情 報 学 基 礎 39−11
（１９９５．９．１４）

研究者のための新しい情報分析ツール『ThinkView』

永田 陽一　　福本 一夫　　田中 勝

富士総合研究所　技術開発センター

本稿では、「ワードマップ」・「テーマエディタ」・「文献分布図」のスパイラル利用により、対象分野に対する利用者独自の体系を生成し継続的に利用することを支援するツール『ThinkView』を提案し、その評価テストの結果などからツールの問題点、可能性等について述べる。

A New Information Analysis Tool for Researchers "ThinkView"

Yoichi Nagata, Kazuo Fukumoto, and Masaru Tanaka

Information Technology Center, Fuji Research Institute Corporation
3-22-51 Nishikasai, Edogawa-ku, Tokyo 134, Japan

E-mail: {naga,kazuo,tanaka}@pier.fuji-ric.co.jp

In this paper, we propose a new information analysis tool named "ThinkView", based on the spiral use of three main functions; "Word Map", "Theme Editor", and "Distribution Map". "ThinkView" assists a user to create his or her own conceptual scheme of analysis of his or her domain of the work continually.
Through the experiments, we found the positive effects and the possibilities of the system as well as the problems we have to resolve for the better usage.

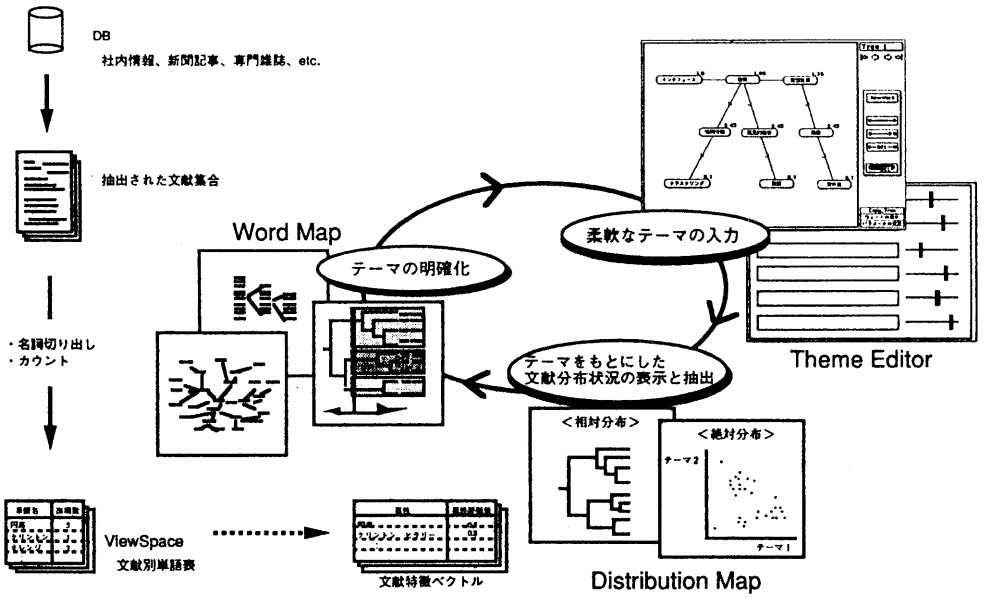


図１：システムイメージ

１．はじめに

調査・研究活動を行う者にとって、これからも増え続けるであろう情報の洪水の中から、いかにして有効な情報を効率よく集めることができるかは、調査・研究活動の効率／成果を左右する大きな問題の１つであると言われている。それには、調査・研究領域に対する個人毎の概念体系をすばやく確立し、その体系を効果的に利用した情報収集を行うことが重要である。

しかし、文献ＤＢに対して現在一般に使われているブール演算等（AND／OR等）の検索方法では、検索対象が明確な場合や検索対象にうまく一致するキーワードを思い付く場合等の特定な場合を除いては、抽出文献が多すぎたり少なすぎたり、イメージに合う文献がなかなか見つからなかったりと、効率的な情報収集ができず、体系把握などの派生的効果もあまり期待できないのが現状である。

そこで我々は、研究者のための新しい情報分析ツール『ThinkView』を試作し、試用によるシステム有効性の検証を行った。

２．ThinkViewの概要／特徴

２．１　ThinkViewとは

ThinkViewは、ターゲットとなる文献を獲得することのみを目的とする従来の情報検索システムとは異なり、ターゲットを絞っていく過程で、対象分野に対する利用者独自の概念体系の生成を支援し、情報フィルタリング・特定分野の理解・コミュニケーション等において、その継続的利用による効果を狙う研究者のための新しい情報分析ツールである。（図１）（図２）

