自己組織化マップによるWWW日本語情報検索システムの評価結果

中村順一　甲斐郷子　村井幸一

九州工業大学情報工学部

1　はじめに

インターネットの発達に伴い，WWW(World-Wide Web)の利用が急速に進んでいる．Web上には多種多様な情報提供ページが日々刻々と登録，更新されている．そこでWeb情報を検索するには，カテゴリーごとに分野別に整理されたイエローページや，robotにより得られたデータベースを利用した全文検索ページ等が利用されている．

これに対して筆者らは，別の観点からWeb情報の検索を支援するため，Web上の任意のページからリンクで辿れる一定数の情報を自動的に整理し，利用者が目的の情報に簡単に辿りつけるようにするWeb情報整理システム[1]の作成を行なっている．これは，特定のサーバ中の情報をカテゴリに分類するものであって，情報整理の方法はT.Kohonen[2]が提案した自己組織化マップの手法を用いている．これまでの中尾[1]が作成したシステムは情報整理の対象が英語のページのみであった．それを，日本語のページに対しても実行できるように改良を行なった．本稿では，システムの概要と，評価実験の結果を述べる．[3]

2　自己組織化マップ

情報整理の手法には，Kohonenが考案した多次元ベクトル化した情報群を二次元マップに整理して配置する学習モデルである自己組織化マップを用いた．まずHTMLファイル中に含まれる単語の種類と数を利用して個々のファイルを多次元ベクトル化することにより，特徴を数値化する[4]．これを入力パターンとし，これと同次元のベクトルを持つユニットの集合である二次元マップとの間で繰り返し学習を行なうことで，マップの各ユニットに特徴を持たせる[5]．

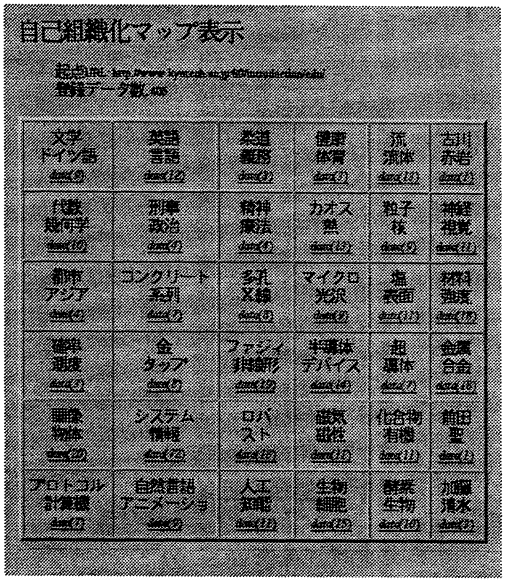

図1: 結果表示画面

学習が終わると，各HTMLファイルは自分の特徴(入力パターン)と最も似た特徴(パターン)を持つユニットに配置されるため，結果として情報内容の近いものどうしがマップの一部分にまとまる．その結果，ユーザーは欲しい情報がありそうなユニット，及びその近傍のユニットを見ることで，目的の情報に近い情報を収得できる．

整理結果は図1のようにHTML形式のファイルとして出力される．マップの各ユニットにはそのユニットの最大の特徴を表す単語の上位二つ，及びそのユニットに配置されたページの数が表示されている．ここでユーザが目的の情報に関するページが配置されていそうなユニットをクリックすると，今度はそのユニットに配

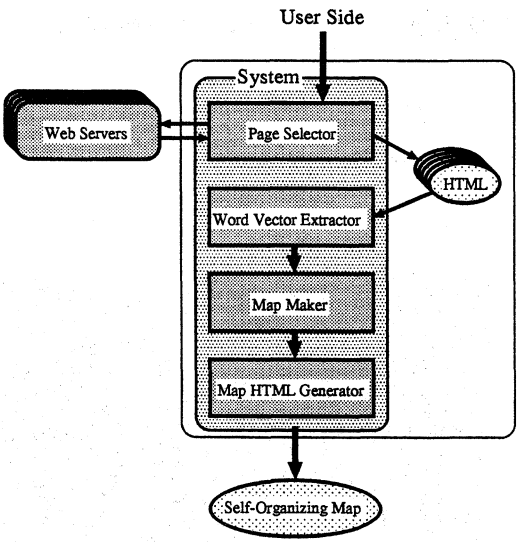

図 2: システム構成

置されたページのタイトル一覧が表示されているページに切り替わる.ユーザはリストされたタイトルから目的の情報に関するタイトルを選び,そのタイトルの部分をクリックすることで,目的の情報にアクセスすることができる.

