

自己組織化手法を用いた情報のフィルタリング†

5Q-3

○ 田中 栄治 宮崎 哲夫 古城 則道‡

学習情報通信システム研究所§

はじめに

近年のコンピュータとそれを取り巻くネットワーク技術の発展は、誰もが世界中の様々な情報(リソース)にアクセスすることを可能とした。しかしその一方で、情報の氾濫によりユーザーにとって本当に意味のある(必要な)情報にたどり着けない、情報の過負荷(Information Overload)問題を引き起こしつつある。この問題が、創造的思考のいわば予備段階である情報収集に過程において、大きな障害となる可能性を秘めていることが指摘されている。

本研究の目的は、ユーザーが行う自由探索学習を支援し、情報の組織化を自動的に行う情報検索システムの開発である。本システムは、ユーザーが Web 上での NetSurfing から得た情報を自己組織化手法を用いて自動的にクラスタリングを行う機能を有する。また、ユーザーによる分類教示を学習する機能を備える事により、ユーザーの意図に応じた分類システムの再構築を可能としている (Fig. 1)。

本稿では HTML 文書をリソースとした情報探索支援システムについて考察し、本システムの概要とその実験結果について報告する。

本システムの概要

キーワードによりベクトル表現された文書に対し、教師なし学習で知られる自己組織化マップを用いて、文書分類や概念構造の抽出を行う提案が多くなされている。この手法の優れた所は、直接視覚的な表現に結びつくこと、分類の粒度の調整が容易なこと、及びその高速学習性にある。

我々は以前に、自己組織化クラスタリングによる文書の自動分類法、および文書の意味空間へのマッピング法についての報告を行っている [1], [2]。そこで、それらの技術を応用し、3層からなるニューラルネットワーク構造の文書分類アルゴリズムを導入した。また、自己組織化クラスタリングで用いる文書の間表現自体を修正することにより、ユーザーの分類に適応させるという学習メカニズムを適用した。

中間表現の修正は、誤差逆伝播により入力層と中間層の間の結合係数に相当する単語表現から中間表現への変換行列の修正を伴う。本システムの中間表現は明確な意味を

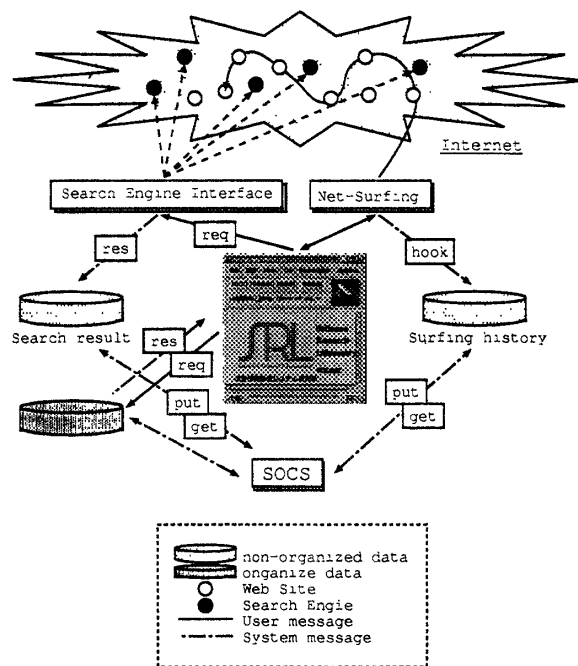


Fig. 1 プロトタイプ (SOCS) の概要

持たないが、このことにより、ユーザーの集めた文書の増加につれて、ユーザーの教示に従い作り上げていく事が可能となる。

文書自動分類システムの構成

提案するシステムの構成を Fig. 2 に示す。

システムはユーザーの教示に適応するネットワーク学習機構の付加機能を持つ分類エンジンで構成される。分類エンジンは、新規文書からキーワードの有無をもとに単語表現を生成、中間表現へのマッピング、そして自己組織化クラスタリングによるカテゴリ決定処理を行う。カテゴリ決定は、逐次的な構築アルゴリズムであり、しかもカテゴリ数が可変の自己組織化ニューラルネットワークを用いる。これにより、ユーザーによるカテゴリの生成や変更が随時可能である対話的な機構を実現できる。