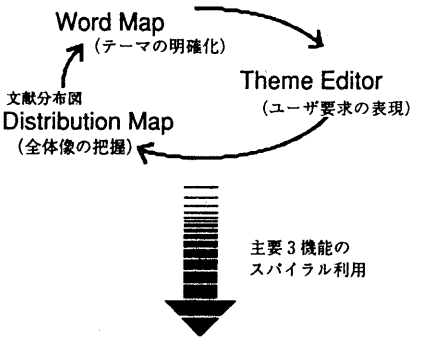


図２：主要３機能の関連

あらかじめ生成しておいたViewSpaceファイルを利用して、ユーザは次のような操作を行う。

①操作対象の文献集合に対するワードマップを表示し、②ワードマップ上の単語を参考にしてテーマエディタによりテーマを表現し、③文献分布図上にそのテーマに対する分布を表示し内容を確認する。　ユーザは①~③の操作をスパイラルに繰り返しながら、研究対象領域に対するユーザ独自の体系を徐々に構築していく。（図１）

（注）ここでいうテーマとは対象領域に対してユーザが抱く抽出要求や分類観点のことであり、複数の名詞とそれらの関連により表現する。

2．2　ThinkViewのねらい

ThinkViewの特徴／ねらいは以下の点である。

・システムに知的なことはさせないで、人間が知的なことをすることを支援するための道具に徹する。

・あまり意味的に踏み込んで処理が遅くなっては道具の価値がなくなるので、表層的なマッチングに限定しても有効な道具が作れることを狙う。

・一時的な利用で終わらせるのではなく、継続的に利用することで価値がより高まる道具にする。

・従来のキーワード検索では生成しにくかったユーザオリジナルな集合の獲得及び全体像の把握の実現をはかる。

・最適な分類をシステムが提供するのではなく、ユーザが試行錯誤を繰り返しながら最適な分類を獲得していく作業を支援する

・情報分析作業の過程や結果（体系等）が他人と共有可能な道具を目指す。

3．主要機能の説明

3．1　ワードマップ（Word Map）

操作中の文献部分集合に含まれる全単語を3種類のソート（出現文献数順／総単語数順／５０音順）機能で表示することによって、以下の点においてユーザを支援する。（図3）

・抽出イメージの明確化
・単語からの全体像の把握
・検索洩れの少ない検索式の作成

現在のワードマップはソートによる構造化のみであるが、今後は階層化表示等の多様な構造化を行うことを検討している。

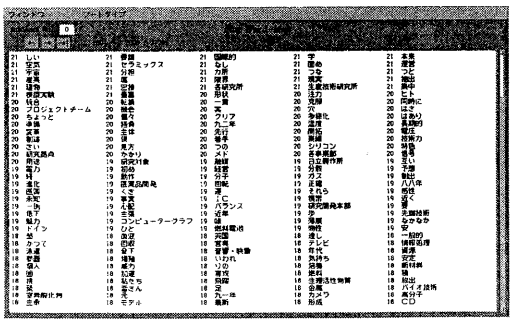


図3：ワードマップ

3．2　テーマエディタ（Theme Editor）

ユーザは抽出したいイメージを図4のようなテーマエディタを用いて作成する。システム側は各フィールドに入力された単語をもとにして文献特徴ベクトルを生成し、ベクトル間類似度による文献分布図を生成する。

任意のフィールドに複数の単語を入力することで、入力されたものを１つの属性値として計算することができる。又フィールド横のスライダーを操作することで、類似度に対する対象属性の寄与率を変更することができる。

