マルチ Web ロボットによるインターネット情報マップの作成

Building the Internet Information Map by multi Web robots

棚野 憲克　　山田 誠二
Norikatsu NAGINO　Seiji YAMADA
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
Tokyo Institute of Technology

Abstract

We propose a method to gather information efficiently use the Internet Information Map built by multi web robots. Web robots generally use simple breadth-first search and integrate their gatherd information. Therefore Web robots don't deal with dynamic environment and biased their gatherd information. Web robots will cope with the problem by our method.

1　はじめに

近年のインターネットの急速な普及に伴い，World Wide Web で公開されるホームページの数は急増し，その内容は頻繁に更新されている．現行では複数の Web ロボットが各々Web ページ中に含まれるリンクを辿ることにより情報を収集する．ほとんどの場合 Web ロボットは横型探索し，動的な環境や集められた情報の偏りに対応する機構は含まれていない．

そこで本研究では，動的な環境に対応し，集められる情報の偏りを少なくするための機構を提案する．

WEBSOM[2] は，Web ページに含まれるキーワードの頻度をもとに 2 段階の自己組織化マップ (SOM) を用いて Web ページを分類し，その結果をハイパーテキストインタフェースを用いて視覚的に表現することにより，ユーザが速く目的の Web ページにたどり着くようにした．

また，人工生命の技術を利用した InfoSpider[4] は，Web ページに含まれるキーワードの頻度とリンクの関係から決定されるエネルギーを各エージェントが得て複製，消滅をすることにより，関連するページにエージェントを集める手法を提案している．

2　WWW 上での情報収集

検索エンジンが用いる情報は，Web ページにアクセスしそのドキュメントの中に含まれるリンクを用いて他の Web ページにアクセスするという動作を繰り返すプログラムにより集められる．このようなプログラムは **Web ロボット** と呼ばれる．

Web ロボットの探索戦略は，ほとんどの場合横型探索を行い，最終的に各々の Web ロボットの収集した情報を統合する [1]．しかし，Web ページに含まれるリンクには関連性があることが多く，つねに全く異なるジャンルのリンクが並べられるということはない．そのため，Web ロボットが収集する Web ページには偏りができる．Web ロボットの収集する Web ページの分野に注目するとき，長い目でみればまんべんなくすべての情報が集められるかも知れないが，短期間に注目すれば特定の非常に偏りのある Web ページを集めていることになる．

また，Web ページは頻繁に更新，削除，追加される．つまり，Web ロボットに対する環境は頻繁に変化する．そのようなとき，ある分野の情報は現在 Web ロボットが探索しているため新しいが，他のある分野にはなかなかアクセスされないため古くなっている，ということになる．この問題に対応する機構もほとんど提案されていない．

3　インターネット情報マップの作成

3.1　インターネット情報マップ

本研究では，Web ロボットが現時点でどのような分野の Web ページをどのくらい集めたかを把握するために，収集した Web ページを分類したインターネット情報マップを作成することを考える．**インターネット情報マップ** は，Web ページ間の関係を表す．Web ページ間の関連度を調べるために，まず Web ページ中に現れる単語の重み付けを行う．本研究では，情報検索において重み付けに用いられる TF·IDF 法を用いる．Web ページ中の単語の内，重要度が高いものを幾つか選択し，その単語をキーワードとして Web ページを分類する．分類の方法には，自己組織化マップ (SOM)[3] を用いる．SOM の各出力ノードの値を成分とするベクトルを考え，出力ノードの数と同じ次元の仮想空間中にそのベクトルを位置ベクトルとして持つ点を配置したものをインターネット情報マップとする．

3.2　自己組織化マップに基づく Web ページの分類

SOM への入力は，重要語の頻度とする．重要語を決めるために TF·IDF 法 [5] により単語の重み付けをする．重要度 w が大きい単語を幾つか選択しその単語の Web

〒 226-8502 横浜市緑区長津田町 4259　nagino@ymd.dis.titech.ac.jp

ページ中における出現頻度を，SOM への入力とする．この時，選択する重要語数を n とし，分類する Web ページ (ここでは，Web ロボットが収集した Web ページ) 数を m とすると，SOM の入力層のノード数は n であり，m の Web ページの分類のための SOM への全入力数は $n \times m$ になる．ここで，選択する重要語数 n により SOM への入力数をおさえ，Web ページの分類にかかる処理の負荷を減らしている．

