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

中村 順一　中尾 学
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, 最近 WWW(World-Wide Web)[4] の利用が急速に進んでおり, 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. 情報内容も, 初期段階では研究機関や学校機関, 情報通信系の会社などの一部が自己紹介するだけのページが大半を占めていたが, 現在では音楽情報やイベント紹介といった趣味・娯楽情報ページ, 地域紹介ページ, 個人のページ, さらにはインターネットショッピングのページなど多種多様になっている.

ユーザは最近まで, このような莫大な Web 上の情報空間から全体像を把握し目的の情報にたどりつくために, robot[6] により得られたデータベースを利用した情報検索ページ (Search Engine) や, あらかじめ情報が分野別に整理されたページ (Yellow Page) を利用してきた. しかし, 現在, Web 空間は爆発的に広がりつつあり, これらのデータベースを利用しても全体像が把握できなくなっている. そこで本研究では, 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 情報整理システムの作成を行なった.

2 システムの構成

システムは図1に示すように, 整理する情報を 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 を多次元ベクトルに数値化する "Word Vector Extractor", 自己組織化マップ [5][1][3][2] という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

2.1 整理する情報の選択

本研究では情報整理の対象を, Web 空間 の情報のうち HTTP(Hypertext Transfar Protocol) で辿れる HTML(Hypertext Markup Language) ファイルとしている. ユーザは, この対象の中から整理する情報の範囲と量を指定する. システムは指定されたページから辿れる HTML を Web 空間から収集する. しかし収集されたこれらの HTML は, それぞれ質的・量的にも多種多様であり, それらすべてに対して整理を行なうことは得策ではない. そこで, システムではあらかじめ HTML に対して制約を設け, 質的・量的に情報の少ないものを削除する. 具体的には, 例えば「ページのサイズが 0.5KB 未満のものは整理対象から削除する」や「意味的に階層構造を持つページ群に対し上位のページに戻るためのリンクは辿る必要がない」といった制約を設けている.

2.2 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. まず情報中に含まれる単語の種類と数を利用して個々の情報を多次元ベクトル化することにより, 特徴を数値化する. 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. 具体的には, それぞれの入力パターンについて, もっとも近いベクトルパターンを持つユニット及びその近隣のユニットを入力パターンに近づけるという作業を繰り返し行なうことにより, 距離の近いユニッ

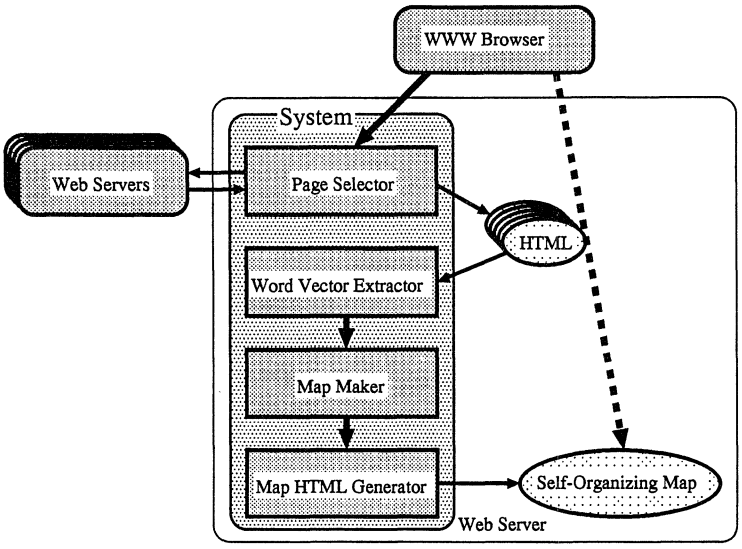

図 1: システム構成

トどうしのパターンが似てくる. 学習が終わると, 各情報は自分の特徴(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.

