データベースシステム 104－6
（1995．7．18）

3次元自己組織化マップに基づく文書のブラウジングと検索

銭 晴† 波多野 賢治‡ 田中 克己†
†神戸大学大学院自然科学研究科知能科学専攻
‡神戸大学大学院自然科学研究科情報知能工学専攻

本稿はKohonenの自己組織化マップの3次元化とそれに基づく文書の概覧、検索機構について述べる。この機構に基づき文書データ群の要約表示や動的な自己組織化、概念発見などの機能を提供できる。また、自己組織化マップを3次元に拡張して表示する機構により、視覚的に優れ段階的詳細化が可能なブラウザや曖昧検索を支援するユーザインターフェイスを実現できる。さらに入力データとしてHTML文書のような構造化文書を用いることもでき、WWWのような環境にも対応できると共に、リンク情報を含む構造化文書群の自己組織化にも適用できる。

Document Browsing and Retrieval based on 3D Self-Organizing Map

Qing Qian† Kenji Hatano‡ Katsumi Tanaka†
†Division of Intelligence Science,
Graduate School of Science and Technology, Kobe University
‡Dept. of Computer and Systems Engineering,
Graduate School of Science and Technology, Kobe University

In this paper, we will describe our 3-dimensional(3D) extention of Kohonen's self-organizing map, and our document browsing/retrieval system based on the 3D self-organizing map. The proposed mechanism makes it possible to provide facilities of a summary-information generation, a dynamic self-organization, and a concept discovery for document data. The 3D extention of the self-organizing map is useful to realize a visually-effective document interface with supporting incremental 'zoom-up' and ambiguous queries. The proposed system can accept structured documents such as HTML documents as its input, and can generate a map which focuses on their hyperlink information. The system can be easily extented to WWW environment.

1 まえがき

近年の社会の情報化にともない、電子化された文書量が増大し、動的に分類を行なうなどの動的、自己組織的な構造化機能を持つ文書データベースシステムの必要性が高まっている。

文書データベースに対して現在用いられている検索方式として、カテゴリやキーワードなどあらかじめ文書に付加した2次情報を利用しながら文字列の一致に基づくAND/OR演算による対話型検索方式やSQLなどによる集合的問い合わせ言語を用いた検索などが挙げられる。しかしこれらの方式では、結果として検索を行なおうとするユーザーに検索用キーワードやシソーラス、文書の分類体系に関するある程度の知識を要求する傾向がある。さらに、検索用キーワードやシソーラス、文書の分類体系は事前に規定されていることが多いが、進展の激しい科学技術分野ではこれでは不十分であり、収集文献群から動的、自己組織的に構造化できる機能が望まれている。このためキーワードに基づく類似度計算、キーワード索引生成、自然言語解析などの従来的なテキスト検索処理や記号処理を経ずに対処する手法として、ニューラルネットワーク情報処理技術の情報検索への適用が始まっている[1][2][3][4]。

このような背景から、本研究は、自己組織化マップを用いて文書データベースの自動的な構造化や分類を行ない、そのデータ集合を3次元自己組織化マップの形で表示し、動的な構造生成、ズームイン等の段階的詳細化、曖昧検索などが可能なブラウジング・検索システムを実現する試みを行なった。

本稿では主に次の事項に関して述べる。

- 文書の自己組織化および文書の特性ベクトル生成のアルゴリズム
- Kohonenの自己組織化マップを基に実現した3次元マップに基づく種々の検索機能

2 基本的事項

2.1 自己組織化マップ

ニューラルネットワークの一種である自己組織化マップ(Self-Organizing(Feature)Map、以下SOM法と呼ぶ)は、1990年にT.Kohonenによって提案された教師なし競合学習モデルである[5]。出力層の各ユニットが層の中で位置を持つという点が他の学習モデルと異なる点である。このモデルの特徴はデータに隠されているトポロジカルな構造を学習アルゴリズムにより発見し、通常2次元空間で表示するというものである。

