データベースシステム 109－6
（1996．7．24）

自己組織化マップを用いたテキスト情報からの知識獲得

仲川 亜希　　小西 修

高知大学大学院理学研究科情報科学専攻
高知大学理学部情報科学科

Email:{nakagawa,konishi}@is.kochi-u.ac.jp

データベースの検索結果の集合や、与えられた問題世界の情報集合から、それらの集合の特徴を表す概念関係が視覚的に表示されれば、非常によいと思ったことはなかろうか。

本論文では、与えられた文献集合に対して、その要素であるキーワード(キーフレーズ)の共出現対に注目し、自己組織化マップ処理によるクラスタから概念関係を自動的に抽出し、視覚的に表示する方法について述べる。

Automated Knowledge Acquisition from Text Information using Self-Organizing Maps

Aki Nakagawa　　Osamu Konishi

Department of Information science, Kochi University

Email:{nakagawa,konishi}@is.kochi-u.ac.jp

We will expect that the conceptual relationship which represents the features of the sets from the results of database search or information set of a given problem world, can be visualized.

In this paper, we describe the approach that aims at the relations of term co-occurrence into a given text, classifies the set of the term pairs using self-organizing map, extracts automatically the conceptual relationship for a cluster, and then displays visually it.

1 はじめに

計算機の処理能力と、大容量記憶能力の発達とともにテキストのオンライン記憶が可能となり、テキストへのアクセス可能性と利用可能性の拡大がもたらされた。その結果、オンライン情報検索システムが発達した。利用可能な文献の数が指数関数的に増えるにつれ、自動分類を行うような、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。また、文献集合から、その内容全体を概覧できるデータベース構造を生成できるような機能も望まれている。[6]

キーワードに基づく類似度計算や索引生成、自然言語解析などのような従来のテキスト検索処理や記号処理を行わずに、大量のデータを取り扱う手法として、ニューラルネットワーク処理技術の情報検索への適用がある。[1]

本研究では、文献に共出現する用語の関係を考慮し、与えられた文献集合に対して、その要素であるキーワードの共出現語対に着目した。そして学習したニューラルネットワークから「知識」としてルールを抽出し、各クラスタ内の概念関係を自動的に抽出し、表示する試みを行った。

2 文献集合からの共出現語対の抽出

ある文脈において、一つの用語と共に出現するもう一つの用語の間には、概念において何らかの共通の世界(その文脈の主題)を持つと考えられる。[3] そこで、文献集合における共出現語に着目し一つの文献に共出現する用語の対を抽出する。この用語対に頻度情報による順序関係を持たせることによって概念の階層関係を導入する。

2.1 共出現語関係

用語間の概念の階層関係の情報を得るために同じ文献中に共出現する用語と用語の対(共出現語対)とその順序関係を求める。

定義 1 文献集合 $D=(D_1, D_2, \cdots, D_n)$
文献 D_i=文献標題、抄録、索引語
文献集合から抽出される用語を
$\mathrm{TERM_k}=(t_{1k}, t_{2k}, \cdots, t_{nk})$,(t_{ik}は文献D_iの用語) とすると、共出現語対は
$C(\mathrm{TERM_k}, \mathrm{TERM_h})$
$=\{[t_{ik}, t_{ih}] | t_{ik} \in D_i, t_{ih} \in D_i\}$
となる。

定義 2 $C(\mathrm{TERM_k}, \mathrm{TERM_h})$ に重みを付けるために、$\mathrm{TERM_k}$ と $\mathrm{TERM_h}$ の間の距離 (結合度) を次のような関数で与える。
$f(\mathrm{TERM_k}, \mathrm{TERM_h})$

$$= \frac{freq.ofC(\mathrm{TERM_k}, \mathrm{TERM_h})}{\sqrt{[freq.of\mathrm{TERM_k} \times freq.of\mathrm{TERM_h}]}}$$

ここで、$freq.of\mathrm{TERM_k}=\sum t_{ik}$
$freq.of\mathrm{TERM_h}=\sum t_{ih}$

定義 3 $C(\mathrm{TERM_k}, \mathrm{TERM_h})$ において、$freq.of\mathrm{TERM_k} > freq.of\mathrm{TERM_h}$ ならば、そのとき $\mathrm{TERM_k}$ は $\mathrm{TERM_h}$ よりも概念の上位関係にあるとする。
$freq.of\mathrm{TERM_k} \geq freq.of\mathrm{TERM_h}$