3 システムの構成

システムは図2に示すように,整理する情報をユーザからの指示により Web 空間から取得し,有用な HTML を選択する "Page Selector", Page Selector で選択した HTML を使用されている単語を元に多次元ベクトルに数値化する "Word Vector Extractor",自己組織化マップ [2][4][5] 学習アルゴリズムを利用することにより多次元ベクトル化した HTML を二次元マップに整理する "Map Maker",結果の二次元マップを HTML 形式に変換し出力する "Map HTML Generator" の 4 つのモジュールで構成した

なお日本語のページも実行できるように,改良を行なった. HTML から単語を抽出する際に, 日本語形態素解析システム JUMAN[6] により単語の品詞分類を行ない, その結果より普通名詞と固有名詞を抜き出し利用単語とするようにした.

4 システムの評価

自己組織化マップの学習アルゴリズムでは,マップの大きさ,学習回数,次元数などのパラメータを決定する必要がある.中尾 [1] の評価実験では,学習時間と情報整理の精度の観点から,各パラメータの有効な値を求めた.その結果によると,マップサイズは,一つのユニットに配置される情報数が平均 2〜5 個になるようなマップが望ましいことが予想でき,おおよそ 6 × 6〜10 × 10 であるという結果となった.また,学習回数では,得られたデータのみでは十分な評価が行なえなかったが,100 回で十分であることが予想できた.

本稿ではシステムの日本語化を行なったので,同様に学習回数,マップサイズを変えて実験を行ない,自己組織化マップの整理結果を評価した.

4.1 評価実験 1

システムの日本語への対応を行なったことより,自己組織化マップの学習成果を調べるため,九工大「教育と研究」のページから辿れる 406 ページに対して,学習回数,マップの大きさを変えることで整理結果がどのように変わるか評価した.「教育と研究」のページは九工大の各教官の講義, 研究内容の紹介のページであり,統一されてフォーマットで書かれており,講座ごとにカテゴリーに分かれているため,評価しやすいと考え,このページを選んだ. [3]

4.1.1 学習回数

実験はマップサイズを 10 × 10 と一定にして,学習回数を 20 〜 200 回と変えていき,情報工学部の各教官のページが学科ごとに, マップのどのユニットに配置されていくか調べた.

知能情報工学科の教官の配置状況を図3に示す.図中の数値は,ユニットに配置された情報数である.図より,学習を増やしていけばいくほど,マップの一部に収束して行くのが確認できた.学習回数が少ない時には, 1 ユニットに多くの情

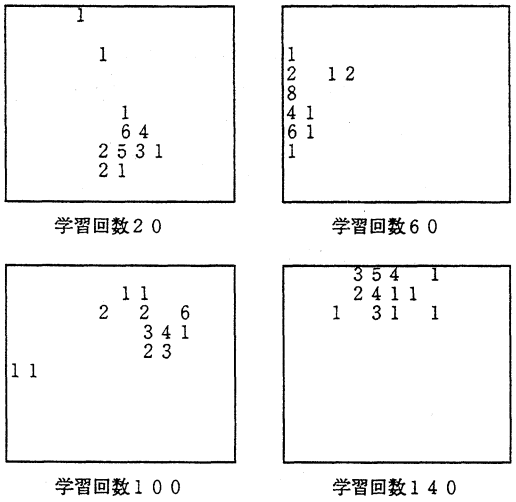

図 3: 知能情報の教官の配置状況

報が配置されたものや, 全く情報が配置されてないユニットがあり, キーワードに合致しないノイズ情報がかなり見られた. それが学習回数が大きくなるにつれ, マップ全体に広がって配置されて行き, 1ユニットの情報数も減ることで, 的確なキーワードのユニットの元に配置されて行った. またカテゴリーごとの情報がマップの一部分に集まって配置されていき, ユニットの属性を表すキーワードも, 学習が増えるにつれ, より, そのユニットの特徴を的確に表すものへ変わっていった.

この結果では, 学習回数が100回を過ぎたぐらいから, マップ全体に大きな変化が見られなくなってきており, 学習回数は100～150回程度あれば有効な整理結果が得られるのではないかと考える.

4.1.2 マップサイズ

マップの大きさが小さくなればユニット数も減り, 1ユニットに配置される情報数が増える. そうなると, ユニットの属性を表すキーワードが配置された情報の上位カテゴリーを示すものが現れて来ると期待できる. それを調べる目的で実験した. 学習回数を200回と一定にして, マップの大きさを10×10～4×4と小さくして行き, 知能情報の各教官のページが配置されたユニットのキーワードの変化を調べた.

図1は6×6の結果であった. 結果は10×10でユニット数が多いとより具体的な単語が多く見られ, それが, マップサイズが小さくなるにつれ, より上位カテゴリーを示すような単語に収束して行った. 例えば, "自然言語" と "アニメーション" は最初は別々のユニットであったが, 6×6で同じユニットに配置されている. [1]最終的には "システム, 情報" という情報工学全般を表すような単語を示すユニットにほとんどのページが配置された. 知能情報学科以外も, 人文社会学科等の他のカテゴリーも同様な傾向を示した.