3.3 Web ロボットの探索戦略

一般に使われている Web ロボットは，横型探索を行うものがほとんどである．本研究ではこのインターネット情報マップ中の各点の粗密状況に基づき，なるべく粗な領域の Web ページを集めるように，マルチ Web ロボットを制御する．どのリンクを辿るかは以下の式により決定する．ここで，Web ページ d_i の重要語の出現頻度を SOM に入力したときの出力ノードの重みを成分としてもつベクトルを $\mathbf{D}_i$，出力ノードの個数を N とする．選択されるリンクは，Web ページ S に含まれるリンクから 1 つ選ばれる．

$$
\begin{aligned}
\mathbf{D}_i &= (x_{i1}, x_{i2}, \ldots, x_{iN}) \\
\mathbf{V} &= -\sum_{i=1}^{m} \mathbf{D}_i \\
s &= \min_{\mathbf{D}_i} |\overrightarrow{\mathbf{V}\mathbf{D}_i}|
\end{aligned}
$$

直観的には，$\mathbf{V}$ はインターネット情報マップにおいて点が密に配置されている場所からなるべく離れる方向と，同じようなスカラをもつベクトルである．s はマップ上に配置されている点の内，V を位置ベクトルに点から最も近い点であり，これはマップ上の粗である部分の点になると考えられる．

この方法では，ある Web ページに含まれるリンクは，その Web ページに関連する Web ページへのリンクであるという考えを用いている．

各 Web ロボットは Web ページを一つ選択し，その Web ページにアクセスした後，その Web ページを SOM に入力し，インターネット情報マップに対応する点を配置する．その後，上記の戦略により次に訪れるリンクを決定する．この動作を再帰的に適用することにより，階層的な構造ができる．Web ロボットはこの動作を，分散して非同期に行う (図 1)．

3.4 動的な環境，情報の偏りへの対応

Web ロボットがリンクを選択し，リンク先の Web ページを獲得したとき，その集めた Web ページ中のリンク先の Web ページのヘッダにアクセスする．この時，更新または削除されている場合は，インターネット情報マップの対応する点へのベクトルをベクトル V を求める際に，ベクトル合成に加えない．これにより，頻繁に更新または削除される Web ページに対応する部分が粗であるとみなされ，Web ロボットにより選択されやすくなる．また，得られた Web ページに偏りがある場合は，粗である部分がはっきり現れ，この場合も Web ロボットにより選択されやすくなる．

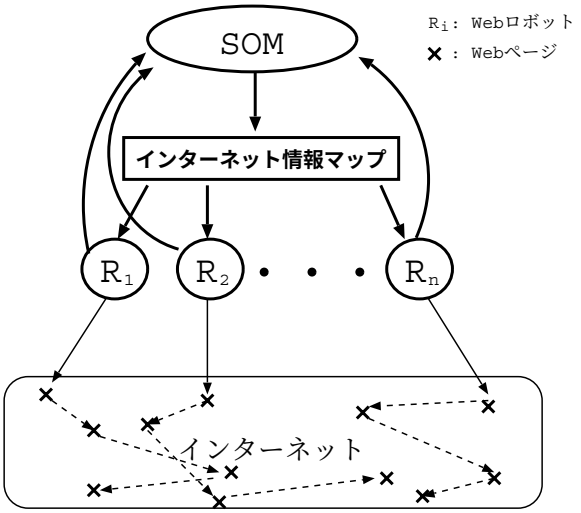

図 1: Web ロボットの流れ

4 まとめ

本研究では，Web ロボットが収集する Web ページを SOM を用いて分類したマップの粗密を考慮した探索手法を提案した．この手法を用いて，WWW の動的な環境や収集する情報の偏りに対応する方法を述べた．

参考文献

[1] Fah-Chun Cheong. *Internet Agents : Spiders, Wanderers, Brokers, and Bots*. New Riders, 1996.

[2] Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. Exploration of full-text databases with self-organizing maps. In *Proceedings of the ICNN96, International Conference on Neural Networks*, volume I, pages 56–61. IEEE Service Center, Piscataway, NJ, 1996.

[3] Teuvo Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, 1995. Second Extended Edition 1997.

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

[5] Gerard Salton. *Automatic Text Processing*. Addison-Wesley, 1989.