このように順序を有する共出現語対の二項関係を共出現語関係と呼ぶ。次にこの定義に基づいた共出現語関係の抽出手順を示す。

step1 出現頻度の降順にソートされた用語候補リストを準備し、頻度に応じて用語をいくつかの (6~7 の) クラスに分類する。

step2 用語候補リストを得た最初の検索結果の文献集合を対象に、用語候補リストの各用語を検索語とした検索を行なう。

step3 その検索結果の集合から用語候補リストと同様にキーワードを切り出し、頻度の降順にソートする。ここで、検索語となった用語 (頻度統計の第1位の用語)

入力ベクトル

	artificial intelligence	back -propagation	learning system	neural net	training	...
artificial intelligence	1.000	0.183	0.167	0.350	0.000	...
back -propagation	0.183	1.000	0.134	0.185	0.000	...
learning system	0.167	0.134	1.000	0.655	0.172	...
neural net	0.350	0.185	0.655	1.000	0.142	...
training	0.000	0.000	0.172	0.142	1.000	...
⋮	⋮	⋮	⋮	⋮	⋮	

図 1: 入力パターン例

とそれ以外の用語 (第 2 位以降からある頻度以上のものまで) との組み合わせが、共出現語対である。このとき、第 2 位以降の用語の頻度は第 1 位の用語との共出現回数を示している。

step4 用語候補リストの全ての用語について、**step2,3** を繰り返す。

step5 得られた共出現語対に対して、**定義 2** による結合度を計算する。

3 Kohonenの自己組織化マップ

3.1 自己組織化マップ

Kohonen の自己組織化 (特徴) マップ (Self-Organizing (Feature) Map) は、1990 年に T.Kohonen によって提案されたパラダイムであり、ベクトルで表される入力パターン間の位相関係を、学習アルゴリズムにより発見、分類して位相地図を組織化する 2 層のネットワークである。このときベクトルの各成分はパターンの要素に対応している。この結果得られた地図は、ネットワークに与えられたパターン間の自然な関係構造を表している。ネットワークは処理ユニットの入力層と競合層の組み合わせであり、教師なし学習により訓練される。

入力パターンは競合層で活性化されるユニットにより分類される。パターン間の類似は競合層のグリッド上の近さの関係に写される。訓練が終了した後、パターン関係やパターングループが競合層で観察される。 [2]

