4R−10

自己組織型ニューラルネットワークによるドキュメントの自動分類に関する一考察

小船 隆一
三菱電機(株)中央研究所

1. はじめに

近年、情報検索の分野におけるニュラルネットワークへの関心が高まってきている。本報告では、情報の検索への1つのアプローチとして、ドキュメントをキーワードの集合として扱い、そのキーワードの出現の統計的な性質を自己組織型のニューラルネットワークを用いて抽出し、ドキュメントの自動分類を行う方法について考察する。
検索を行うために、そのドキュメントの性質を文章の中のキーになる単語の出現のパターンによって特定することとした。そしてそのキーワードのパターンをauto-associative learningを行うニューラルネットを用いて分類させ、その分類をドキュメントの分類とし、また、そのネットワークの連想機能を用いたキーワードによるドキュメントの検索を実験した。

2. ドキュメントとキーワード

先にも述べたように、ドキュメントの性質を文章の中のキーになる単語の出現のパターンによって特定することとした。また、キーワードとなる単語に対して、その意味を示す情報は与えない。この条件から、我々は、情報検索に関して単語の意味に関する単純化を行った。

単語の意味：検索を行うときに検索空間をしぼるための制約条件。
制約条件：検索空間にバイアスをかける他の入力、すなわち、単語から連想される他の単語。

ここで、検索空間をニュラルネットワークの状態空間とすると、単語の意味はニューラルネットワークに学習されたキーワードのパターンの中の単語の相関関係に対応する。この単語の意味づけが人間の認識している意味と近くなることを期待している。

3. 自己組織化、連想、検索

単語の意味の単純化から、ドキュメントの自動分類には、キーワードのパターン以外の情報なしでキーワード間の相関関係を抽出し、その相関によるキーワード間の連想を行える連想メモリが必要である。そして、自己組織型のニューラルネットワークと連想メモリは深いつながりを持っており[1]、自然な流れとしてドキュメントの自動分類への自己組織型ニューラルネットワークによるアプローチをとることになった。
連想学習においては、入力パターンの中の個々のデータは明示的な意味を持たず、特徴的なパターンの学習によってデータが相互に条件づけを行うことにより、入力パターンがエンコードされ、個々のデータはその条件づけによりその意味が最終的に規定される。
ニューラルネットワークは、基本的には連想学習を行うものであるが、ドキュメントの分類においては、次に示す条件を満足するものを必要とする。

1) autoassociative mapping
学習データの間に入出力の対応はない。ドキュメントのすべてのキーワードは、基本的にすべての他のキーワードと相互に条件づけを持つ可能性がある。

2) online learning
学習アルゴリズムは、逐次的な学習データの入力を可能とする。膨大なドキュメントを逐次的に蓄積する場合には、全ドキュメントのパターンをそのつど学習することは不可能である。

検索は、検索キーに対応するキーワードのパターンを与えてニューラルネットワークの

Document Classification by Self-Organized Neural Network
Ryuichi Kobune
Mitsubishi Electric Corp.

associative recallを行い、検索キーから連想されたキーワードのパターンの連想の強度によってドキュメントの検索空間が絞り込まれることにより処理される。

4. 目標とするシステムのイメージ

ここで、ドキュメントの自動分類を行う検索システムの最終的なイメージをしめす。最終的なシステムは、ユーザごとにユーザの語彙のバックグラウンドに対応した、キーワード連想メモリ、格納されたドキュメントから抽出されたキーワードの意味を持つ連想メモリ、キーワードのパターンとドキュメントを結び付けるドキュメントベースからなる。

ユーザごとのキーワード連想メモリは、ユーザの認識している意味とドキュメントから抽出された単純化された意味とのマッピングを行い、ユーザごとにドキュメントベースに対する異なった視点を与えることを可能にする。

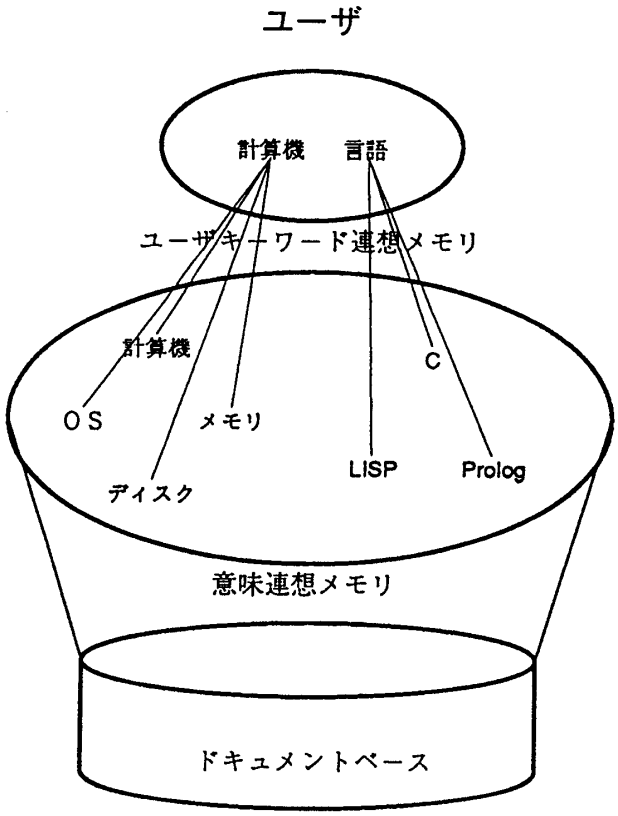

図1　最終的なシステムイメージ

5. 実験

日経ＡＩのエキスパートシステム、Prolog、LISPに関する記事を集め、その記事の中の単語を切り出して、キーワードとし、ネットワークモデルとしてＡＲＴ１(Binary Adaptive Resonace Theory) [2]を用いて自動分類の実験を行った。

キーワードに対応する入力ノード数：
　　約６００
分類に対応するクラスのノード数：
　　３～１０
記事の数：それぞれの内容に関して７つづつ
各記事のキーワード数：
　　５０

記事からのキーワードの切り出しは、人間の手でそれぞれの記事の中で重要であると思われる単語を取り出すことによって、統計的にキーワードの集合に対してバイアスが加わっているため、分類結果は人間による分類に近くなった。

キーワードの共起関係の強さが人間の意識している単語の意味に会った形に学習されたキーワードの集合を用いた検索では、与える検索キーに対してヒットするドキュメントは予想できる範囲にある。

6. 今後の課題

さまざまな側面からドキュメントを分類するためのドキュメントからのキー情報の抽出の仕方を実験し、柔軟な検索を可能にするためのそれぞれの分類を行うネットワークの相互作用の制御方法の検討が必要である。また、検索のコンテクストをいかににネットワークに反映させるかも問題となる。

参考文献

[1] Kohonen, T: Self-Organization and Associative Memory, Springer-Verlag.

[2] Simpson, P. K.: ARTIFICIAL NEURAL SYSTEMS: Foundations, Paradigms, Applications, and Implementations, PERGAMON PRESS.