特　集　「Web システムにおける情報獲得支援技術」

# 知的 Web 情報システム

## Intelligent Web Information System

山田　誠二　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
Seiji Yamada　CISS, IGSSE, Tokyo Institute of Technology.
yamada@ymd.dis.titech.ac.jp, http://www.ymd.dis.titech.ac.jp/~yamada/

村田　剛志　国立情報学研究所情報学基礎研究系認知科学研究部門
Tsuyoshi Murata　Cognitive Science Research, Information Foundation Research Division, National Institute of Informatics.
tmurata@nii.ac.jp, http://research.nii.ac.jp/~tmurata/

北村　泰彦　大阪市立大学大学院工学研究科情報工学専攻
Yasuhiko Kitamura　Department of Information and Communication Engineering, Graduate School of Engineering, Osaka City University.
kitamura@info.eng.osaka-cu.ac.jp, http://www.kdel.info.eng.osaka-cu.ac.jp/~kitamura/

Keywords: the world wide web, information gathering / search / integration, hyper-link structure.

## 1. はじめに

人工知能の技術を Web の情報システムに応用する研究が活発に行われている．最近注目される AI の応用分野には，ペットロボット，ロボカップに代表される知能ロボットや，遺伝子解析やバイオインフォマティクスなどのゲノムの分野がある．それらと比較しても，Web は，基本的に記号で記述された膨大な情報の世界で，また一般の人々が日常的に触れる世界でもあり，さらには，ビジネスとも結び付きやすいまぎれもない実世界であることから，現在 AI 技術にとって最も敷居が低く，かつ実用性が高い重要な応用分野であることは間違いないだろう．当然，このような認識をもつ研究者も多く，さまざまな応用研究，あるいは，Web への適用という局面ではじめて必要になる新しい方法論の研究が行われている．

本稿では，このように近年活発になっている Web の情報システムにおける研究を，Web における情報収集，情報検索，情報統合，情報管理，そして Web における構造の発見という幅広い視点に立って，AI の直接的な応用からそれほど AI 色のないものまで，筆者らにとって興味深いものを紹介していく．

## 2. Web における知的情報収集

Web 上にある膨大な情報を有効利用することは非常に重要であるが，どうやって欲しい情報を集めるかという問題は簡単ではない．現在の Web の規模から考えると，大きな検索エンジンが一つあれば十分というのは非現実的であり，そのため，AI 技術が Web における情報収集に応用されている．ここでは，オンラインで情報収集するマルチエージェントシステムである ARACHNID と，組織的な知識を収集できるナビゲーションプラニングを紹介する．

### 2･1　ARACHNID

ARACHNID [Menczer 97] は，情報収集エージェントが Web 上で分散してオンラインで情報収集を行うシステムである．適合した Web ページを見つけたエージェントは，エネルギーを獲得して繁殖し，そうでないものはエネルギーを失い個体数も減少して，選択，進化が行われる．なお，類似システムの InfoSpider [Menczer 99] が，Web で公開されている[*1]．

#### §1　全体の仕組み

ARACHNID の基本手続きを図 1 に示す．まず，分散して複数のエージェントがばらまかれ，それぞれがリンクをたどり，近傍の Web ページを検索していく．そし

各エージェントをエネルギー$E=\frac{\varepsilon}{2}$で初期化.
(1)ランダムにエージェント$a$ を取り出す.
(2) $a$ のもつリンクを一つを選択する.
(3)そのリンク先のWeb ページ$D_a$を獲得する.
(4)エージェント$a$ のエネルギー$E_a$を下式で更新.

$$E_a \longleftarrow E_a - cost + \begin{cases} r(D_a) & D_a \text{未処理.} \\ f(D_a) & D_a \text{処理済.} \end{cases}$$

(5)$D_a$に処理済の印をつける.
(6)強化学習を行う.
(7)以下のようにして，繁殖，死滅を実行する.
　a　$E_a$が$\varepsilon$より大きいと，以下のように繁殖.
　　・$a$を複製，突然変異の後，子供 $a'$を作る.
　　・$a$と$a'$のエネルギーを $\frac{E_a}{2}$ で初期化.
　b　もし$E_a$ が負の値なら，$a$ は死滅する.

