問応答部，そしてデータ相関関係抽出の際に必要となる事前知識を構築するためのカスタマイズ部である．

定期的処理部では，各データの情報を自動で収集・解析し，インデックスに登録する．さらにそれらの解析結果を元に，データ間の相関関係を算出する．このとき解析された各種情報 (インデックス，相関関係情報) はシステム内に蓄えられ，管理される．

ユーザ質問応答部では，システムが解析・蓄積した各種情報を元に要求されたデータを探し出し，検索結果として出力する．

3　システムの構成

以下に，本研究で開発されたシステムの構成各部について考察する．

3.1　データ収集と管理

本システムでは，多種のデータを管理の対象とする．具体的には，以下のようなデータを扱う．

- ファイル利用履歴
- ユーザのスケジュール (google calendar を使用)
- Web 閲覧履歴，メール，メッセンジャー履歴等

システムはユーザに関するこれらのデータを定期的に収集・管理する．

本システムでは，スケジュール情報も管理の対象とする．これにより，各データ・スケジュール間の時間的な相関関係の発見が可能となり，システムはユーザの行動とコンピュータ内データとの関わりを把握できるようになる．

本システムの利用により，ユーザの行動や意思に密接に関わった，新しい形の情報検索が可能となる．

3.2　スケジュールイベントの種類判定

カレンダー上に記述されたイベントのデータは，時間帯情報やタイトル，コメント等によるキーワード情報しか所持していない．

データ間の相互関係を発掘するためには，これらのイベントがどのような性質のデータなのか，つまりイベントのタイプが何であるかを事前に判断する必要がある．

本研究では，各イベントの種類ごとのサンプルを用意し，それらのサンプルと解析対象のイベントとの類似距離を計算することにより，イベントの種類を判定する仕組みを導入した．実際のサンプルが図 2 である．

```
<授業 count="3">
- <time>
    <所要時間>{under90=3}</所要時間>
    <時間帯>{ごぜん=1, ごご=2}</時間帯>
    <曜日>{Tue=2, Fri=1}</曜日>
  </time>
- <keyword>
    <場所>{コンピュータ=1, 階=2, 理学部3号館=1, 3号館=1, 室=1, 共
      通講義棟1号館=1, 共通講義棟=1}</場所>
    <タイトル>{ドイツ語上級=1, 実習=1, 人工知能論=1, プログラミング
      =1}</タイトル>
    <詳細>{ドイツ語=1, 提出=1, 演習=1, 著=1, 毎週=1, レポート=1,
      人工知能論=1, 授業=2, プログラミング=1}</詳細>
  </keyword>
</授業>
```

図 2: イベントサンプル例（授業イベント）

例えばこの例では，時間的特長として，授業イベントは所要時間が９０分であり，火曜金曜など平日に起こりうるイベントであることがわかる．あるいは内容

的特長としては，学問に関するキーワードが多い，などが挙げられる．

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，インデックスに記入される．

これらのインデックスは，相関関係の自動抽出，及び，ユーザ質問応答時のデータ検索に使用される．具体的には以下の 3 つのインデックスである．

- keyword インデックス
- time インデックス
- data-type インデックス

keyword インデックスはファイル名，スケジュール内キーワードなどを解析し，保存する．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 インデックスにより，時系列によるデータ検索，及び相関関係の発掘が可能となる．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 インデックスは，データの種類のインデックスである．ファイルの拡張子の種類の他に，例えばスケジュールデータの種類 (旅行，授業，飲み会，等) を登録する．これにより，検索時に類似データの提示ができる．

3.4 相関関係抽出部分

ここでは主に，データの種類は違うが時間的に相関関係のあるデータ等を自動抽出する．

具体的には「旅行」と「旅行中撮った写真」,「会議」と「会議の資料」などの関係である．

詳細情報が解析・登録された各種インデックス，及び，事前に与えられたヒューリスティックを用いて処理が行われる．

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は，あらかじめユーザが事前知識としてシステムに与えておく．

図 3 はそれら事前知識の一部の例である．この表では，イベント授業に関係する可能性の高いファイルが pdf ファイルであること等がわかる．

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，各ユーザのカスタマイズによって，数値を自動算出する仕組みが必要となる．

		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	0.8	0	0	0	0
人と会	1		0	0	0	0	0	0	0	0	0	0	0
演奏	2			0	0	0	0	0	0.1	0	0	0	0
授業	12				0	0	0	0.04	0	0.19	0.04	0.25	0
ゼミ	3					0	0	0	0	0	0	0	0.67
講演	2						0	0	0	0	0	0	0
締め	2							0	0	0.07	1	0.2	0
jpg	10								1	0	0	0	0
txt	7									0.08	0.07	0	0
doc	4										1	0.2	0
pdf	5											1	0
ppt	6												1

図 3: 相関関係スコア表

3.5 検索結果表示部分

ユーザが直感的に使用できるデータ検索を実現するためには，結果の表示にも工夫が必要となる．

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た状態で結果を提示する．実際の検索結果出力画面を図 4 に示す．

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に線を延ばす．システムが時間的な関係と，内容の関係，それぞれの相関関係を提示し，ユーザはアイコンをクリックしていくことにより，データを参照していく．

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掘することができる．

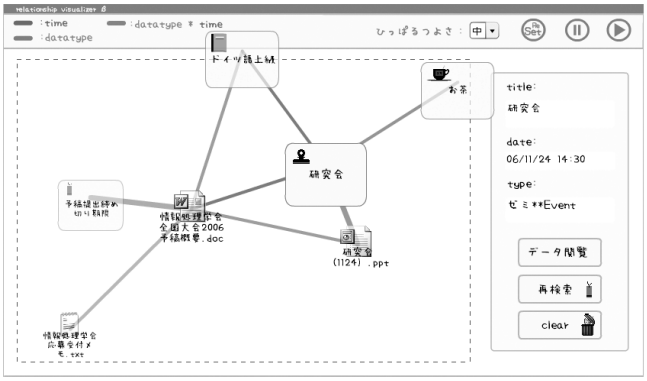

図 4: 出力結果画面

4 まとめ

本研究において，各データ間の関係性を元に新たなデータを参照するという手法を提案した．本システムを利用することにより，ユーザはネットサーフィンをするような感覚でデータを渡り歩くことが可能となる．

視覚的にデータを参照できるということは非常に操作性に優れている反面，一方で相関関係の的確な抽出を実現しなくては本システムは効力を発揮しない．

更なる精度の向上が，本システムの本格的な実用化に向けた課題であるといえる．

参考文献

[1] 超整理法, 野口悠紀夫, 中公新書 2006.

[2] オントロジー工学, 溝口理一郎, 人工知能学会,2005.

[3] ベイジアンネットワーク技術, 木村陽一 岩崎弘利,2006

[4] 形態素解析システム茶筌 Chasen(松本研究室), http://chasen.naist.jp/hiki/ChaSen/

[5] 法造：オントロジーエディタ HOZO(溝口研究室), http://chasen.naist.jp/hiki/ChaSen/

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