論　文

マルチ Web ロボットによるユーザの興味を反映した情報収集

梛野　憲克[†]　　山田　誠二[†]

Information Gathering Based on User's Interest by Multiple Web Robots

Norikatsu NAGINO[†] and Seiji YAMADA[†]

あらまし　WWW の普及に伴い，WWW 上で情報収集を行うことが一般的になってきた．検索エンジンを用いて情報収集する方法以外に，個人の興味に応じた最新の Web ページを個人の計算機上に収集することも重要である．本研究では，ユーザの興味を反映した情報収集を行う PWM（Personal Web Map）システムを提案する．PWM は，ユーザにとって興味のある Web ページのデータベースであり，ユーザとインタラクティブに構成する．Web ロボットは，ユーザが入力したキーワードをもとに，関連する Web ページを収集する．このとき，ユーザが任意の時点で収集状況を把握するために Web ロボットに一様な情報収集をさせ，より詳細な情報がほしい領域を集中的に情報収集させるために，収集された Web ページを把握するための密度黒板を階層的に構成し，Web ロボットを制御する．収集状況は，自己組織化ネットワークにより構成される 2D マップとしてユーザに提示され，ユーザはより詳細な情報を知りたいクラスタを選択することで，PWM システムにフィードバックを与える．最後に，PWM システムの有効性を実験により示す．

キーワード　WWW での情報収集，マルチ Web ロボット，任意時間制御，自己組織化マップ，HCI

1. まえがき

近年，計算機資源の低コスト化などにより，個人的に興味がある Web ページを個人の計算機上に収集することが可能になってきている．このようなデータベースは，最初からユーザの興味を反映して収集された Web ページで構成されるので，後にその Web ページを活用する際に，既にフィルタリングがされていることから質の高い情報検索が期待できる．そこで，本研究では，ユーザ個人の興味を反映した Web ページの収集を実現するシステムを提案し，実験的にその有効性を確認する [12].

前述のような情報収集の一つの方法として，AltaVista, Yahoo，goo などに代表される検索エンジンのヒットリストに，ユーザの興味に応じたフィルタリングを施し，得られる URL に対応する Web ページを実際に取得する方法が考えられる．しかし，検索エンジンが用いる Web ページのデータベースは，実在する Web ページ集合の一部分でしかない．実在する Web ページの絶対数の変化は大きいが，文献 [5] によると，1999 年 2 月時点で公開されている Web ページは約 8 億であり，その約 16%を超える Web ページを蓄積する検索エンジンはなく，その被覆率は 1997 年 12 月から減少している．また，代表的な検索エンジンにより新しい Web ページ，または更新された Web ページが得られるまでに 1 か月以上かかる [5]．そのため，検索エンジンの結果には，ユーザが本当にほしい Web ページが検索結果に現れない場合や，得られる URL に対応する Web ページは既に存在しない場合が多々ある．ユーザの興味に応じて，その検索エンジンの結果にフィルタリングを施したものは，更に一部分でしかない．また，検索エンジンが検索結果をユーザの検索意図どおりにフィルタリングすることは困難なため，提示される Web ページは多くの関連のないページを含んでいるのが現状である．よって，本研究では，このような検索エンジンを用いた方法は採用しない．

ユーザ個人の Web ページ収集において重要な点の一つは，ユーザの興味をできるだけ反映することである．そのための有効な機能として，ユーザが Web ページの収集過程をモニターすることが可能であり，更に任意の時点でその収集を制御できる必要がある．

[†] 東京工業大学大学院総合理工学研究科知能システム科学専攻，横浜市
CISS, IGSSE, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama-shi, 226–8502 Japan

本研究では，効率化のためにマルチ Web ロボットを用いた Web ページの収集を行い，情報収集の過程を PWM（Personal Web Map）と呼ばれる表現で記述する．PWM は，収集された Web ページデータベースと入力キーワードに関連した Web ページがどのように分布して収集されたかを表す密度黒板から構成される．更に，PWM システムは，情報収集過程を自己組織化マップ SOM（Self-Organizing Map）[4] を用いて，2 次元マップを表示でき，ユーザはより詳しい情報を知りたいクラスタを選択することにより，任意の時点で Web ロボットの情報収集を制御できる．このようなユーザの任意の時点での割込みを可能にするために，PWM システムは，いつ停止されても各キーワードについて偏りのない一様な情報収集をする必要がある．このような情報収集は，一般に検索エンジンのデータベース作成に使われているマルチ Web ロボットの横型探索では不可能であるため，本研究では，密度黒板を参照しながら一様な情報収集を行うようにマルチ Web ロボットを制御する手法である任意時間制御を提案する．そして，最後に，PWM システムを用いることにより，ユーザの興味を反映した情報収集が行われることを実験により示す．

分散型マルチ Web ロボットを用いて負荷分散を用いた研究 [2] など，検索エンジンのデータベース構築の効率化を目指す様々な研究が行われている．しかし，効率化されたとしても，サービスの対象が一般の不特定多数のユーザであるため，やはり Web ロボットは網羅的な探索を行って Web ページを収集する．よって，ユーザの興味を反映したデータベースを構築することはできない．