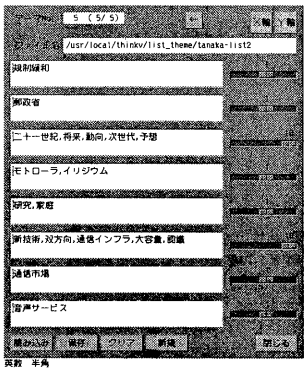


図4：リストテーマエディタ

頭の中のイメージとのギャップをより少なくするようなユーザ抽出概念の表現方法として、ツリー型のものも現在作成中である。

3．3　文献分布図（Distribution Map）

3．3．1　絶対分布図

２次元マップ上に各文献の分布を点の色と大きさで度数表現する。（図5）

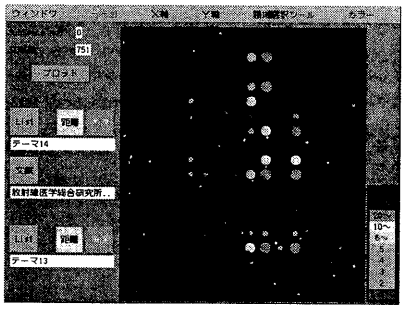


図5：絶対分布

X及びY軸には以下のものを指定することができる。

・指定テーマと各文献の類似度
・指定文献と各文献の類似度
　(a)テーマをもとにした類似度
　(b)テーマによらない類似度
・指定テーマに対する各文献の評価値

・各文献のもつ既存属性値
　発行年月　等

類似度を表示する場合は、(a)ユークリッド距離と(b)コサイン角度のどちらかを状況に応じて選択できる。
作成した分布図に対して、以下の領域ツールによる文献抽出及びタイトル・本文の確認が行える。
・四角形　　・楕円　　・同心円

この絶対分布図により、従来のブール演算では不可能である微妙な境界の作成、抽出集合内外の文献分布状況の把握が実現でき、情報空間の認識向上に役立つことが期待できる。

3．3．2　相対分布図

操作対象集合内に含まれる全文献の文献間類似度をもとに、階層的クラスタ化法で生成したクラスタ分類をデンドログラムにより表示する。　絶対分布と同様にテーマの設定有無及び、(a)ユークリッド距離と(b)コサイン角度のどちらかを状況に応じて選択できる。（図６）

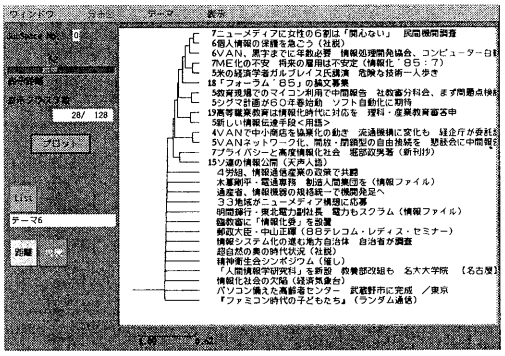


図６：相対分布

スライダーで表示距離（＝類似度）の指定を変えることで分布をマクロ的に見たり、ミクロ的に見たりすることが可能である。クラスタの内容を象徴するものとして(a)クラスタ中の１タイトル又は(b)クラスタ内の特徴的な単語のどちらかを枝の右側に表示する。

4．主なロジックについて

4．1　ViewSpaceファイルの作成

ViewSpaceとは、文献間類似度の算出を迅速に行うため、対象とする文献集合に出現する全単語について、文献別の出現数や出現文献数などをあらかじめカウントしファイルに格納したものである。ViewSpace作成プログラムは次のステップで実行される。
・テンプレートによるタイトル、発行年月などの自動抽出
・単語（名詞）の切り出し
　※単語の長さが極力長くなるように切り出しを行う
・文献別の単語カウント
・ViewSpaceファイルの作成（必要統計情報の算出）

4．2　文献特徴ベクトルの生成

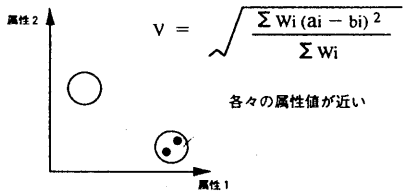
類似度算出に用いる文献特徴ベクトルは、各単語（名詞）の文献別出現数をもとに作成されるが、テーマ指定の有無によって下の２つのパターンに分かれる。
・部分比較用ベクトル　．．テーマの指定がある場合
　次元数：指定したリストテーマの構成グループ数
　属性値：各グループを構成する名詞の出現数合計を標準化した値
　（注）出現数合計は、ユーザが指定した単語のViewSpace作成時に切り出された各単語に対する部分一致でカウントする。