SOM法で用いられるネットワークは、ユニットを2次元上に配置したものである。入力データは通常高次元の特徴(feature)ベクトル$\boldsymbol{x}$にパターン化され、ネットワーク中の各ユニットiは入力パターン$\boldsymbol{x}$と同次元のベクトル$\boldsymbol{m}_i$をもっている。学習はこれらのユニットを入力パターンに選択的に近付けることによって進行する。SOM法では入力パターンに一番近いパターンを持つ出力ユニットcおよびその近傍(図1)のユニットの集合N_cのみを入力パターンに近付ける学習アルゴリズムを採用している。また、統計的に正確な学習効果を得るため、一定の学習回数Tをとらねばならない。

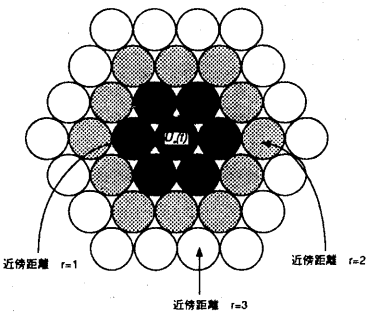

図1: 近傍の取り方

すなわち、SOM法による自己組織化アルゴリズムは以下のようになる。

自己組織化アルゴリズム(SOM法)

1. **各入力データをパターン化する**

   $X = \{ x_1, x_2, \ldots, x_p \mid x_k \in R^n, k = 1, \ldots, p\}$

2. **出力層にある各ユニットの持つパターンを初期化する**

   $M = \{ m_1, m_2, \ldots, m_q \mid m_i \in R^n, i = 1, \ldots, q\}$

3. **入力パターンx_kに一番近いパターンを持つ出力ユニットcを探す。**

   つまり、次式のような$\boldsymbol{m}_c(t)$を持つユニットcを求める。

$$\|\boldsymbol{x}_k - \boldsymbol{m}_c(t)\| = \min_{\text{for all } i} \{\|\boldsymbol{x}_k - \boldsymbol{m}_i(t)\|\}$$

   $\|\cdot\|$は距離を表し、ユークリッドノルム等が用いられる。

4. 各出力ユニット c とその近傍のユニットの集合 $N_c(t)$ を入力パターン x_k に近付ける。

$$m_i(t+1) = \begin{cases} m_i(t) + \alpha(t)[x_k(t) - m_i(t)] & (i \in N_c(t)) \\ m_i(t) & (i \notin N_c(t)) \end{cases}$$

$$\alpha(t) = \alpha_0(t) exp(-\|r_c - r_i\|^2/\sigma(t)^2)$$

ここで$\alpha(t)$ は学習率であり時間とともに 0 へと単調減少し、

$$o_i(t) = x \cdot m_i(t)$$

である。また、$\| r_c - r_i \|$ が、ユニット c と i との距離である。さらに、$N_c(t)$ の大きさも時間とともに単調に減少する。

5. $k = k+1$ とし、3〜4 を繰り返す

6. $t = t+1$, $t \leq T$ (Tはあらかじめ設定された学習回数) とし、$N_c(t)$ とαを次第に小さくしながら 3〜5 を繰り返す

この結果、入力ベクトル空間で近くにあるものは、ネットワーク上でも互いに近傍のユニットへと射影されるような写像が完成することになる。

2.2 SGML と HTML

SGML(Standard Generalized Markup Language) は、文書の論理構造や参照構造などを規定する言語で、1986 年に形式定義が ISO 標準 8879 として標準化されている。

ハイパーテキストはコンピュータの支援により、複数の断片的な情報をリンクによりネットワーク状に結合し、それらを非線形情報として表示、構築、および管理するシステム [7] であるが、WWW において採用されているハイパーテキストを記述する言語が HTML(HyperText Markup Language) で、文章のフォーマット情報、グラフィックや音声、テキストの一部や他の文書などのリソースへのリンク情報や文の整形を SGML の DTD(Document Type Definition) によって定義されたタグを平文に挿入する構造化タグ言語である。これは、SGML をネットワークワイドに発展、応用したものと言える。

3 文章群からの 3 次元自己組織化マップの生成

図 2に、我々の 3 次元自己組織化マップによる文書群の自己組織化とそれに基づくブラウジング・検索の流れを示す。未分類状態の文書データベースからキーワード検索などによって興味の対象となる文書集合を設定し、この文書集合から 3 次元自己組織化マップを生成する。以下では、文書の特徴ベクトルの生成法および 3 次元マップの生成法について述べる。