Fish Search [1] は，人工生命の技術を利用して，関連する情報を集める探索アルゴリズムである．探索を行うエージェントは，質問単語に関係ある Web ページの集合を同定する．また，InfoSpiders [8] は，Fish Search を改良した，動的な WWW 環境に適応するアルゴリズムである．残念ながら，これらの研究は，ユーザとのインタラクションがなく，PWM システムのような，任意時間においてユーザによる明示的な Web ロボットの操作が可能な枠組みにはなっていない．

WebWatcher [3] や Letizia [6] は，ユーザのブラウジング履歴からユーザの好みを学習し，ユーザの興味に関連する Web ページへのリンクを提示することができる．これらのシステムは，ユーザの興味を学習することができるが，その目的はユーザのブラウジングの支援であり，情報収集ではない．

ブックマークエージェント [9] は，ユーザによりフィルタリングされた URL 情報であるブックマークファイルを複数のユーザで共有するシステムである．Web ページ間の類似度を計算して，他のユーザのブックマークファイル中にある，類似したページの URL をエージェントが教えてくれる．ユーザの興味によるフィルタリングが可能だが，エージェントが自ら情報収集するわけではない点が本研究と異なる．

ナビゲーションプランニング [13] は，目標概念を理解するために必要な Web ページの系列をプランニングにより自動生成する．索引付けアルゴリズムとタグ構造を利用して，Web ページからオペレータを自動生成しながらプランニングが行われる．ナビゲーションプランニングでは，システムが自ら情報収集を行うが，ユーザとのインタラクションはなく，ユーザの興味を対話的に取り込むことができない．

2. Web ロボットによる情報収集

図 1 に，システム全体の概観を示す．以下にシステムの動作をユーザ，PWM，Web ロボットの三つに分けて示す．

- ユーザ：ユーザは，はじめに自分の興味に関連するキーワードをシステムに入力する．ユーザは，Web

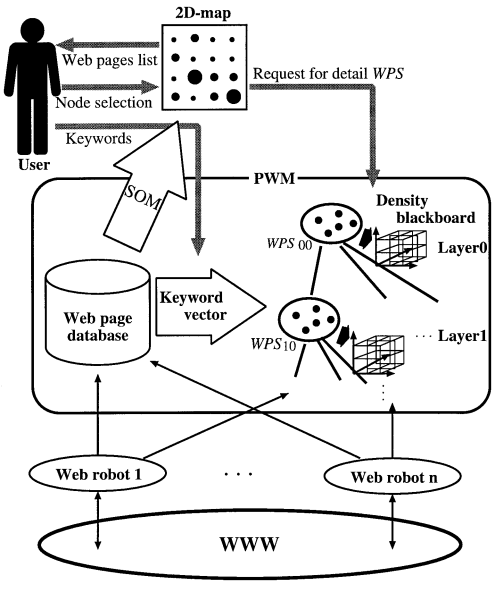


図 1　システムの概観
Fig. 1　Overview of the system.

ロボットが収集した Web ページの情報を得るためにいつでもシステムに指示を与えることができる．WPS に含まれる Web ページは SOM により 2D マップとして視覚化される．ユーザはそれを参照し，自分の興味に関係するクラスタを選択する．

- PWM：ユーザから与えられたキーワードをもとに **2.1**で述べる WPS に対応する密度黒板の各軸を設定し，キーワードの個数を次元数とする多次元空間を構成する．Web ロボットにより集められた Web ページの情報はデータベースに登録され，含まれているキーワードの頻度から密度黒板にプロットされる．ユーザが 2D マップ上のクラスタを一つ選択すると，そのクラスタに対応する Web ページ間の関連性をより詳細に表す，新たな WPS を構成する．
- Web ロボット：Web ロボットは WPS を参照し，収集する Web ページの URL を選択する．Web ロボットは複数で，非同期に Web ページを収集する．

2.1 PWM（Personal Web Map）

本研究では，ユーザの興味を反映した情報収集のための PWM（Personal Web Map）を提案する．図 1 の PWM は，収集された Web ページのデータベース，WPS，そして密度黒板から構成される．WPS は，収集された Web ページのキーワードベクトルの集合であり，密度黒板は，WPS 中の Web ページに含まれるキーワードの頻度の分布を表す．密度黒板は，収集された Web ページの特徴を把握するために用いられる．Web ロボットは密度黒板を参照し，次に収集する Web ページの URL を決定する．WPS_{00} は根となる最初の階層であり，ユーザからの入力によって与えられたキーワードに直接的に関連する Web ページを含む．ここで，WPS_{ij} は，階層 i における j 番目の WPS を意味する．