・全体比較用ベクトル　．．テーマの指定が無い場合
　次元数：文献別の名詞種数
　属性値：名詞の文献別出現数を標準化した値

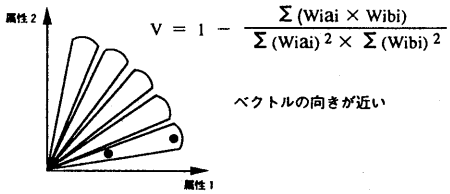
4．3　類似度算出方法

文献特徴ベクトルをもとにする類似度の算出方法として、本システムでは２つの算出式　(a)ユークリッド距離　(b)コサイン角度　を提供している。又、類似度とは異なる尺度として　(c)テーマに対する評価値　を提供している。
各算出方法による類似度の性質の違いが直感的にわかりやすいように、次元数＝２の場合の類似島の違いを下に示した。（図７）

（a）ユークリッド距離



（b）コサイン角度



（c）テーマに対する文献評価値

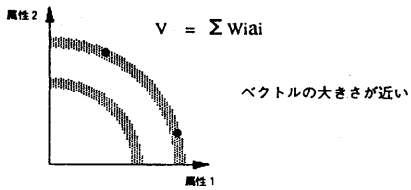


図７：各算出方法の類似に対する考え方
　　及び　属性空間上における類似島

（注）テーマに対する類似度の算出の場合は、テーマベクトルの各属性評価値として、各々の最大値を与えて計算を行う

ここでユーザが注意しなければならないは、ベクトルを構成する各属性の特徴量が、表面上のパターンマッチからなるという性格上、本システムにおける類似度が示唆するものは内容の類似性ではなく、各文献が言及しているいろいろな分野の重なり度合いを示しているにすぎないということである。

4．4　クラスタ特徴語の抽出

相対分布の表示において、生成されたデンドログラムの各クラスタにおける特徴語を表示することで、以下ような点でユーザを支援する。

・標本分布と因子（特徴語）の関連を提示することによる対象情報空間の認知の向上
・特徴語から各クラスタ内容を予測することによる不要なクラスタの排除

各クラスタ特徴語の抽出は次のように行う。

(a)事前に以下の処理を行っておく
・各文献別に、名詞を切り出す。
・切り出した名詞には、名詞の直後の語（助詞、句読点など）から決定するランク（重要度）を各々に付加する。

(b)該当クラスタ内の文献に含まれる全ての名詞（重要度が指定値以上のもの）について、特徴語レベル値を算出し、それの上位のもの（５～１０個）をクラスタ内の特徴語とする。

各々の単語の特徴語レベル値は次のように求める

$$V = \Sigma WiPi$$

（W：ウェート、P：各クラスタ別特徴量）

P（任意の単語のクラスタ別特徴量）として現在以下のものを検討・実験中である。
①クラスタ内ランク合計
②単語出現文献数
③１文献内最大単語出現数
④タイトル内出現有無
⑤①の標準偏差
：

5．評価テスト

5．1　目的

・ThinkViewの有効な利用方法／場面の確立
・各機能の有効性の検証
・情報収集／情報分類における新たな問題点の発掘

5．2　テスト方法

試験者（４名）の関心の高い分野に関するキーワードにより商用ＤＢから収集した新聞記事について、試験者の思いのままに分類、抽出、全体把握などのテストを行ってもらい、後日に座談会を実施した。又、同じデータを用いて通常の全文検索システムによる比較テストも行った。

5．3　使用データ

テストは、商用DBからキーワード検索で収集した新聞記事により行った。各試験者のデータ量は100～500記事程度である。

5．4　テスト環境／処理時間

IBM_PowerStation520（15.9 SPECint、53.1 SPECfp）を使用して開発、テストを行った。

主な処理に掛かった時間は以下の通りである。

・ViewSpaceファイルの生成

		総名詞種数
113文献	12分	7368
289文献	24分	15574
482文献	36分	18074

・分布図の生成、表示

	絶対分布図	相対分布図
113文献	2秒	4秒
289文献	15秒	65秒
482文献	18秒	257秒
（参考）751文献	21秒	896秒

・クラスタ特徴語の抽出
１クラスタ（113文献の場合）につき30秒程度

「絶対分布図」に比べ「相対分布図」及び「クラスタ特徴語の抽出」については、処理速度において現状での実用性は低いが、今後のCPUパワーの増大により実用性も向上していくと思われる。