**図 1**　ARACHNID のアルゴリズム

＊1 http://dollar.biz.uiowa.edu/fil/IS/

て，得られたページ入力キーワードと適合すれば，自分のエネルギーを増大させ，そうでなければエネルギーが減少する．そして，エネルギーがしきい値以上であれば，子供を複製，突然変異させて，繁殖が行われる．また，エネルギーが負になれば，死滅する．このように，適合するWebページを多く含む近傍を多くのエージェントが探索するような制御が実現される．

図1のステップ7aで，繁殖を決定するパラメータが定数なので，各エージェントは，ほかのエージェントとは独立に繁殖を決定でき，RACHNIDが分散アルゴリズムになっている．また，図1のオンラインアルゴリズムに加えて，獲得されたWebページをユーザがオフラインで評価するメカニズムも用意されている．

なお，ARACHNIDの入力は，キーワードのリスト，出発点となるURLの集合，たどるページの上限であり，適合度でランキングされたURLリストが出力される．図2に，InfoSpiderの各エージェントがWebページを収集する軌跡を示す．

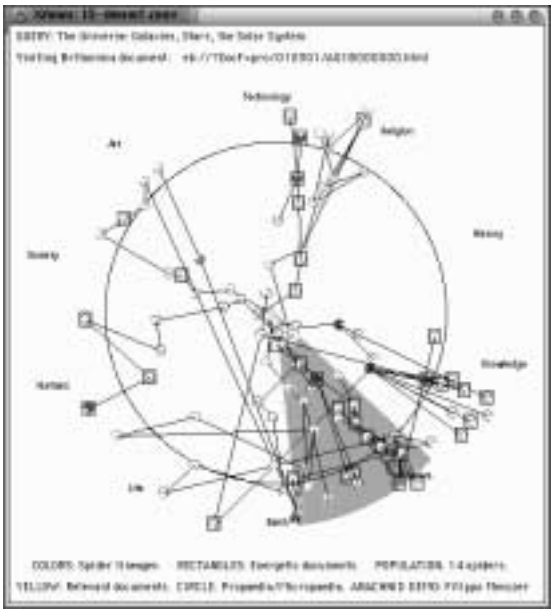

**図2** InfoSpiderの情報収集エージェントの軌跡

### §2 近傍ページの評価

図1のステップ2では，エージェントが現在のWebページからリンクを張られているページの適合度を評価し，次にたどるページが決定される．現在のWebページ $D_a$ のリンク $l$ の評価値 $\lambda_l$ は，次式により計算される．この式は，入力キーワードのうちのWebページ $D_a$ にも含まれるキーワードについて，$D_a$ においてそのキーワードとリンク $l$ との間に挟まれている他のリンク数を距離としているところが興味深い．つまり，直観的には，近くに他のリンクに遮られていない入力キーワードがたくさんあるリンクほど，評価が高くなる．そして，評価値 $\lambda_l$ が決まると，その値に比例した確率分布により，確率的な行為選択が行われる．

$$\lambda_l = \sum_{\text{キーワード} k \in D_a} \frac{match(k, Q)}{disk(k, l)}$$

$$match(k, Q) = \begin{cases} 1 & k \in Q \\ 0 & \text{それ以外.} \end{cases}$$

$disk(k, l) = D_a$ で $k$ と $l$ の間にあるリンクの数.

Webの部分集合と考えられるブリタニカ百科事典コーパスを用いて，最初に適合文書を見つけるのに必要だった文書数により評価実験がされた[Menczer 97]．その結果，単純な幅優先探索を使った探索システムよりも，100倍以上ARACHNIDのほうが優れている結果が得られた．

## 2・2 ナビゲーションプラニング

ナビゲーションプラニング[山田 99]は，ユーザが目標概念を理解するために有用なWebページをブラウズする手続きをプラニングと捉え，階層的な説明に役立つWebページの半順序系列を自動生成するシステムである．WebページからSTRIPS-likeオペレータを自動生成しながら，プラニングを行うという特長をもつ．