修正学習機能は、ユーザーが分類エンジンが作成した分類に対して修正を施したときに起動し、ユーザーの与えた分類を実現するように中間表現生成の適応学習を行う。

本システムは、基本的には文書のベクトル表現があれば、自己組織化クラスタリングにより、実行される。中間表現

†An Approach on Information Filtering with Self-Organizing Mechanism

‡Eiji TANAKA, Tetsuo MIYAZAKI, Norimichi KOJO

§Software Research Laboratory

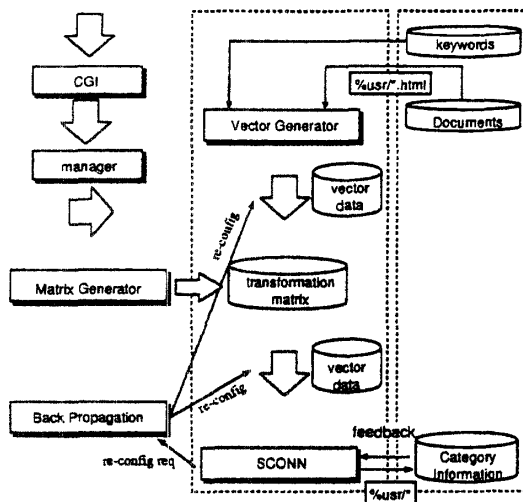


Fig. 2 プロトタイプ処理手順

の役割は、ユーザの要求に適応できる柔軟性をシステムに提供するとともに、次元を縮約することにより分類例を瞬時学習する高速化を可能にすることにある。この中間表現は、後に述べるように、学習者の分類例から動的に変更される。

文書の自動分類プロセス

本システムでの分類プロセスでは、カテゴリ数が可変の逐次的分類を実行し、かつ各時点での分類結果が最適化条件を満たしているようにするため、閾値設定によるクラスタ生成と K-means 法 [3] をベースとするアルゴリズムを用いた。

本システムで用いた文書自動分類アルゴリズムは次の特徴を持つ。

- 逐次的なアルゴリズムであること
- 予め定義されたカテゴリを持たないアルゴリズムであること
- カテゴリ数は可変であること
- K-means 法に従うアルゴリズムであること
- カテゴリ数や分類の粒度を制御できること

分類例の教示による分類機構の学習

ここで、システムの与えた分類に対して、ユーザが修正する場合を考える。ユーザの修正をユーザによるシステムへの分類例の教示とみなすと、例からの学習法を用いて分類機構の学習アルゴリズムを作ることができる。

本システムの分類アルゴリズムの中には、分類を制御するパラメータとして閾値がある。従って、ユーザの分類修正に対して、第一に考えなければならないことは、閾値の変更で対応できないかどうかである。また、我々の分類ア

ルゴリズムでは、文書の入力順序により異なる分類結果を作る。これは、K-means 法による結果が初期配置に依存するという良く知られた事実に対応している。すなわち、分類結果の最適解は一つとは限らない。従って、ユーザの分類修正への対応として、以下の処理を行う。

- ユーザの指定した配置が最適解のとき、各文書が閾値内であれば終了、閾値外のものがあれば閾値を更新して終了
- 最適解でなければ以下に述べる分類機構の変更を行う

ここで、ユーザの変更に対する適応として中間表現の修正を考える。中間表現の修正は、単語表現から中間表現への変換行列の修正で表現される。この修正学習を、誤差逆伝播の手法に従い以下のように定式化する。そのため、分類機構を Fig. 3 のような 3 層のネットワークとして捉える。

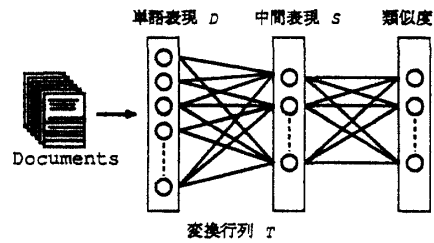


Fig. 3 3層ニューラルネットワーク

おわりに

本研究では、計算機の発達に伴って生じた情報の過負荷問題を、計算機自身によって解決させるために、自己組織化手法を用いた情報のフィルタリングシステムを構築した。

逐次的なクラスタリング手法とユーザからの教示による分類例の教示機構により、ユーザが行う自由探索学習の学習履歴を分類することができた。

今後の課題として、まず単語の自動抽出機構やキーワードラベルの自動生成 / 修正機構の追加や各処理 / アルゴリズムの最適化、ユーザインターフェイスの改善などが挙げられる。

参考文献

- [1] 宮崎哲夫, 田中栄治, 古城則道: 分類例からの分類視点の学習機構, 1997 年 電子情報通信学会ソサイエティ大会, D-5-2 (1997).
- [2] 田中栄治, 宮崎哲夫, 古城則道: 情報探索支援システムの構築 (2), 信学技報, ET96-124, pp.81-86 (1997).
- [3] 河口至商: 多変量解析入門, 森北出版 (1978).