Web ページの収集過程で，ユーザの指示により現在の WPS 中の Web ページが SOM を用いて複数のクラスタに分類され，その結果が 2D マップとしてユーザに提示される．ユーザが WPS_{nj} の分割されたクラスタの一つを選択すると，階層 $n+1$ に選択されたクラスタについてより詳しく収集した Web ページによって $\mathrm{WPS}_{(n+1)k}$ が作られる．このようにして，ユーザが興味のある Web ページの詳細な情報を獲得するために，PWM の構成を制御することが可能になる．

2.2 WPS の構成

各 Web ページに含まれるキーワードの出現頻度に基づいて WPS を構成する．Web ページの類似度を計る方法として，情報検索の分野では一般にベクトル空間モデル [11] が用いられる．Web ページ p のキーワードベクトル V_p を，このベクトル空間モデルに基づき式 (1) で定義する．

$$
\begin{aligned}
V_p &= (v_{p1}, v_{p2}, \cdots, v_{pN}) \\
&= \left(\frac{f(p,t_1)}{m(D,t_1)}, \frac{f(p,t_2)}{m(D,t_2)}, \cdots, \frac{f(p,t_N)}{m(D,t_N)} \right) \\
&\text{ここで } m(D,t) = \max_{p \in D} f(p,t) \qquad (1)
\end{aligned}
$$

Web ページデータベース D は，Web ページ p の集合を表し，v_{pi} は，ユーザからの入力キーワード t_i の Web ページ p における頻度 $f(p,t_i)$ を正規化したものであり，t_i の重要度を表す．また，N は，ユーザから入力されたキーワード数である．v_{pi} は，Web ページデータベース D の Web ページの各キーワードの最大値で正規化され，値の範囲は $[0,1]$ となる．このキーワードベクトル V_p の集合が WPS である．

2.3 密度黒板

ユーザとインタラクティブに PWM を構成するにあたって重要なことは，ユーザが PWM の構成プロセスにいつでも割り込めることである．ユーザは情報収集が終了するまで待ちその結果を得るのではなく，必要なときにいつでもある程度の結果が得られることが重要である．ユーザが任意時間において Web ロボットを制御することにより，より効率的に興味を反映することができる．

一般の Web ロボットのように網羅的な横型探索により情報収集をすると，収集される Web ページの特徴に偏りができる．この理由として以下の二つが考えられる．一つは，ある Web ページからリンクされている Web ページは，もとの Web ページに類似している場合が多いため，大域的に見ると収集された Web ページの特徴には偏りが生じる．もう一つは，実際に存在する Web ページの特徴に偏りがあるためである．例えば，流行の話題に関連する Web ページは多く存在する．この場合，ある時点ではユーザの興味に関連する Web ページが極端に少ない場合が存在し，その場合 Web ロボットがたどることができるリンクの数がほとんどない可能性がある．

収集された情報に偏りのある PWM と一様な PWM の密度黒板の例を図 2 に示す．

システムに二つのキーワードが与えられ，X 軸，Y 軸はそれぞれキーワード 1，キーワード 2 の関係を表

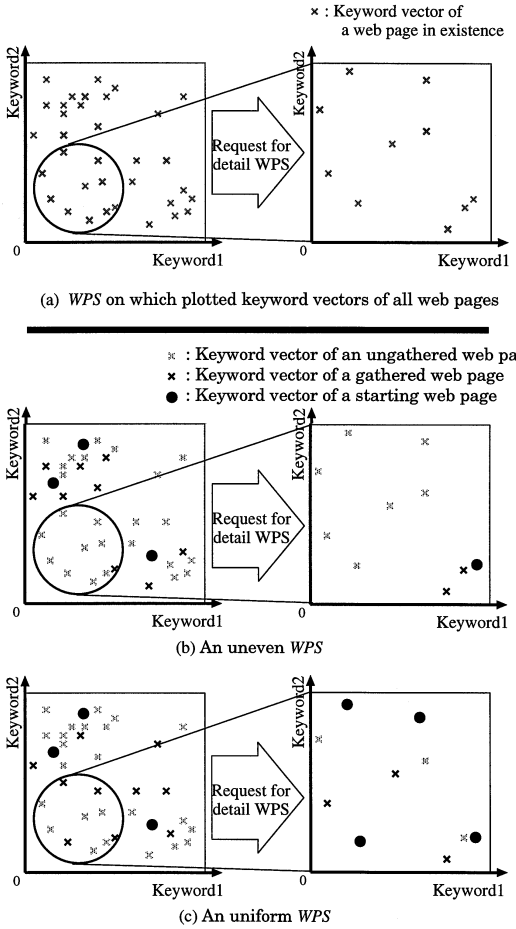


図 2　偏りのある WPS と一様な WPS
Fig. 2　An uneven WPS and an uniform WPS.