### §1 プラニングとしてのブラウジング

ある概念を理解しようとするユーザのブラウジングは，以下のようにまとめられる．この手続きは，ユーザが停止するまで繰り返される．

（1）サーチエンジンを使って，目標概念に関連のあるWebページを検索する．
（2）検索されたWebページのうち，役に立ちそうなページを見て理解する．
（3）そのWebページにおいて，未知の概念を目標概念として（1）にいく．

そして，この手続きは，以下のような対応で，プラニングとして定式化できる．

- 行為：Webページに記述されている概念を理解する．
- 状態：ユーザの知識状態．既知の概念を表す単語の集合により記述．
- 初期状態：ユーザの初期の知識状態．
- 目標状態：ユーザが理解したい目標概念．
- オペレータ：Webページを見て知識を獲得するU-オペレータは，以下の条件と効果からなる．
  - ・条件：そのWebページの理解に必要な条件知識．
  - ・効果：そのWebページから得られる効果知識．

ただし，あらかじめU-オペレータを世界中のWebページについて用意することは不可能なため，必要になったときに逐次的にU-オペレータを自動生成する方法を採る．

### §2 Webページからの U-オペレータの自動生成とプラニング

ナビゲーションプラニングでは，Webページから条

件知識と効果知識を自動抽出することにより，U-オペレータを自動生成する．そのため，タグ構造による抽出と *KeyGraph* による方法を併用する．

アンカータグの単語を条件知識の候補とし，タイトルタグ，ヘッディングタグの単語を効果知識の候補とする．そして，Web ページの意味構造をテキストだけの情報から推定し，その構造を基に文書での主張のこめられた単語をキーワードとして抽出する手法である *KeyGraph* [大澤 99] により，条件知識と効果知識の候補を求め，先の候補と総合的に評価する．

### §3 プラニングと生成されたプラン

ナビゲーションプラニングのプラニング手続きは，目標状態から後向きのビーム探索である．なお，状態ノードの展開（図 3）では，サーチエンジンを使った関連 Web ページの検索と，U-オペレータ生成が行われる．

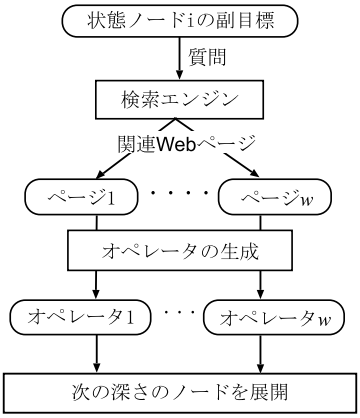


**図** 3 ノードの展開

図 4 に，目標概念が“ concept formation ”の場合に生成されたプランを示す．ページ 1 は，目標概念を直接的に ILP を基にして説明している．そして，ILP については，ページ 3 で詳しく述べられ，4 には関連する機械学習の論文アブストラクトが紹介されている．


**図** 4 “ Concept formation ”を説明するために生成されたプラン

ナビゲーションプラニングシステムは，Perl で実装されている．そして，サーチエンジンで収集した Web ページによる表層的な概念理解よりも，より深い理解が得られることが被験者を使った実験により確認されている．

## 3. Web におけるハイパーリンク構造の発見

Web からの知識発見（Web Mining）は，Web content mining，Web structure mining，Web usage mining の三つに大まかに分類できる [Kosala 00]．これらのアプローチの中で，ハイパーリンクのグラフ構造に基づいてページの重要度を見いだしたり，ページ間の関連性を見いだす Web structure mining の研究が近年盛んになってきている．ハイパーリンクはときとしてコンテンツが表現する以上の情報を提供するものであり，また大規模な Web データを高速に処理する上でもハイパーリンクに基づくアプローチは有効である．ハイパーリンクの意味としては，同一サイト内のページを見る閲覧者の利便性のためや，広告のためなどさまざまなものがあるが，全体としてはリンク先のページ内容に対する支持を表すものが一番多いと考えられる [Clever 99]．ハイパーリンクのグラフ構造の視覚化が盛んに研究されているのも，Web ページの重要度やページ間の関連性を見いだすうえで，ハイパーリンクが重要な手がかりであると考えられているためである．