5．5　評価テストの結果について

5．5．1　有効な利用方法／場面

評価テストの結果、断片的な利用や体系化が進んだ研究対象領域に対して利用するのではあまり効果がなく、むしろ全く新しく取り組むような領域に対して、テーマ及び分布の生成を繰り返しながら研究者独自の概念体系を徐々に固めていく．．というような場面において利用することが最も効果的であることがわかった。　又、その過程の中でも相対分布図は序盤に適し、絶対分布図は中盤以降の段階でよりその効果を発揮することもわかった。

分布図の利用においては次に示す利用例のように、１度文献集合からサンプリングした文献を読んで評価してから分布を生成するという使い方が効果的であることもわかった。

＜サンプル読みを行った場合の分布図の利用＞

①サンプリングした文献を１件ずつよんで、気に入った文献をチェックする。

②分布生成のもとになるテーマを一からつくる代わりに、チェックした文献集合を指定する。システムはチェックした文献集合に含まれる単語からのテーマの自動生成と編集を支援する。

③チェックされた文献が任意の領域にかたまるように分布の生成とテーマの修正を繰り返す。このときチェックした文献は分布上で明示されている。

④良いテーマと分布が得られたら、それを関係式として保存し、以後のより大きな集合の分類／フィルタリングに利用する。再利用や共有がしやすいように、テーマと分布がマンガ化／略図化できれば素晴らしい。

次に、ThinkViewの開発者及び試験者により実施した座談会から出た「機能別評価」及び「新機能案」に関する意見を参考までに紹介する。

５．５．２　機能別評価

(a)絶対分布図について
- ・絶対分布はオリジナルな境界面を生成するのに良い
- ・特に指定文献との類似度による分布が良好
- ・分布図を掌握しにくい
  ブール演算操作に比べて結果が予想しにくい
- ・マークした特定の文献を分布図上に明示する機能がほしい

(b)相対分布図について
- ・分布結果に対するギャップは相対分布のほうが少ないように感じる
- ・デンドログラムは解釈しにくい
- ・明らかに不要な文献は分布図作成の非対象にしたい
  サンプル読みによる不要クラスタの非対象化　等

(c)ワードマップについて
- ・マクロ的表示機能がほしい
  グループの代表値のみの表示　等

(d)クラスタ内特徴語抽出について
- ・上位クラスタにおいては、文献集合全体を象徴するような単語を抽出することに成功した
- ・各クラスタの特徴語の抽出に30秒程度かかってしまい、一度に全クラスタの特徴語を抽出するにはパフォーマンスが非常に悪い
  ↓
  特徴語を算出するクラスタの制限等が必要
  （例）クラスタ内文献数が一定値以上のもの
- ・精度とニーズの関係が比例するわけでもないので、適当な精度を選択するべきである

５．５．３　今後ほしい機能について

(a)分類成果の再利用
- ・テーマと分布、抽出結果、操作履歴の保存機能
- ・概念マップ機能
  相対分布／絶対分布を簡略化した図で表現し、抽出結果の保存・再生・共有を容易にするもの

(b)例示と学習
- ・サンプル分布からのテーマ生成（逆演算）
- ・テーマからのブール演算式生成
  → 商用ＤＢからの効率よい収集を支援
- ・サンプル提示による自動分類機能（図８）
  提示するサンプル（階層構造、離散的分布）に対して文献間類似度を用いて母集団をマップする

  サンプルの提示例
  - ・個人のＨＤなど既存の分類（階層構造）を提示
  - ・分類したいぶんのサンプルを母集団から提示
    各サンプルは１～３程度の文献からなる
  - ・他人又は本人の完成した分類式の提示

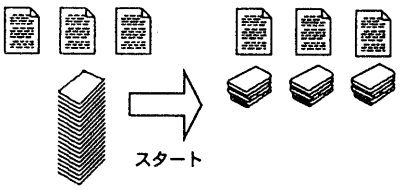


図８：サンプル提示による自動分類機能

(c)マーク機能
- ・任意文献の非対象化
- ・任意文献を分布上に明示した操作
- ・相対分布／絶対分布の比較表示（図９）

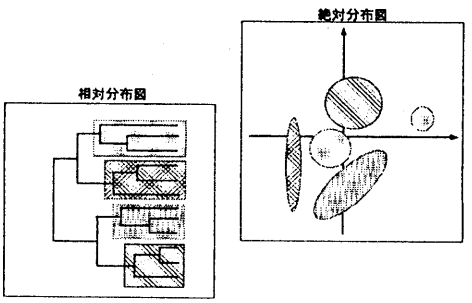


図９：相対分布／絶対分布の比較表示

6．考察

現在のプロトタイプは速度等の点での問題はあるが、ブール演算等を用いた通常の検索システムに比べ、ユーザ独自のきめこまかい分布を作成しやすい、又ツールを操作すること自体が対象領域に対する概念体系の確立に役立つ等の有効性が確認できている。