Kohonen の自己組織化マップのアルゴリズムは以下の通りである。

自己組織化アルゴリズム

step1 入力パターンを与える。

$E = [e_1, e_2, e_3, \ldots, e_n]$

step2 この入力から競合層の各ユニット i への結合の重みを与える。

$U_i = [u_{i1}, u_{i2}, \ldots, u_{in}]$

step3 その重みが入力パターンと最もよく一致する競合層のユニット c を定める。すなわち、ベクトル E と U_iの間の距離が最小となるものを探す。

$$\| E - U_c \| = \min_j \| E - U_i \|$$

$$= \sqrt{\sum_j (e_j - u_{ij})^2}$$

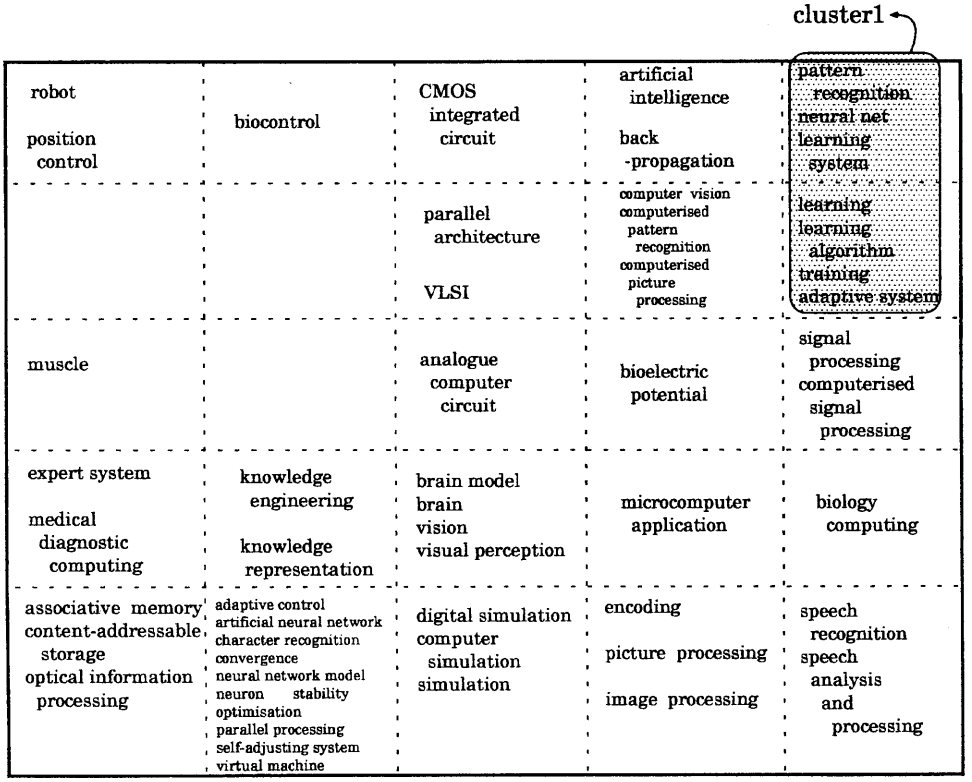

図 2: 競合層 5 × 5 、学習回数 10000 回、3.2により得られたコホーネンマップ

step4 このユニット i とその近傍 N_cで重みを調整して一致を増大させる。

$$\Delta u_{ij} = \begin{cases} \alpha(e_j - u_{ij}) & (i \in N_c) \\ 0 & (i \notin N_c) \end{cases}$$

また

$$u_{ij}^{new} = u_{ij}^{old} + \Delta u_{ij}$$

$$\alpha_t = \alpha_0 \left(1 - \frac{t}{T}\right)$$

ここで、αは学習率でその値は訓練が進むにつれて 0 へと減少していく。また、t は現在の訓練回数であり、T は行われるべき訓練の全回数である。

step5 学習反復が進むに連れて近傍のサイズと重みの変化の量を次第に減少させる。

3.2 文献集合からの自己組織化マップの生成

用語の特徴ベクトルを生成する際に、第 2 章で得られた共出現語関係を用いる。ここでは *neural net* という用語に注目し、 *neural net* と共出現語関係をもつキーワードと、さらにその用語と共出現語関係をもつキーワードのうち、出現頻度のある程度高いものを取り出したところ、 55 のキーワードが得られた。これらの共出現語対の結合度をキーワードの重みとして与える。入力パターン例の一部を図 1 に示す。こうして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。

実際に 5 × 5 のマップを用いてマップ生成を行なった。文献データベースの索引情報 (論文標題、抄録、Subject Index,Free Index) を上記の方法に乗っ取ってパターン化し、入力とした。学習回数は 10000 回、学習式には、

3.1 のものをそのまま用いた。この実験例では、表示の複雑さを避けるために、結合度のしきい値が 0.14 以上の、用語間の結合度が高い共出現語関係が処理の対象になっている。

4 クラスタリング

3.2 で得られた図 2 のようなコホーネンマップは関連の強い用語が近くにまとめられている。マップ上の用語は、その専門分野の概念体系を表している主要な用語でありこれらの代表的な用語に連なって他の多くの用語があると考えられる。そこで、これらの用語間の関係からその専門分野の知識構造を把握するために、マップ上の用語をクラスタリングする。

いくつかの出力ノードをひとつのあるクラスタにグループ化し、ルールを定義して概念関係を識別する。クラスタは、それぞれの出力ノードに関するルールの条件文により決定する。すなわち、条件文に同じ属性群を含むルールの出力ノードは同じクラスタにグループ化される。そして、第 5 章で述べる方法により、これらの概念 (クラスタ) は階層化される。

5 ルール抽出と概念階層

ルール抽出のアルゴリズムは、以下の通りである [5]。

ルール抽出のアルゴリズム

step1 すべての入力から競合層のあるユニット b_k への結合の重みの中で、最大のものを探す。
$W_{max} = (w_{1k}, w_{2k}, \ldots, w_{nk})$

step2 $W_{ik} \geq \beta W_{max}$ となる入力 a_i をすべて選ぶ。ここで β は 0 と 1 の間の定数とする。

step3 **step2** で選んだすべての入力を AND でつなぎ、ルールの条件文とする。例えば、**step2** で選ばれた入力を a_{i1}, a_{i2}, a_{i3} とするとルールは
IF (a_{i1} AND a_{i2} AND a_{i3}) THEN(b_k)
となり
(a_{i1} AND a_{i2} AND a_{i3}) $\Rightarrow$ (b_k)
と表す。