している．存在する Web ページをすべて収集し密度黒板にプロットした結果，図 2 (a) のようになったとする．左図の黒丸で囲まれた部分（クラスタ）にユーザは興味をもっていると仮定し，右図はその部分を詳細に表した新たな WPS である．ここで図 2 (b) のように収集される Web ページの特徴に偏りが存在する場合，興味に関係する Web ページがほとんど収集されておらず，たどるべきリンクがほとんどない．一方図 2 (c) では新たな WPS の始点となる Web ページが存在し，興味に関連する Web ページを集めやすい．

これを実現するために，WPS を図 3 のようにメッシュ状に分割して，密度黒板を構成する．キーワード t 個のベクトル空間における各軸の範囲 $[0, 1]$ を s 個の区画に区切る場合，t 次元空間の中に合計で s^t 個の t

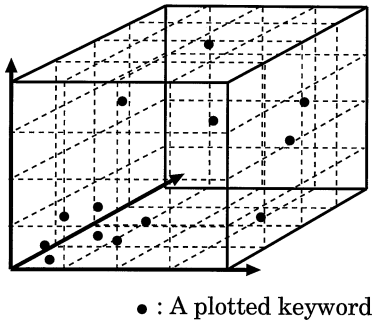


図 3　密 度 黒 板
Fig. 3　Density blackboard.

次元の区画ができる．Web ページデータベース D からキーワードベクトルに基づき密度黒板にプロットした後，各区画の密度を計算する．ここで，$1 \leqq s \leqq 10$ の範囲で s を変化させて密度を調べたところ，**2. 4**で示す任意時間制御が **4.** で示す一般的な Web ロボット探索に比べて，$s = 5$ で最も一様になることが実験的にわかった．これは，s が大きすぎると区画が細かく，小さすぎると区画が粗くなり，たどるべきリンクが適切に選択されていないためであると考えられる．**4.**で示す実験では，$s = 5$ としている．

2. 4　任意時間制御

密度黒板を利用し，あまり収集されていない Web ページの特徴を決定し，その区画に含まれる Web ページのリンクをたどることにより WPS の一様性を実現する．つまり，密度の低い区画のページのリンクをたどると，類似のページが得られ，それが同一区画にプロットされると予想されることから，その区画の密度を上げることが期待できる．これは，ある Web ページのリンク先の Web ページは，リンク元の Web ページに類似しているという仮定に基づいている．検索エンジンである WebCrawler の探索アルゴリズム [10] は，この仮定に基づいて設計されている．また，ARACHNID [7] では，あるページが他の Web ページからリンクされているとき，その Web ページがリンク元の Web ページと関連しているという条件付き確率を R, をランダムに Web ページをとってきたときに関連している確率を G とするとき，$R > G$ であることが実験により示されている．

ここでは，以上の方法により一様性を得るための Web ロボットの制御を，Web ロボットの任意時間制御と呼ぶ．Web ロボットの任意時間制御手続きを以下

に示す．

（1） WPS を参照し，最も密度の低い区画を決定する．

（2） その区画に含まれる Web ページからランダムに一つの Web ページ p を選択する．

（3） 選択された Web ページ p に含まれるリンクからランダムに一つのリンク l を選択する．

（4） リンク l に対する Web ページの html ファイルを取得する．

（5） html ファイルから式 (1) によりキーワードベクトルを生成し，密度黒板に書き込む．また，Web ページデータベースに html ファイルを登録する．

（6） （1）に戻る．

3. 2D マップを用いたユーザからのフィードバック

3.1 2D マップ

キーワードベクトルはユーザから入力されたキーワードの数の次元をもち，それは 4 以上の場合がある．よって，ユーザに PWM を把握しやすいように提示するためには，次元を 2 次元に落とすことが望ましい．よって，本研究では，SOM を用いて PWM の 2D マップを構成する．

まず，各ベクトル値がランダムである N 次元単位ベクトル（N はユーザが入力したキーワード数）を訓練ベクトルとして生成し，SOM に入力して学習を行う．SOM の学習の詳細は，文献 [4] を参照されたい．

学習後，収集された Web ページを複数のクラスタに分類するために，式 (2) のような各 Web ページに対する文書ベクトル E_p を SOM に入力する．そのときの勝者ノードが，その Web ページの属するクラスタを表すノードとなる．

$$E_p = \left(\frac{f(p,t_1)}{l(p)}, \frac{f(p,t_2)}{l(p)}, \cdots, \frac{f(p,t_N)}{l(p)} \right) \quad (2)$$

$$\text{ここで}, l(p) = \sqrt{\sum_{i=1}^{N} f(p,t_i)^2}$$

最後に，それぞれのキーワードに対応するクラスタにラベルを付ける．キーワード k に対する式 (3) のような単位キーワードベクトル E_k を SOM に入力する．

$$E_k = (e_1, e_2, \cdots, e_N) \quad (3)$$

$$e_i = \begin{cases} 1 & (i = k) \\ 0 & (i \neq k) \end{cases}$$

そして，勝者となった 2D マップ上のクラスタは，キーワードクラスタとしてそのキーワードをラベルとして表示し，各キーワードクラスタは異なる色（異なる原色など）で色付けする．二つのキーワードクラスタ間のクラスタの色は，それらのキーワードに対応する色の距離に応じた中間色を用いることによりグラデーションをかける．またクラスタの大きさは，そのクラスタに分類された文書ベクトルの数に比例した大きさにする．2D マップ上のクラスタをラベル付け，色付けすることにより，ユーザはキーワードに関連する Web ページがどのクラスタに分類されているかを容易に把握することができる．また，ランダムな単位ベクトル E_r で学習することにより，キーワードクラスタが分散されて 2D マップ上に配置されることにより，どのキーワードに関連する Web ページが多く収集されているかも容易に把握することができる．以上の手順によって構成された 2D マップの例を図 4 に示す．ここで，ユーザから入力されたキーワードは，“application”，“database”，“foil”，“learning”，“ontology”，“progol”，“relation” の 7 単語であった．