ここでは，ハイパーリンクの構造に基づいた研究として，query dependent ranking の手法である HITS [Kleinberg 98] と，query independent ranking の手法である PageR-ank[Page 98]について説明する．また，リンクのグラフ構造から関連ページ集合であるコミュニティを発見する Web Trawling [Kumar 99] や，リンクの構造の意味を見いだす試みである ParaSite [Spertus 97] についても紹介する．

### 3･1 HITS

HITS [Kleinberg 98] は，特定トピックに関する有用なページを出力するアルゴリズムである．このアルゴリズムにおいては Web ページの有用性の評価基準として，特定トピックに関する情報の豊富さを表すオーソリティと，オーソリティへのリンクの豊富さを表すハブを導入している．オーソリティとして価値の高いページへリンクを張っているページはハブとしての価値が高く，またハブとして価値の高いページからリンクを張られているページはオーソリティとしての価値が高いといえる．オーソリティとハブの具体的な計算方法は以下のとおりである．

(1) 特定トピックに関する有用なページを含んでいる Web ページ集合を取得する．まず通常のサーチエンジンでのキーワード検索によって 200 ページ程度を取得し（これを root set とする），その root set のページからのリンク先のページや，root set のページへリンクを張っているページを追加する（これを base set とする）．このようにして得られるおよそ 1 000 ～ 3 000 ページの

base set の中に，特定トピックに関するオーソリティやハブが含まれていると考えられる．

（2）次に，この base set に含まれている各ページ $p$ のオーソリティ（$x_p$）とハブ（$y_p$）を計算する．相互再帰的な関係から，両者は以下のように定義できる．

$$x_p = \sum_{q \ such \ that \ p \to q} y_q$$

$$y_p = \sum_{q \ such \ that \ p \to q} x_q$$

（3）個々のページの $x_p$ と $y_p$ を計算する方法として，base set に含まれている $n$ 個のページの隣接行列を用いる．隣接行列 $A$ はページ $i$ からページ $j$ へのハイパーリンクが存在する場合に $(i, j)$ の成分が 1 で，それ以外は 0 であるような $n \times n$ の正方行列である．base set の $n$ 個のページのオーソリティとハブをベクトル $\boldsymbol{x}, \boldsymbol{y}$ で表現するとそれぞれ $\boldsymbol{x} = A^T\boldsymbol{y}$，$\boldsymbol{y} = A\boldsymbol{x}$ となる．$\boldsymbol{x}$ と $\boldsymbol{y}$ を正規化することにすると，両者はそれぞれ $A^TA$，$AA^T$ の主固有ベクトルとして求めることができる．

HITS は，最初の root set の取得以外はハイパーリンク情報のみに基づくシンプルなアルゴリズムであり，その振舞いの特徴についての研究も数多くなされている．HITS において入力として特殊なトピックが与えられた場合，それを一般化した内容のオーソリティとハブが得られる傾向がある [Gibson 98]．また入力トピックと関係のないページが base set の大半を占める場合には，最終的に得られるハブやオーソリティが入力トピックと関係のないページとなってしまう現象が指摘されている [Chakrabarti 99, Henzinger 01]．この問題を解決するための工夫として，リンク周辺のテキスト情報を用いたり，リンクに対する重みを導入するなどして HITS を改良した Clever [Chakrabarti 98, Clever 99] の研究が進められている．

### 3・2 PageRank

PageRank [Page 98] は「多くの良質なページからリンクされているページはやはり良質なページである」という考えに基づくランキング手法である．ページ $A$ の PageRank $R(A)$ は，以下のように定義される．

$$R(A) = \frac{\varepsilon}{n} + (1-\varepsilon) \sum_{B \ such \ that \ B \to A} \frac{R(B)}{outdegree(B)}$$

ここで，

- $n$ は対象とする Web ページの総数．
- $\varepsilon$ は定数（0.1 から 0.2 の範囲の値）．
- $outdegree(B)$ はページ B から出るリンクの総数．