step4 **step3** をすべての出力に対して行い、初期ルール集合をつくる。

step5 条件文の中の入力属性が最も少ないルールを選ぶ。

step6 初期ルール集合に、**step5** で選んだ概念を代入する。

step7 代入がそれ以上できなくなるまで、**step5,6** を繰り返す。

step8 最終ルール集合から、概念階層をつくる。

このアルゴリズムを適用し、実際にクラスタ 1 の中の用語のルールを抽出し、階層化を行うと、以下のようになる。$\beta = 0.7$ とすると、初期ルール集合は、
(neural net AND learning system) $\Rightarrow$ (**learning system**)
(neural net AND pattern recognition) $\Rightarrow$ (**pattern recognition**)
(neural net) $\Rightarrow$ (**neural net**)
(learning AND learning system AND neural net) $\Rightarrow$ (**learning**)
(learning system AND learning algorithm) $\Rightarrow$ (**learning algorithm**)
(learning system AND training AND neural net) $\Rightarrow$ (**training**)
となり、代入を繰り返した後の最終ルール集合は、
(**learning system**) $\Rightarrow$ (**learning system**)
(**pattern algorithm**) $\Rightarrow$ (**pattern algorithm**)
(**neural net**) $\Rightarrow$ (**neural net**)
(learning AND **learning system**) $\Rightarrow$ (**learning**)
(learning algorithm AND **learning system**) $\Rightarrow$ (**learning algorithm**)
(training AND **learning system**) $\Rightarrow$

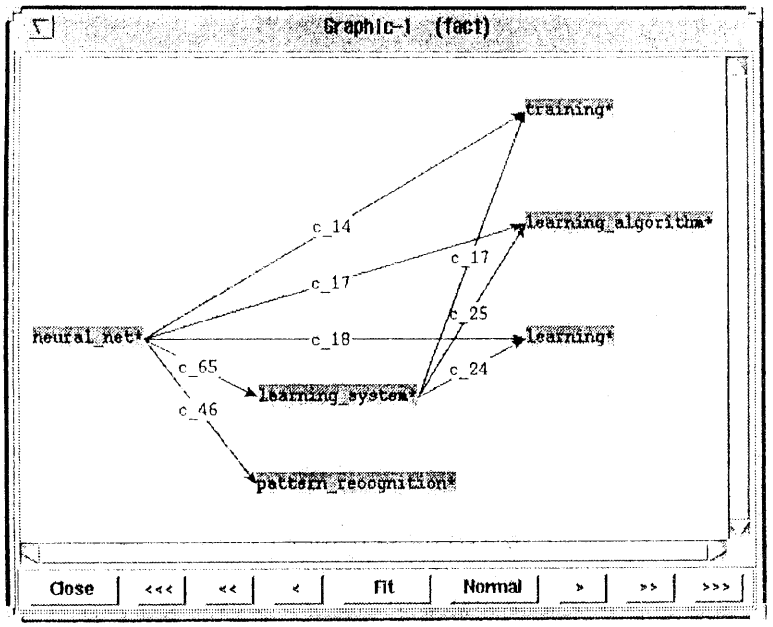

図 3: クラスタ 1 の概念階層

(training)

となる。これらの最終ルール集合から得られた概念階層を図 3に示す。矢印についている数字は、その共出現語対の結合度を表す。初めに、条件部分の属性の数が最小のものを選ぶことによって、概念階層の中で最も総括された概念が撰択される。階層の中で低い位置にある概念は、それを専門化したものである。

6 おわりに

本研究では、文献集合の要素であるキーワードの共出現語対に注目し、自己組織化マップ処理によるクラスタから知識としてルールを抽出し、概念関係を自動抽出し表示することを試みた。これは、データベースの検索結果の集合や、与えられた問題世界の情報集合から、それらの集合の特徴を表す概念体系を表示できるので、内容全体を概覧することができその分野の知識構造を把握するのを助けるアプリケーションに応用することができると考えられる。

参考文献

[1] 銭 晴、波多野賢治、田中克巳「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」: データベースシステム 104-6、pp41-48、1995 年

[2] J. ディホフ原著、桂井浩訳「ニューラルネットワークアーキテクチャ入門」、森北出版、1993 年

[3] 小西 修：自動構築型知識に基づく専門用語集形成システム、情報処理学会論文誌、vol 30、No2、pp179-189、1989 年

[4] Philib D.WASSERMAN 著、石井直広/塚田 稔共訳「ニューラルコンピューティング」、森北出版、1994 年

[5] S.Sestito & T.S.Dillon ; Automated Knowledge Acquisition, PRENTICE HALL,1994

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

[7] Xia Lin、Dagobert Soergel、Gary Marchionini、; A Self-organizing Semantic Map for Information Retrieval, : SIGIR 、pp262-269、1991

[8] 矢野洋嗣「自己組織化ニューラルネットワークアルゴリズムの解析」: 高知大学理学部情報科学科平成 6 年度、卒業研究