3.1 文書の特徴ベクトル生成

文書の特徴ベクトルの生成のために以下のような処理を行なう。

1. ストップワードの処理

   あらかじめ用意しておいたストップワード辞書によってキーワード候補となり得ない単語を削除する。複数形などは同義語として処理される。

2. キーワード候補の重みづけ

   文書 D_i から切り出されたキーワード候補 T_k に対して、その重み w_{ik} を次のいづれかの方法で決定する。

   - w_{ik}: D_i 中の単語 T_k の出現頻度
   - $$w_{jk}: \frac{tf_{jk} \cdot log(N/n_k)}{\sqrt{\sum_{j=1}^{t} (tf_{jk})^2 \cdot (log\,(N/n_j))^2}}$$

   ここで、tf_{ik}(term frequency) は、D_i 中の単語 T_k の出現頻度、N は総文献数、n_k は単語 T_k を含む文献数である。

   また、HTML 文書の場合、アンカー (ハイパーリンクの出発点) となっている単語は、その文献を特徴づける度合いが高いものであると判断し、さらに重みを大きくする方法をとっている。

3. 特徴ベクトルの生成

   各キーワード候補 T_k に対して、重み $w_{1k}, \ldots, w_{Nk}$ が生成される。あるしきい値以上の重みの平均値をとり、その値の最大のものから順に定数個のキーワード候補を選択し、このキーワード候補群からなる特徴ベクトルを生成する。

これらの作業によりテキストから文字列を抽出すると図 3 のようになる。

3.2 3 次元マップ生成アルゴリズム

3.1で得られた特徴ベクトルを用いて Kohonen の自己組織化アルゴリズムを適用して学習を行なう。学習が終了した時点で、各出力ユニットにはそのベクトルと距離的に最も近いベクトルを持つ文書が対応づけられる。