この章では、次に示すような各項目における、より具体的な問題点や解決案について私なりの考えを述べる。

6．1　文献ファイルの解析の深さ

本システムは文献特徴ベクトルの属性として、“名詞の意味する概念に関する言及度合い”のみを採用し、その要素としての全名詞の文献別出現数のみを事前に一括解析しているが、ユーザが文献分析の際に使いたくなる情報はこれ以外にも次のように多様なものが考えられる。

(a)文献自体のもつ情報
- ・名詞の出現場所
  タイトル、アブストラクト、結論．．
- ・名詞の使われ方（格）
- ・単語ペアの出現数
  １文中、１段落中．．
- ・表、写真、絵図、数式の数
- ・使用フォント、レイアウト、構成バランス
- ・語尾、語感
- ・参照文献（参照数、参照文献種）

(b)関連情報
- ・著者実績、著者所属
- ・発行年月、出典
- ・文献の引用度
- ・文献に対する他人の評価
  （類似例）競馬　：オッズで選択
  歌謡曲：オリコンで選択

又、文献分布図生成時のパフォーマンスを向上させるためには、事前に各文献の特徴情報の数値化処理を行っておくと良いが、それにより次のようなデメリットも生ずる。

・分析の自由度とのトレードオフ
　事前の数値化、ベクトル化により多くの情報が失われる

・データ量とのトレードオフ
　事前の数値化、ベクトル化によりオリジナル文献に対するデータ量が増加する（単語ペアの出現数などは、事前に解析すること自体が現実的ではない）

・データの鮮度とのトレードオフ
　事前の解析にある程度の時間が掛かってしまうので、鮮度が重要となる情報には向かない

本システムにおいては、かなり浅いレベルでの解析のみを事前に行うことを現実的な１つの解として採用したが、コンピュータパワーの増大につれて、将来的には事前解析は行わない方向に向かうであろう。

6．2　ユーザ要求（テーマ）の表現方法

現在の検索システムのほとんどはユーザ要求をブール演算式で表現するものばかりであるが、本システムではリストによる表現を試みた。今後とも多様なユーザ要求をギャップ少なく表現できるようなモデルの研究が重要になると考える。

私個人としては自然言語によるあいまいな表現よりは、図形等を用いた表現のほうが有効なのではないかと思う。又、単に結果を表現するのではなく、テーマ生成時の思考過程をも反映するようなモデルが良いのではないかと思う。現在考えられる表現モデルとしては、ツリー、テーブル、立体的ベン図、スパイラルなどがある。

6．3　類似度算出方法

本システムにおいては、文献特徴ベクトルから文献間類似度を求める演算式として「ユークリッド距離」と「コサイン角度」を提供しているが、それらは共に相違加算方式である。

一般的に、このてのシステムにおいては相違加算方式が用いられることが多いようだが、現実社会において人が似ていると感じるもの同士においては、多くの相違点がある際だった類似点によって無視されていることも少なくなく、類似加算方式による類似度のほうがユーザの感覚によりフィットするようにも思われる。

相違加算方式では値０が最も類似することを意味するのに対し、類似加算方式においては値０が全く類似しないことを意味する。参考までに類似加算方式の演算案を下に示す。

類似加算方式の演算案
　属性値間の差を下の用にランク値化したものの和
　　差が0.1未満　→　10
　　差が0.2未満　→　5
　　差が0.5未満　→　3
　　差が0.5以上　→　0

6．4　クラスター化方法

本システムにおける相対分布図の生成処理は以下の様に行っているが、その処理時間のほとんどが③のステップに費やされている。

①対象文献集合内の全ての文献特徴ベクトルの生成
②対象文献集合内の全組み合わせの文献間類似度の算出
③階層的クラスタ化法の実施（UPGMA法）
④デンドログラムの描写

しかし、階層的クラスタ化法を文献数が多い場合に用いることは、評価テストにおける処理時間からも明らかなように現在のCPUパワーにおいては実用的ではなく、又デンドログラムによって表現される相対分布も一般ユーザが把握しやすいものとは言い難い。