2.3 ユーザインターフェイス

システムはWWWのもつ機能であるCGI(Common Gateway Interface) [7] を利用することにより Web サーバ上に実現した. ユーザは Netscape のような Web ブラウザを利用することでシステムの起動及び出力整理結果の確認を行なうことができる.

起動時には, ユーザは起動するためのページ(図2)にアクセスし, Web 空間のどのあたりの HTML をどれくらい整理したいかという探索条件を入力する. 指定できるパラメータには, 探索起点の URL(Uniform Resource Locater), 探索 HTML 数, 起点からの探索の深さ, ドメイン名による探索範囲の限定がある.

起動してしばらく待つとブラウザは自動的に結果のページにアクセスする. 最初のページには, 図3(a)のような二次元マップが表示される.

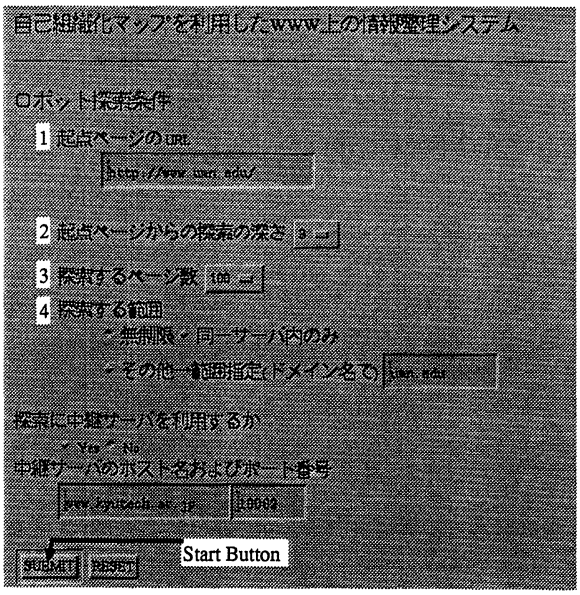

図 2: 入力画面

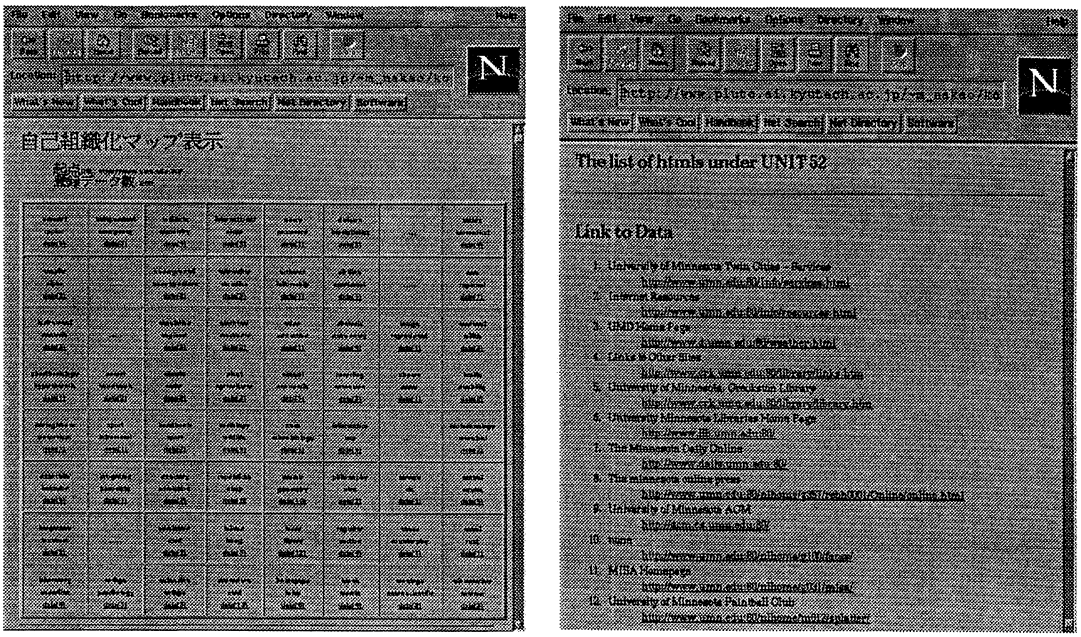

図 3: 結果表示画面

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及びそのユニットに配置された HTML 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 HTML 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置された HTML のタイトル一覧が表示されているページに切り替わる (図 3(b). 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 評価実験