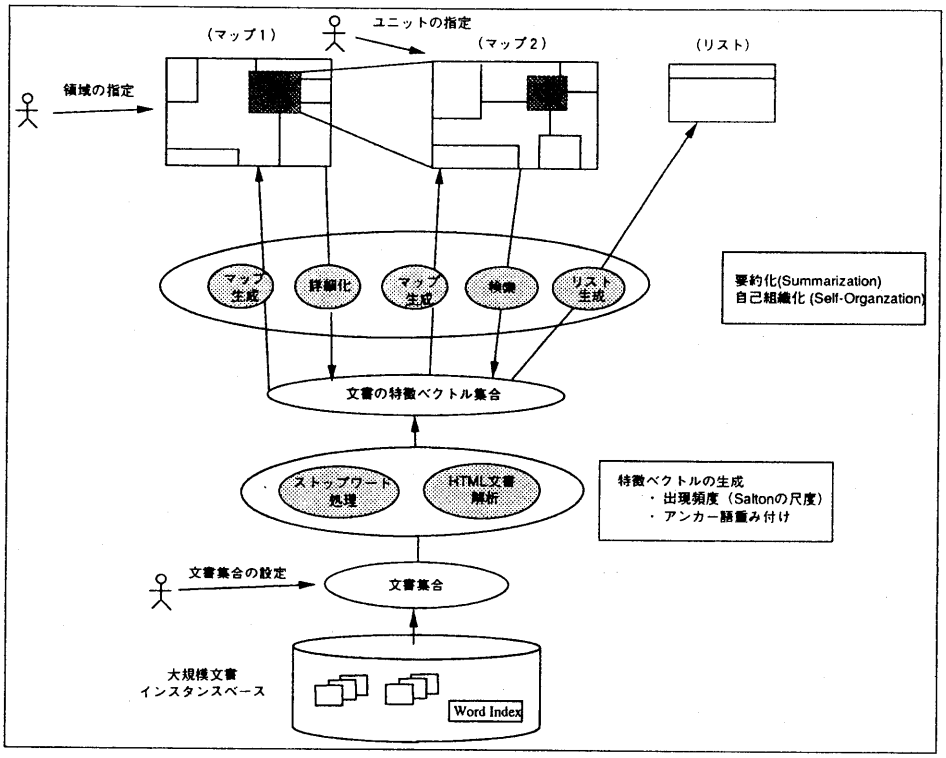

図 2: 文書データベースの自己組織化機構

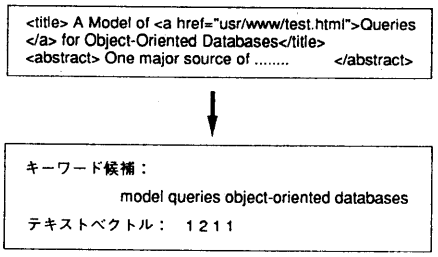

図 3: ハイパーテキストの特徴ベクトル化

3.2.1 キーワード選定のアルゴリズム

単一の出力ユニットによって構成される領域のキーワードには、その出力ユニットの要素中もっとも大きな値を持つ要素、すなわちその出力ユニットにマッピングされた文書群に含まれる単語の中で最も出現頻度が高いものを選ぶ。

なお、キーワード抽出の際には各出力ユニットの持つパターンについて要素の大きいものから 3 つ抽出し、それらを単語に変換することにより、その出力ユニットに分類されたテキストを象徴するキーワードを得ている。この方法はテキストの内容、すなわち切り出された単語の数が多ければ多いほどよりテキストの詳細な情報をベクトルに持たせることができ、キーワードマップの曖昧検索機能や詳細化機能の実現に必要となっている。

複数の出力ユニットによって構成される領域のキーワードには、その領域に含まれる各出力ユニットの内部ベクトルをすべて加えてできた疑似ユニットをまず求め、その疑似ユニットに含まれる単語群の中で最も出現頻度が高いものを選ぶようにする。

3.2.2 3 次元マップ表示アルゴリズム

これまで自己組織化マップで学習し、得られたキーワード出力マップは 2 次元平面上で作成されたものであった。しかし、学習した後のテキストの分布は一ヶ所に集中し 2 次元平面上のすべてのユニットにテキストが写像されるとは限らないため、情報検索に支障をきたす。ここで、2 次元平面より 3 次元にマップを進

化させれば、よりユーザーが容易に情報検索することができるような環境を整えることができる[3]。つまり、キーワード出力マップ上の各々のユニットに対し、それに写像されたテキストの数により、高さを決められる円筒形がそのユニットの外観として、つまり、円筒の高さが各ユニットに射影されたテキストの数に比例するというふうにとらえられているということにより、より良いユーザーインタフェース環境が提供されている。これにより、ウォークスルー機能も実現している。

この3次元マップにおいて、学習したマップを見やすく表示するために、距離の近いベクトルを持つユニットが同じ領域に、距離の少し離れているベクトルを持つユニットが異なる領域に分割する。異なる領域は異なる色づけによって区別されており、それら異なる領域は3次元キーワードマップの上に配置されている。ベクトル距離の遠近の判断はユーザの指定によるものであり、指定された数字より小さいのはより近いという意味であり、大きいのはより遠いという意味である。また、分割された領域に一番強い特徴、つまり最も出現頻度が高いキーワードを選び、その領域のラベルとして付けられる。

3.3 実行例

実際に 20×20 のマップを用いてマップ生成を行った。対象はオブジェクト指向データベース関連の論文のタイトルおよびアブストラクト200本を3.1の方式に乗っとってパターン化し、入力とした。学習回数は11,000回、学習式には2.1のものをそのまま用いた。

このもとで学習させるとユニット数は400のマップは4分間で生成された。

データベース関連の論文という特定の分野のハイパーテキストを自動分類したため、それほどマップ全体に文書が分散したという感は受けなかったが、テキストが一箇所に集中しているわけでもなく随所に分布の隔たりが見られるため、自動分類の試みは成功しているといえる。

1. キーワードマップ

   図4はキーワードマップの実例で、その上に2桁数の異なる色づけの領域がある。各々の領域上にはその領域を代表する"language"、"programming"、"model"というようなキーワードも出力されている。