従って実用的な相対分布の生成・表示方法の検討が必要であるが、現時点では下に示すように、ユーザが提示する指標を用いた「この指止まれ」的な非階層的クラスタ化方法が良いのではないかと考えている。

非階層的クラスタ化案
①ユーザあるいはシステムが互いに類似しないような８個程度のクラスタの核となるような文献を抽出
②残りの文献は最も類似する核のクラスタに分類

6．5　文献分布の表現（視覚化）方法

本システムにおける情報の視覚化方法には、改良の余地がまだ大いにある。　私は、必ずしも情報空間を立体的に表現することがユーザにとって使いやすいとは思わないが、コンピュータパワーの増大により、VRMLに象徴されるような3D的なユーザインタフェースで、大量の情報をスピーディに操作できるようになれば、情報空間認知の問題は一歩前進するのではないかと思う。

以下は本システムにおけるユーザの情報空間認知の向上を図るための改善案の一部である。

（絶対分布図）

現在、分布上には度数を表す点が描写されているのみであるが、タイトル等文献の中身を特徴づける情報も同時に表示すべきである。しかし、適当に表示するタイトルの数を制限しないと文字が重なって滅茶苦茶になってしまう。

（相対分布図）

デンドログラムは、4次元以上の属性空間の各標本位置を平面上に表現する1手段ではあるが、標本のもつ各属性値が分布上に全く表現できないために分布図の意味を捉えにくいという面もある。この問題を克服する工夫として、特定の2属性からなる平面マップ上に、立体的なデンドログラムを描写することも行われているようである。

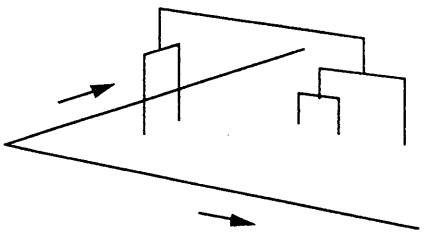


図10：平面マップ上の立体的なデンドログラム

6．6　ThinkViewの応用

ThinkViewのエキスの応用としては、以下のようなものが考えらる。

(a)自分が蓄積してきたメモ、アイデアを再統合／整理するのを支援するツール

(b)ページ数が爆発していくWWW情報を分類し、有効に活用していく環境

WWWのBookMarkの自動分類　等

(c)顧客に情報を整理して提供するサービスをするための支援ツール

顧客にとって最適なメニュー構造を作成することを支援するツール。階層構造のメニュー及びメニューに含まれる情報は、顧客の情報空間認知が安定するまでは、顧客のオペレーションに伴って進化していくようなものを提供することが望ましい。

(d)オーダーメードのレポートを短期間に高品質で作成するための、『プライベートブック』

ある分野についてのレポートを書く場合、まず文献を数多く読んでその分野に対するキーワード体系を構築することが重要であり、経験的に3回ぐらい作業を繰り返すと、ほぼ漏れの無い体系が構築できると言われている。

又、作成したキーワード体系に従って、ネットワーク上等の電子情報からオリジナルなプライベートブックを自動的に生成できれば、短期間で高品質なレポートを作成することが期待できる。

7．おわりに

今回提案した『ThinkView』は、まだまだ解決しなければならない問題点も多いが、対象分野に対する利用者独自の概念体系の生成を支援するツールとしての可能性を示すことができたのではないかと思う。　今後はより研究者の活動全体を見据えた支援ができるようなツールとして検討していきたい。

WWW、メール、NetNews、テレビ、雑誌、．．．氾濫する情報の中で、どの程度情報収集に時間をさけば良いのか？（いっそ情報など収集せずに、山にでもこもって思考すべきなのでは！？）、蓄積している情報は本当に必要なのか？（もしかしたら不安の解消をしているだけなのでは！？）．．．我々の悩みはつきないが、今後とも自分がいかに情報とつきあっていくべきか悟りが開けるまで研究を続けていきたい。

参考文献

[1]P．イングベルセン：「情報検索研究－認知的アプローチ」：監訳　藤原鎌男：訳　細野公男、後藤智範、岸田和明：トッパン、1995年

[2]H.C.Romesburg：「実例クラスター分析」：訳　西田英郎、佐藤嗣二：内田老鶴圃、1992年