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. これらは学習時間や情報整理の精度に大きく影響を及ぼすが, 本システムは対話的に利用できることを目標としているため, 時間と精度の両方の面で適度な値を求めなければならない. そこで, これらのパラメータそれぞれについて値を変化させて実験を行ない, 評価した. 評価の元になるデータはミネソタ大学のホームページ <URL:http://www.umn.edu/> から辿れる 200 ページとした.

3.1 マップの大きさについて

まずマップの大きさについて, 他のパラメータをすべて一定にしておき, マップの大きさを 4×4 〜 12×12 と変化させ, それぞれ生成された二次元マップを評価した. 評価は, それぞれ配置された情報の内容がそのユニットの最大属性(単語) と合致するかどうか, つまり的確に配置されたかどうかを判断することにより行なった. その結果, マップの大きさに比例して的確に配置された情報数は一次元的に増加した. 5×5 と 6×6 は, マップに表示された単語が抽象的なものが多かったため, 例外的に的確に配置された情報数が多くなった. また 5×5 以下では, 一つのユニットに何十もの情報が配置され, とても整理されているとはいえなかった. さらに 10×10 以上になると, 的確に配置されている情報が一つもないユニットが増えてしまい, ユーザが情報を探すのに苦慮するだろうと予想できた. これらから, 適度なマップの大きさについては本実験においては $6 \times$

6～10 × 10 であるといえた. またこの結果より一般的には, 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た.

3.2 学習回数について

次に学習回数についてもマップの大きさ同様, 他のパラメータを一定にしておき, 学習回数を 20～120 と変化させ, それぞれ生成された二次元マップを評価した. 評価は, 特定の話題に関する情報をグループとし, 同グループの情報が二次元マップ上でどのくらいうまくまとまっているかを判断することにより行なった. その結果, 情報により学習回数が増えるにつれまとまるものもあれば, 同グループの情報が二箇所に分散していったもの, あまりまとまらないままのものもあった. 本実験に関しては, 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.

3.3 利用単語数について

最後に利用単語数についても, 他のパラメータを一定にしておき, 1 情報あたりの利用単語数を 5～すべてと変化させ, それぞれ生成された二次元マップを評価した. 評価は, それぞれ配置された情報の内容がそのユニットの最大属性(単語) と合致するかどうか, つまり的確に配置されたかどうかを判断することにより行なった. その結果は, 的確に配置された情報の数は生成された二次元マップに現れるそれぞれのユニットの最大属性になった単語に左右され, 利用単語数に比例した結果は得られなかった. しかし, 単語数 5 であっても一定の分類ができることが分かった.

4 おわりに

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, 利用単語数について適度な値を探るための実験, 評価を行なった. しかし, 最適な値について十分な結論を得るには, さらに情報ソースをかえて実験を行なう必要がある.

参考文献

[1] 阿江 忠 : “もうひとつのニューラルネット学習法ー自己組織化”, 共立出版 bit, Vol.24, No.9-11 (1992).

[2] 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).

[3] 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).

[4] T.Berners-Lee, R.Cailliau, N.Pellow, A. Secret: “The World Wide Web Initiative”, In proceedings of INET '93, Internet Society (1993).

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

[6] M.Koster: “World Wide Web Robots, Wanderers, and Spiders”, <URL:http://info.webcrawler.com/mak/projects/robots/robots.html>.

[7] NCSA HTTPd Development Team: “The Common Gateway Interface”, <URL:http://hoohoo.ncsa.uiuc.edu/cgi/>.