2. 3次元マップ

   図5は図4の異なる視点のものである。その上のノードは円筒の形を持つ、それぞれの円筒の高さはそれらに射影された文書の数に比例する。

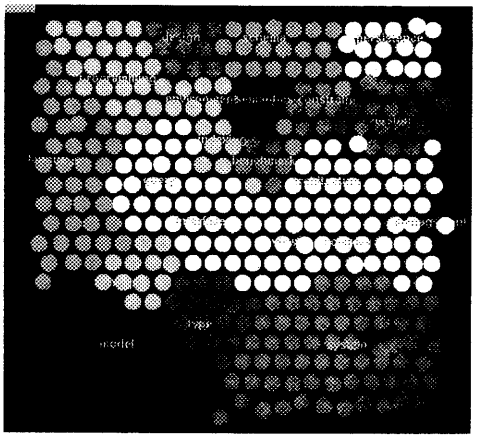

図4: キーワードマップ

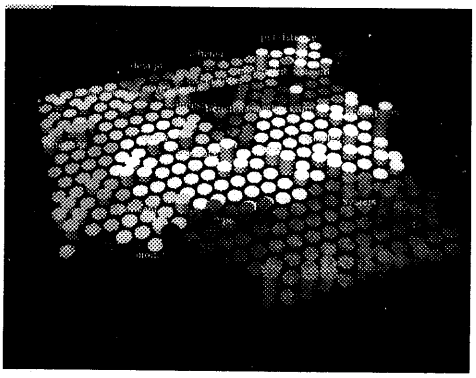

図5: 3次元マップ

4 システムの実現

4.1 システムの全体構成

このシステムはSiliconGraphics社のIndy上で実現[1]されて、次の2つの部分からなる。

1. 特徴ベクトル生成部

   3.1節で述べられたベクトル生成アルゴリズムがCで実現されている。

[1]マップ表示部分以外はCコンパイラ環境が整っていればよい

2. **自己組織化マップ (SOM パッケージ) による学習部分**

1で作られたテキストベクトルを入力データとして自己組織化マップに学習させる。

ここでは、T.Kohonen が率いるプログラム開発チームにより開発されたSOMパッケージを用いて、ユニットに対応するベクトルを初期化し、それらのベクトルを上で生成された入力ベクトルに学習させる。これは、すべての自己組織化マップの応用に対応するプログラムを含んでおり具体的には以下のようなものがある。

- **初期化プログラム (Initialization program)**
  このプログラムでは、出力空間のユニットに対応するベクトルは初期化される。
- **学習プログラム (Training programs)**
  このプログラムでは、出力空間のユニットに対応するベクトルとそれの近傍にあるパターンを入力パターンに近づける処理は行われている。
- **分量精度プログラム (Quantization accuracy program)**
  このプログラムで、分量誤差の平均値が計算される。つまり、学習後、各入力ベクトルに対しマップ中の最も距離が近いパターンを持つユニットが見付け出され、同時にそれらのユニットを持つパターンの分量と入力ベクトルの分量間の誤差の平均値が計算される。
- **グラフィックプログラム (Graphics program)**

3. **3次元マップの生成部とユーザインタフェイス部**

2で得られた学習結果をC++言語で書かれた3次元オブジェクト指向ツールキット IRIS Inventor で作成されたマップ生成部で処理して表示する。さらに文書検索やブラウジング OSF/Motif を用いて、文書表示用のテキストウィジェットを作成している。

以上がマップ生成におけるおおまかな流れである。

4.2 文書のブラウジング機能

- **点検索**

キーワードマップ上でマウスにより任意のノードを選択すると、そのノードは赤色の四角に囲まれることによりそれがもう既に選択されているということを示している。同時にリストウィジェットを持つウインドウが開く。そのノードにテキストが写像されていれば、そのテキスト群のタイトルが全てリストウィジェットに表示される。さらに、リストウィジェットにある項目を選び、ポップアップボタンを押すと、対応するテキストがテキストウィジェットに表示される (図6)。

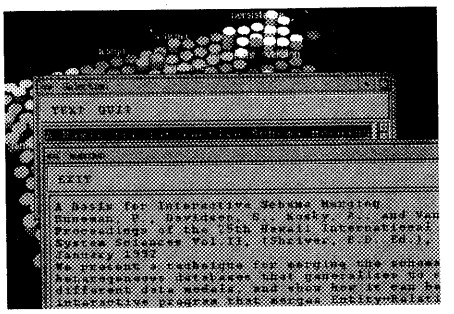

図 6: 点検索

図4上で"schema"と言う領域中のある点を選択し、図6のようにテキストのタイトルを含むリストウィジェットが現われる。さらに、そのリストウィジェット中の項目を選択することにより、文書を含むテキストウィジェットが出力され、一次検索は実現している。