3.2 ユーザからのフィードバック

Web ページが分類された後，既に収集された Web

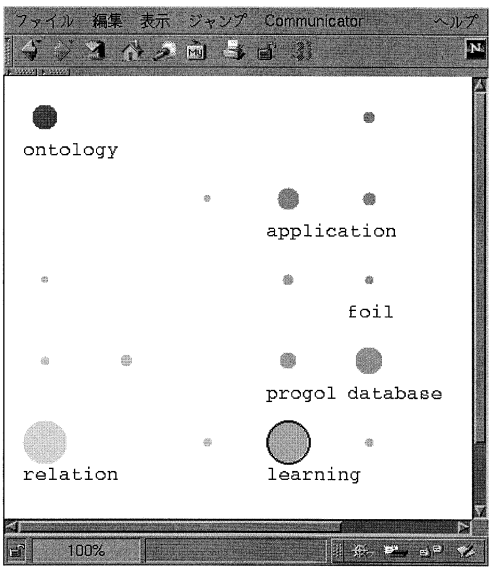


図 4 2D マップの例
Fig. 4 An example of 2D-map.

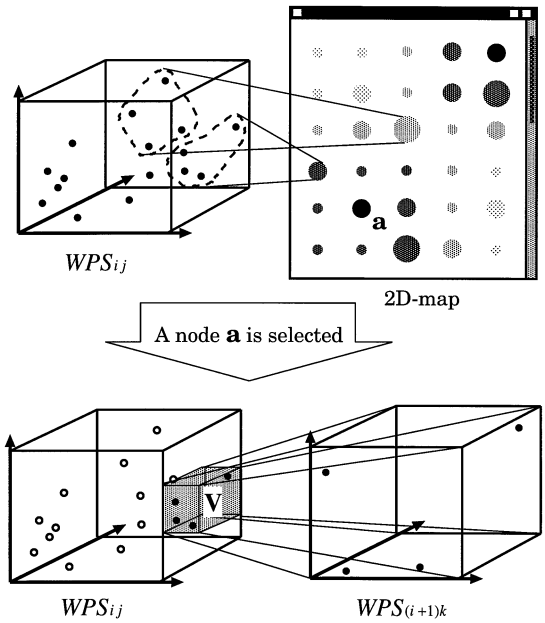


図 5　ユーザからのフィードバック
Fig. 5　Feedback from a user.

ページ間の関係を表した図 5 の上図のウィンドウがユーザに提示される．例えば，図 5 の上図のように，WPS_{ij} の 2D マップでノード a を選択するとそのノードに対応する WPS の部分空間が特定でき，その部分を詳細に表した新たな $\mathrm{WPS}_{(i+1)k}$ が作られる．左下図の V は，ユーザが選択したクラスタに含まれるキーワードベクトルをすべて包含し，$v_1^{\min} \leqq v_1 \leqq v_1^{\max}, \cdots, v_N^{\min} \leqq v_N \leqq v_N^{\max}$ を満たす領域である．ここで，$v_i^{\min} = \min_{p \in S} v(p,i)$，$v_i^{\max} = \max_{p \in S} v(p,i)$，$N$ は入力キーワードの数（各 WPS の次元数），$v_1, v_2, \cdots, v_N$ は，V の各軸の変数，S はユーザの選択したクラスタが含むキーワードベクトルの集合，$v(p,i)$ はキーワードベクトル p の i 番目の要素である．

詳細な WPS が新たに形成されたら，それを現在の WPS とすることを Web ロボットに伝える．これにより，ユーザが詳細な WPS の形成を指示した直後から Web ロボットはそれを参照し，ユーザの興味を反映した情報収集を行う．

4. 実　　験

任意時間制御の効果を調べるために，収集された Web ページの特徴における一様性，及び収集された興味のある Web ページの数について一般的な Web ロボット探索と比較する．ここで Web ロボット探索とは，網羅的な横型探索を意味し，効果的な制御を行わないものとする．Web ロボット探索の手順を以下に示す．

（1） Web ページのリストを $L = [\ ]$ で初期化．
（2） 始点となる Web ページ p_0 を現在の Web ページとし，p_0 に含まれるすべてのリンクを L に追加．
（3） L の先頭のページ p を L から削除．
（4） Web ページ p を取得．
（5） p に含まれるすべてのリンクを L の再後尾に追加．
（6） （3）に戻る．

