6E−4

# ニューラルネットワークを用いた自然語文書検索方法

牧　秀行　　松葉育雄
(株)日立製作所　システム開発研究所

## 1．緒言

従来の代表的な文書検索方法としてはキーワード検索とフルテキストサーチが挙げられる。キーワード検索は文書の内容を少数のキーワードで記述し、検索時にはキーワードのみを参照するため、キーワードとして登録されなかった情報が失われるという欠点がある。一方、フルテキストサーチは検索時に文全体を参照できるので、情報の欠落はないが、文の意味内容を反映した検索が困難であるという欠点がある。そこで、文の持つ情報を欠落させることなく、しかも、意味内容を反映した検索が可能な文書検索方法を、ニューラルネットワークを用いて実現する。

## 2．検索装置の概要

検索装置は自然語で書かれた文を検索対象文として保持している。そして、検索条件文字列を入力とし、各検索対象文の適合度を出力とする。適合度は検索対象文が検索条件をどの程度満足するかを表す数値である。基本方針として、検索条件文字列と類似した文字列を含む検索対象文には高い適合度が与えられるが、各検索対象文がどの程度の適合度をとるべきかの判断は人間の主観による。検索装置内の適合度計算部は検索条件文字列と検索対象文を比較し、適合度を計算する。この適合度計算部をニューラルネットワークを用いて構成し、人間の判断による適合度を教師として学習を行う。

## 3．特徴情報

検索条件文字列と検索対象文を比較する際に参照される情報を文字列、及び文の特徴情報と呼ぶことにする。本方法では文中の自立語の出現確度とこれら自立語間の構文上の距離を特徴情報とする。出現確度はそれぞれの単語が文中に出現するかどうかを表す数値である。構文上の距離は文中の任意の2単語間について定義され、以下のように求められる。まず、適当な方法により構文解析を行い、構文木を生成する。そして、この構文木上で、2つの単語を表す節を結ぶ経路上に存在する枝の数をこれらの単語間の構文上の距離とする。構文上の距離を特徴情報として用いることにより、検索条件として指定された単語が偶然同時に文中に出現することに起因する検索雑音を防ぐことができる。

## 4．適合度計算部の構成

適合度計算部は検索条件文字列と検索対象文の特徴情報から各検索対象文の適合度を計算する。特徴情報から人間の判断に近い適合度を求める関数は容易には決定できないので、適合度計算部をニューラルネットワークで構成し、教師付き学習によって、適合度を求める関数を決定する。

適合度計算部の構成を図1に示す。図中、二重線で描かれた4つの部分がニューラルネットワークで構成されている。このうち、類義語連想部は相互結合型ネットワークで構成され、検索対象文中に出現する単語について類義語の連想を行い、その情報を検索対象文に付加する。これによって、文中に出現していない単語に対しても、類義語が文中に出現していればある程度の出現確度が与えられ、類義語や異表記の使用

---

Text Retrieval Method for Natural Language Using Neural Networks
Hideyuki MAKI, Ikuo MATSUBA
Systems Development Laboratory, HITACHI, Ltd.

に起因する検索漏れを防ぐことができる。この相互結合型ネットワークは各単語の類義語情報を結合重みとして持っている。その他の3つの部分は多層型ネットワークで構成されている。

ニューラルネットワークの学習は誤差逆伝播法で行う。教師値は適合度統合部に提示される。出力誤差は適合度統合部から各ニューラルネットワークに伝播される。図中、破線で示したのが誤差の流れである。個々のネットワーク内だけでなく、ネットワーク間でも誤差逆伝播を行うことにより、適合度統合部に教師値を提示するだけで4つのネットワークの学習を行うことができる。

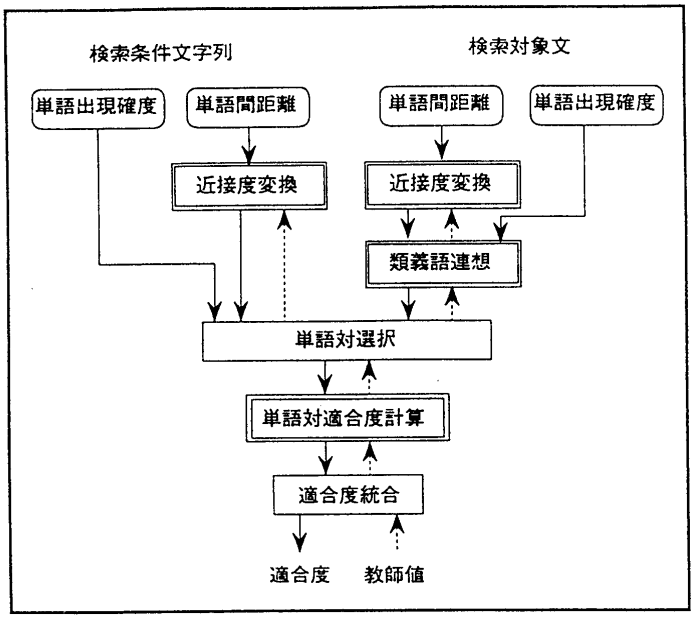


図1　適合度計算部の構成

## 5．実験

検索対象文の例を表1に示す。「高い建物」を検索条件文字列としたところ、1から5の文には低い適合度、6から10の文には高い適合度が出力され、人間の判断に近い検索結果が得られた。

## 6．結言

自然語文中の自立語に関する構文情報と類義語情報を用い、また、ニューラルネットワークの学習能力を利用することによって、意味内容をある程度反映し、人間の判断に近い検索結果を得た。

表1　検索対象文の例

| 文番号 | 検索対象文 | 教師値 | 適合度出力 |
|---|---|---|---|
| 1 | その建物は高い山の麓に建造された。 | 0 | 0.0028 |
| 2 | ビルの外はかなり温度が高い。 | 0 | 0.0085 |
| 3 | この家の持ち主は背の高い人です。 | 0 | 0.0065 |
| 4 | このあたりは土地の値段が高いので家を借りるのも大変である。 | 0 | 0.0065 |
| 5 | その建物のある町は最近若者の間で人気が高い。 | 0 | 0.043 |
| 6 | 向いの家があまりに高いので、日当たりが悪くなった。 | 1 | 0.99 |
| 7 | このビルはおそらくこの町で一番高い。 | 1 | 0.99 |
| 8 | 高い建物が建つと古都の景観が損なわれる。 | 1 | 0.99 |
| 9 | ここも高いが、まわりの建物の方がもっと高い。 | 1 | 0.95 |
| 10 | この程度のビルでも田舎ではかなり高い方です。 | 1 | 0.95 |

## 参考文献

1）Rumelhart,D.E. et al. : Parallel Distributed Processing Vol.1, 2, The MIT Press (1986)

2）長尾真　監修：日本語情報処理，電子情報通信学会 (1984)