- **領域検索**

マップ上のの領域に写像された文書を検索するには、マウスでその領域を代表するキーワードを選択することにより、そのキーワードは赤色の四角に囲まれそれが選択されているということを示している。同時にリストウィジェットを持つウインドウが開く。その領域に文書が写像されていれば、その文書群のタイトルが全てリストウィジェットに表示される。さらに、リストウィジェットにある項目を選び、ポップアップボタンを押すと、対応する文書がテキストウィジェットに表示される (図7)。

図6上で"design"を選択すると、図7のような全領域検索ができる。

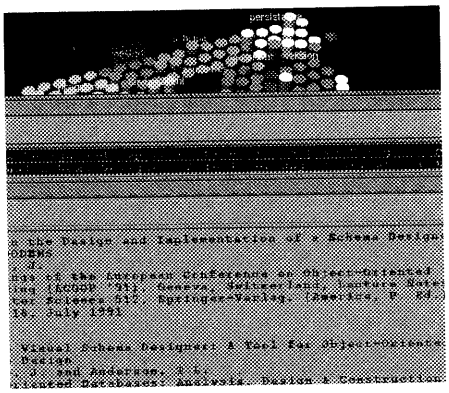

図 7: 領域検索

4.3 曖昧検索機能

2つの異なる領域 (異なる色づけにより区別されている) の接しているところのノードを選ぶことにより、両領域にも関連したテキストは選出される。つまり、両領域を代表するキーワードを含むようなテキストを検索することができる (図 8)。

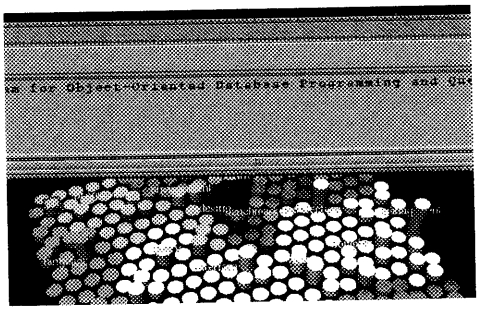

図 8: 曖昧検索

図 8のように、領域"language"と"programming"の境界線上にあるノードを選択することにより、"language"と"programming"の両方にも関連したオブジェクト指向関連の論文を検索することができる。

4.4 段階的な詳細化機能

- **再計算のない詳細化**

  マップ上の任意の領域を代表するキーワード (第1のキーワード) を選択することにより、新たなマップを作り出す。そのようなマップの上に新たなキーワード (ベクトル群中の出現頻度2番目高い単語) が現われる。このことは、すでに最初のマップ上に存在して、まだはっきり表現されていない情報をユーザにみせることができる。つまり、選択された領域上のいままでみえなかった、より詳しい情報をみることができる。このような詳細化はもう一度SOM法で学習することがなく、マップの多重表現だけで実現している (図 9)。

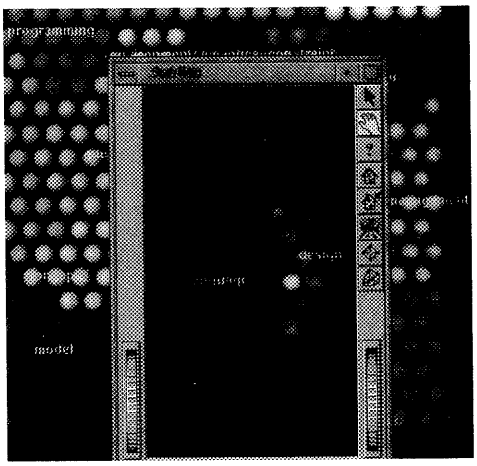

図 9: 再計算のない詳細化

図 9は計算のない検索を行なったもので、"programming"と言う領域を選ぶと、この領域に関するより詳しい情報が得られた。

- **再計算のある詳細化**

  出力マップは2次元空間上のもので、最初にテキストベクトルが不正生成されたり、学習過程中の学習回数などの悪影響などの問題により、ユーザが望む情報が表現できなかったり、表現の効果が悪かったりすることが起こりうる。この対策として、ある領域を選んで、その領域を代表するキーワードを取り除いて、領域に射影されたすべてのテキストをベクトル化しSOM法で再学習させ、新たなキーワードマップを作り出す。これにより1回目の学習で表現できなかった重要知識を表現することができる。