PWM システムでは，Web ページにキーワードが一つも含まれていない場合，そのキーワードベクトルは $(0, 0, \cdots, 0)$ となる．このキーワードベクトルがプロットされる区画は，他の区画と比べて経験的に密度が高いため，キーワードを一つも含まない Web ページのリンクをたどることはほとんどない．実験を公平にするために，本研究では Web ロボット探索においてキーワードを一つも含まない Web ページのリンクはたどらないという制限を加えた．

Web ロボット探索と任意時間制御において，始点となる Web ページ p_0は，代表的なメタ検索エンジンである MetaCrawler にキーワードを入力し，それぞれ各単語を一つずつ入力した結果の上位 3 個，すべてのキーワードを and 検索した結果の上位 5 個，すべてのキーワードを or 検索した結果の上位 5 個からなる Web ページ集合とした．それぞれにおいて，四つの Web ロボットを用いた．

実験は，研究室のメンバー 10 人に行ってもらった．それぞれの被験者が選んだキーワードは表 1 のとおりである．

4.1　一様性の比較

図 6 は，横軸は収集された Web ページの数を，縦軸は密度黒板の各区画に含まれる Web ページの個数の標準偏差を表したものである．密度黒板は各軸を 5 等分した．図から，Web ロボット探索に比べて任意時間制御の方が標準偏差の値が低いことがわかる．これは，任意時間制御の方が各区画に含まれる Web ページの個数にばらつきが少ないことを意味し，一様に集められていることがわかる．

次に，この WPS の一様性が実際のユーザインタ

表 1　被験者の選んだキーワード
Table 1　Keywords given by subjects.

被験者番号	キーワード
p1	director, mystely, actress, movie, actor
p2	php3, postgreSQL, JDBC, JAVA
p3	genetic, programming, GP, intron, tree
p4	cribbage, crib, card, board, game, run, skunk, double
p5	linux, kernel, install, distribution, debian, dpkg
p6	agent, www, navigator, personalize, profile, AI
p7	pet, zoo, food, cat, fishing
p8	relation, learning, ontology, database, progol, foil, application
p9	multiple, robot, agent, cooperate, strategy
p10	slither, link, puzzle, faq, tips

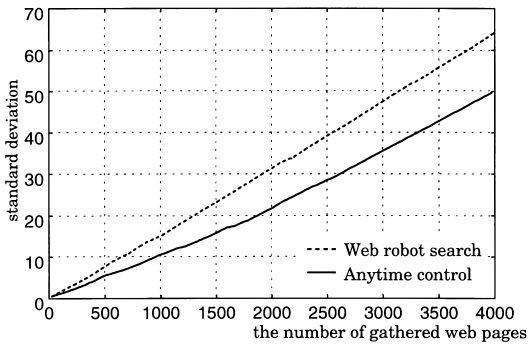


図 6　WPS における一様性の比較
Fig. 6　Uniformity on a WPS.

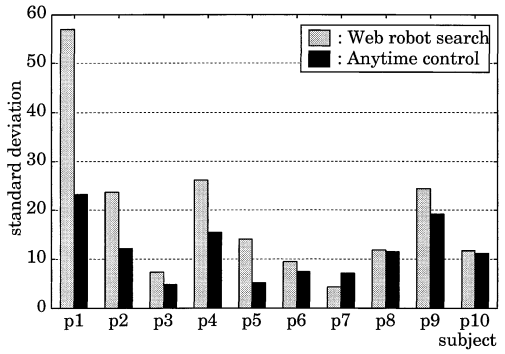


図 7　2D マップにおける一様性の比較
Fig. 7　Uniformity on a 2D-map.

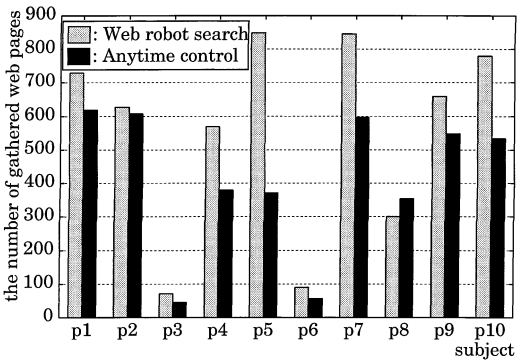


図 8　収集された Web ページの数（1）
Fig. 8　The number of gathered Web pages (1).