4.2 評価実験2

比較的フォーマットが統一されたページに対する自己組織化アルゴリズムの有効性は確認できたが, 一般に Web ページは様々な書かれ方をしている. そこで, 一般的なページでも情報整理がうまく行なえるか調べるため, 早稲田のサーチエンジン「千里眼」[7]でキーワード検索を行なった結果のページに対して, 情報整理を行なった.

4.2.1 実行結果

実行結果の一例として, キーワード「料理」で検索した結果の197ページに対し学習回数を200回と一定にし, マップサイズを8×8～4×4と変えることで, マップの各ユニットを表す単語がどのように変化するか調べた.

4×4の結果を図4に示す. 結果は, 評価実験1のような, 明確な分類は見分けにくかったが, ユニット全体的には, カテゴリーごとにまとまって配置された. 例えば, 「農業」というカテゴリーで分類できる14ページの, マップサイズごとのユニットの配置数は表1のようになった. マップ

[1]筆者の所属する講座では, 物語文章からのアニメーションの生成の研究を行なっている. 中村は, この内, 自然言語の部分を担当し, 別の教官がアニメーション部分を担当している. このため, "自然言語" と "アニメーション" が最初は別ユニットであったものが次に同じユニットに配置された.

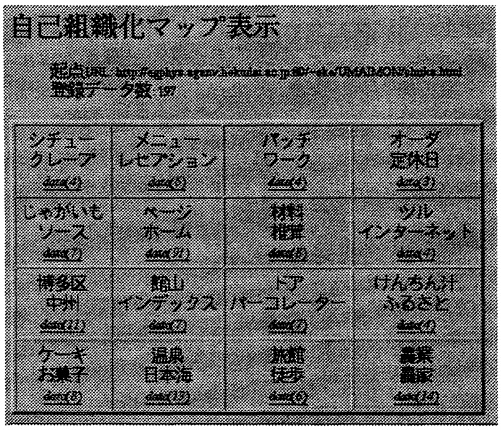

図 4: 「料理」での整理結果

サイズ	ユニットの単語	ページ数
8×8	連絡先, 農業	8
	農協, 農家	5
	農家, 農協	1
6×6	連絡先, 農業	8
	農協, 農家	6
4×4	農業, 農家	14

表 1: ユニットの配置数の変化

サイズが小さくなるにつれ, ユニット数も減少し, 最後は 1 つのユニットに全て配置された. ユニットの属性を表す単語も, よりそのユニットの代表的な単語にまとまっていった.

また, 約半数のページ (91 ページ) は最終的に 4×4 で, 単語が”ページ, ホーム”である 1 ユニットに配置された. しかし, マップサイズが小さくなるにつれ, ユニットのキーワードの単語に合致しないで配置されたページもかなり見られるようになった. これは, Web ページの性格上, 情報提供者が自由にページを作成でき, 1 ページには「料理」以外の情報もある場合があるため, うまく配置できなかったと考える.

5 まとめ

本研究では, WWW 空間上の情報から目的の情報を探すのが困難である現状に対応するために, 自己組織化マップというニューラルネットを用いた Web 情報整理システムを作成した. また, 自己組織化マップの学習過程で利用するパラメータであるマップの大きさ, 学習回数について情報ソースを変えて評価実験を行なった. その結果より自己組織化マップの有効性を確認できた.

参考文献

[1] 中村 順一, 中尾 学: “自己組織化マップを利用した Web 情報整理システムの作成と評価”, 言語処理学会第 2 回年次大会, pp.425-428(1996)

[2] T.Kohonen: “The Self-Organizing Map”, Proceedings of the IEEE, Vol.78, No.9, pp.1464-1480 (1990).

[3] 中村 順一, 甲斐 郷子, 村井 幸一: “自己組織化マップを用いた WWW 情報検索システムの評価”, 自然言語シンポジウム「大規模資源と自然言語処理」, 電子情報通信学会「言語理解とコミュニケーション」(1996),

[4] 有田 英一, 安井 照昌, 津高 新一郎: “単語集合の自動構造化機能を持つ「情報散策」方式”, 電子情報通信学会技術研究報告, 95-NLC-17 (1995).

[5] 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究報告, 99-DB-22 (1994).

[6] 松本裕治, 伝 康晴, 宇津呂 武仁, 妙木 裕, 長尾 真: 日本語形態素解析システム JUMAN 使用説明書 (1994), 奈良先端科学技術大学院大学.

[7] 千里眼 : http://www.info.waseda.ac.jp/search.html