4.5 マップ/文書間の巡航操作機能

マウスで1回目の検索で得られたテキストウィジェット上のテキスト中の任意の単語 (キーワード) を選択すると、この単語を含む、マップ上に射影されたすべて

のテキストのタイトルは新たなリストウィジェットに現われる。同時にマップ上にこれらのテキストに対応するすべてのユニットは赤色の四角に囲まれる。このことにより、このような単語を含むテキストに対応するユニットは全マップ上にどれほどあるか一目でわかる。つまり、これらのテキストの分量はテキスト全体に対する相対的な量として知ることができる。また、リストウィジェット中に項目(タイトル)を選択することにより、対応するテキストを検索することができる。このような過程は再帰的な形式をとって、何回も実行することができる(図10)。

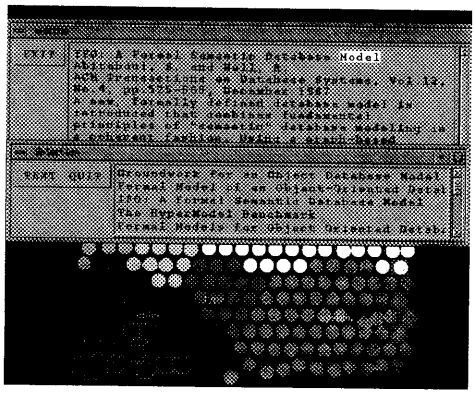

図 10: 巡航操作機能

図 10では、一次検索で得られたテキスト中で、"Model"という単語を指定すると、この単語を含むすべてのテキストが対応するノードはマップ上でマークされ、同時にそれらのテキストのタイトルが現われる。つまり、ハイパーテキストとしての検索ができる。

5 あとがき

本研究では、自己組織化マップをデータベースのブラウジングツール及び問い合わせインターフェースと見なして実際に情報を検索をするシステムにおいて、テキストとしてハイパーテキスト文書を自動分類するという機能を付加した。

また、実際に検索システムを試作し、データベース関連のハイパーテキスト文書の分類を試みた。

結果として利点をあげると次のものが挙げられる。

- 各領域のマップ上での高さをみることによってその領域に分類されているテキストの分量をテキスト全体に対する相対的な量として知ることができる見通しの良さ
- マップ上のキーワード同士の関連を見ながら検索できる柔軟性
- 曖昧検索と段階的な詳細化のよりよい実現
- マップ/文書間の巡航操作の実現により，ユーザが望む情報が得られる
- ハイパーテキスト文書を使って情報検索できるため、今後ネットワークワイドにこの機能を拡張できる

今後の課題として、

- アンカー定義語の重みづけアルゴリズムの改良
- キーワードを抽出する際に文脈を考慮した単語、熟語類の抽出方法の確立[4]
- 段階的詳細化に応じたベクトル生成方法の変更
- 3D マップ上での段階的な詳細化、とくに再計算を伴う詳細化機能の実現
- ネットワークワイドにおける情報検索ができるような機能を持たせ、WWW にこの機能を結合する
- 日本語のテキストへの対応

などが残されている。

参考文献

[1] 仁木 和久、田中 克己: 「ニューラルネットワーク技術の情報検索への応用」: 人工知能学会誌、vol.10 NO.1、1995 年 1 月

[2] 銭 晴、史 欣、田中 克己: 「自己組織化マップと語彙索引を用いたデータベースの抽象化機構」: 情報処理学会データベースシステム研技報 99-22、pp.163-170、1994 年

[3] 史 欣: 「自己組織化マップを用いたデータベースの概覧および検索機構に関する研究」: 神戸大学大学院工学研究科計測工学専攻修士論文、1995 年 3 月

[4] 豊浦 潤、有田 英一: 「単語の連想関係に基づく意味マップによるテキスト表現の試み」: 情報処理学会第 45 回全国大会、文冊 3、pp.247-248、1991 年

[5] Teuvo Kohonen: *The Self-Organizing Map*: Proceeding of the IEEE Vol.78, No.9, pp.1464-1480, 1990 年

[6] Martin Brayan: *SGML AN AUTHOR'S GUIDE*: 監訳 山崎 俊一、訳 福島 誠: アスキー出版局、1991 年

[7] Conklin, J.: *Hypertext: An Introduction and Survey*: IEEE Computer Magazine, pp.17-41, 1987 年