フェースとなる 2D マップ上の一様性にどのくらい影響を与えるかを調べた．各被験者に対し，Web ロボット制御と任意時間制御を用いてそれぞれ 3 時間の間 Web ページを収集し，2D マップ上の各クラスタに分類した．2D マップは，5×5 の 25 個のクラスタを 2 次元に配置した．図 7 の横軸は被験者（10 人）を，縦軸は 2D マップ上の各クラスタに分類された Web ページの個数の標準偏差を表したものである．この結果から，任意時間制御の方が Web ロボット探索よりも 2D マップ上の各クラスタに分類される Web ページの数に偏りが少ないことがわかる．

4.2　興味のある Web ページの比較

PWM システムが，ユーザにとって興味のある Web ページの収集に与える効果を実験で示す．被験者のキーワードをもとにして PWM の 1 階層目で 3 時間収集した．その後，2D マップ上のクラスタを一つ選択してもらい，2 階層目で再び 3 時間収集した．なお，収集には三つの Web ロボットを用いた．図 8 は，2 階層目で実際に収集された Web ページの数を表している．被験者 p3 と p6 は収集された Web ページが極端に少ない．経験的に，たどるリンク先のホストが混雑している場合，このような結果になった．

収集された Web ページがそれぞれ被験者にとって興味があるものであるかどうかを調べるために，収集された Web ページに一つの Web ページ当り，自分の興味に全く関係のない Web ページは 0 点，少し関係のある Web ページは 1 点，強く関係のあるページは 2 点として点数付けを行ってもらった．図 9 は，集められたすべての Web ページの合計点を計算し，各被験者の主観的な評価基準や評価した Web ページの絶対数によりそれぞれの合計点に大きな差がでるため，Web ロボット探索と任意時間制御のそれぞれの合計点を，収集されたすべての Web ページにそれぞれ満点

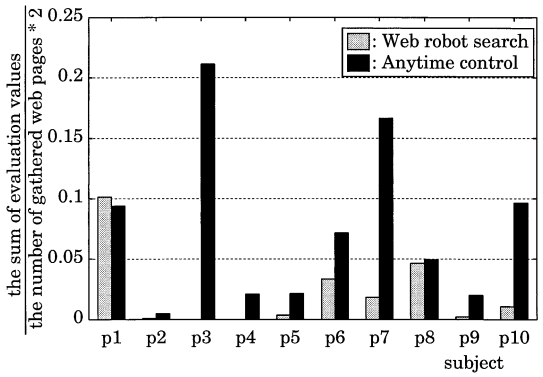


図 9　興味のある Web ページの評価（1）
Fig. 9　The relevance of gathered Web pages (1).

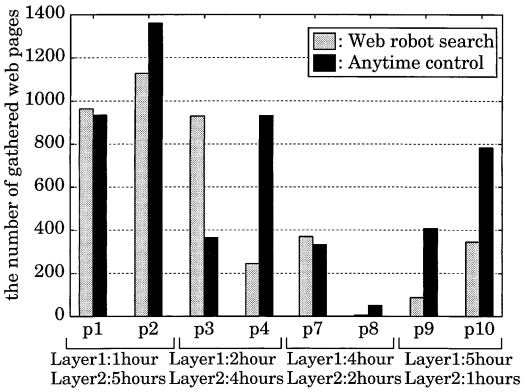


図 10　収集された Web ページの数（2）
Fig. 10　The number of gathered Web pages (2).

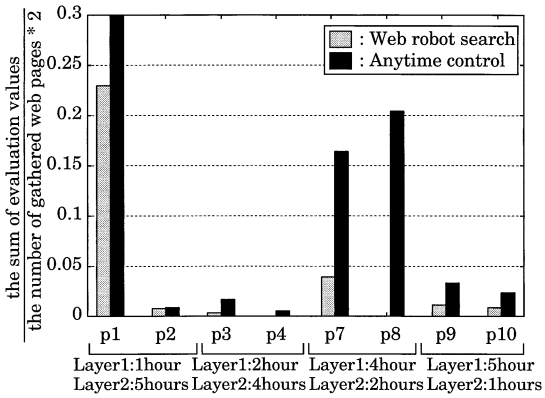


図 11　興味のある Web ページの評価（2）
Fig. 11　The relevance of gathered Web pages (2).

の 2 点をつけたときの合計点で割ることにより正規化し，グラフにしたものである．

図 9 は，被験者の主観的な評価であり，図 9 の p1 のように任意時間制御の方が悪い場合もあるが，ほとんどの被験者において，任意時間制御の方が収集された Web ページが少ないにもかかわらず，興味のある Web ページが多く収集されている．更に，図 8 から，密度計算などの処理のオーバヘッドや，ネットワークの負荷などにより，任意時間制御の方が Web ロボット探索よりも収集される Web ページ数は少ない場合でも，任意時間制御の方が興味のある Web ページが多く収集されていることがわかる．

また，任意時間における制御の効果を調べるために，被験者 8 名で 1 階層目と 2 階層目に割り当てる時間の比を変えて実験を行った．図 10，図 11 は，1 階層目と 2 階層目の収集時間に，それぞれ被験者 p1，p2 には 1 時間と 5 時間，被験者 p3，p4 には 2 時間と 4 時間，p7，p8 には 4 時間と 2 時間，被験者 p9，p10 には 5 時間と 1 時間を割り当てた場合の結果である．

収集の時間配分を変化させた場合でも，ほとんどの被験者において任意時間制御の方が Web ロボット探索よりも興味のある Web ページが多く収集されることがわかった．しかし，図 11 からわかるように，時間配分の変化による評価の合計点の違いに，最適な時間配分を決定し得るような特徴は見られなかった．

各階層での収集時間の配分を変えるとともに，WPS の階層の深さも変えて実験すべきであるが，実験を行った結果，ほとんどの場合において 3 階層目ではたどるべきリンクがほとんど残っていなかったため，2 階層までとした．

同一被験者の，Web ロボット探索と任意時間制御によって収集された Web ページの評価の合計点の間に極端な差が存在するのは，各被験者が主観によって Web ページを評価しているためだと考えられる．

以上の実験結果より，各階層において一様な情報収集を行うことにより任意時間におけるユーザの制御を可能にし，その制御によりユーザの興味を効果的に反映した効果的な情報収集が可能であることがわかった．

5. むすび

本研究では，PWM を用いてユーザが任意時間において Web ロボットを操作し，収集の早い段階からユーザの興味を反映することにより，無駄な Web ページの収集を省いた，効果的な情報収集を行う手法を提案

した．また，収集されたWebページが，実際にユーザの興味に関連しているかどうかを実験により調べた．実験の結果，ユーザにとって興味のあるWebページは，Webロボット探索に比べて任意時間制御の方が多く収集されることがわかった．本手法を用いることにより，検索エンジンのための一般的なWebロボット探索のように網羅的な情報収集をすることなく，効果的にユーザの興味に関連するWebページを取得できることを確認した．代表的な検索エンジンのデータベースに蓄積されたWebページ数の割合が，実在するWebページ数に比べて年々減少してきている今日，この方法は重要になると考えられる．

謝辞 本研究は，日本学術振興会 未来開拓学術研究推進事業研究プロジェクト「生物的適応システム」の支援のもとに行われました．記して，感謝致します．

文　献

[1] P. De Bra and R. Post, “Information retrieval in the World-Wide Web: Making client-based searching feasible,” Computer Networks and ISDN Systems, vol.27, no.2, pp.183–192, 1994.

[2] H. Yamana, K. Tamura, H. Kawano, S. Kamei, M. Harada, H. Nishimura, I. Asai, H. Kusumoto, Y. Shinoda, and Y. Muraoka, “Experiments of collecting WWW information using distributed WWW robots,” Proc. 21st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp.379–380, 1998.

[3] T. Joachims, D. Freitag, and T. Mitchell, “Webwatcher: A tour guide for the World Wide Web,” Proc. 15th Int. Joint Conf. on Artificial Intelligence, pp.770–775, 1997.

[4] T. Kohonen, Self-Organizing Maps, Springer, Berlin, Heidelberg, 1995. Second Extended Edition, 1997.

[5] S. Lawrence and L. Giles, “Accessibility and distribution of information on the Web,” Nature, vol.400, pp.107–109, 1999.

[6] H. Lieberman, “Letizia: An agent that assists web browsing,” Proc. 14th Int. Joint Conf. on Artificial Intelligence, pp.924–929, 1995.

[7] F. Menczer, “ARACHNID: Adaptive retrieval agents choosing heuristic neighborhoods for information discovery,” Machine Learning: Proc. 14th Int. Conf., pp.227–235, July 1997.

[8] F. Menczer, R.K. Belew, and W.Willuhn, “Artificial life applied to adaptive information agents,” Working Notes of the AAAI Symposium on Information Gathering from Distributed, Heterogeneous Databases, AI Press, 1995.

[9] M. Mori and S. Yamada, “Bookmark-agent: Information sharing of urls,” Poster Proc. 8th Int. World Wide Web Conf. (WWW-8), pp.70–71, 1999.

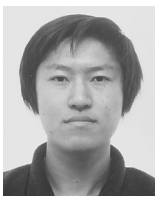
[10] B. Pinkerton, “Finding what people want: Experiences with the WebCrawler,” Electronic Proc. 2nd Int. World Wide Web Conf., 1994.

[11] G. Salton, Automatic Text Processing, Addison-Wesley, 1989.

[12] S. Yamada and N. Nagino, “Constructing a personal web map with anytime-control of web robots,” 4th Int. Conf. on Cooperative Information Systems, pp.140–147, 1999.

[13] 山田誠二，大澤幸生，“WWWにおける概念理解のためのナビゲーションプラニング，” 人工知能誌，vol.14, no.6, pp.1125–1133, 1999.

（平成11年8月2日受付，12月27日再受付）



梛野　憲克　（学生員）

1997名工大・工卒．1999東工大大学院総合理工学研究科修士課程了．現在同大学院博士課程在学中．情報検索，人工知能に興味をもつ．人工知能学会，AAAI各会員．

山田　誠二　（正員）

1984阪大・基礎工卒．1989同大大学院博士課程了．博士（工学）．同年大阪大学基礎工学部助手．1991同大学産業科学研究所講師．1996東京工業大学大学院総合理工学研究科助教授，現在に至る．工博．人工知能，特に，Webでの情報検索，ロボット学習に興味をもつ．人工知能学会，情報処理学会，日本ロボット学会，AAAI，IEEE各会員．