とする．あるページの PageRank の値は，そこにリンクを張っているページの PageRank によって決まるため，任意の初期値から開始してこの計算を反復して行うことで，各ページの PageRank の値を求める．この値は，ランダムにリンクをたどる閲覧者がページを訪れる確率に対応しているが，$\varepsilon/n$ は任意のページにジャンプする確率を表しており，この項を導入することによって，外に対してリンクを持たないページの PageRank の値が不当に大きくなってしまうことを防いでいる．

PageRank は，サーチエンジン Google（http://www.google.com/）における検索結果のランキング手法の一部として利用されている．文章中の単語の出現頻度に基づいた従来の情報検索でのランキング手法では，Web ページ上の見えない部分に単語の羅列を追加するなどして，検索結果の上位にページが配置されるよう Web ページ作成者によって恣意的な操作がなされる危険性がある．PageRank においては，あるページにハイパーリンクを張っているほかのページの重要度によってそのページの重要度を決定するためにそのような操作が行われにくく，サーチエンジンにおける妥当なランキングを実現する上で好ましいものである．

### 3・3 Web Trawling

関連性のあるページ集合であるコミュニティをハイパーリンクのグラフ構造を利用して見いだす研究として，Kumar らの Web Trawling [Kumar 99] がある．Web Trawling は，HITS のように入力された特定トピックについてのページ集合を見いだすのではなく，Web 全体に含まれているコミュニティをすべて数え上げることを目標としている．

数え上げる際の単位として，$i$ 個の頂点のおのおのから $j$ 個の頂点すべてに対する有向辺が存在するような完全 2 部グラフ（complete bipartite graph）$K_{i,j}$ と，少なくとも一つの $K_{i,j}$ を含んでいるような $i+j$ 個の頂点からなる 2 部コアグラフ（bipartite core）$C_{i,j}$ を定義する．Kumar らは，Web において十分に表現されているトピックにはそれに対応する（適切な $i$ と $j$ の）$C_{i,j}$ が存在するという仮説に基づき，$C_{i,j}$ を Web のグラフ構造から高速に探索するためのアルゴリズムを提案している．実際に Web のスナップショットから bipartite core を探索した結果として，$i = j = 3$ 程度の bipartite core が Web 中に数十万個が存在することや，bipartite core を構成するページ集合は，トピックが明確なコミュニティに対応している場合がほとんどであることを示している．

これに関連する研究として，Web のスナップショットを用いずにサーチエンジンで backlink 検索を行って完全 2 部グラフを探索することで，入力 URL を含んでいるような Web コミュニティを発見するアプローチ [村田 01]もある．

### 3・4 ParaSite

ハイパーリンクによるグラフ構造において，個々のハイパーリンクに重みづけをすることによってその重要度を表現するアプローチは上述の Clever などでも採用さ

れているが，参照先によってハイパーリンクを分類し，その結合関係からリンクの構造の意味を見いだす試みとして ParaSite [Spertus 97] がある．ParaSite においては，ハイパーリンクによる参照先が同一サイト内のファイル階層の上位のもの，下位のもの，同階層のもの，そしてほかのサイトへのものの 4 種類に分類し，その組合せによって参照元と参照先のページの関係を推定するヒューリスティックを導入している．このヒューリスティックだけでページ間の関係を完全に決定することは困難であるが，ハイパーリンクによる構造の意味を見いだす第一歩として注目される．

### 3･5 ハイパーリンクによるグラフ構造の解明に向けて

ハイパーリンクによって構成されるグラフ構造を解明することの利点として，Broder らは，以下の五つをあげている [Broder 00].

（1）ロボットが Web ページを収集する際の戦略決定.
（2）Web コンテンツ生成における社会学の理解.
（3）リンク情報を用いる Web アルゴリズム（Page-Rank など）の振舞いの分析.
（4）Web 構造（bipartite core など）の発展の予測とそれを発見するアルゴリズムの開発.
（5）Web グラフにおける重要な新現象の出現の予測.

この分野の研究はまだ始まったばかりであり，これらの壮大な目標を達成するために今後一層の研究がなされることが期待される．

## 4. Web における知的情報検索

Web における情報検索は，従来の情報検索研究が対象としていた新聞や特許などの文献データとは異なり，以下のような特徴がある．

- 非常に大規模である．
- コンテンツの形式や質が多様である．
- ハイパーリンクによる参照関係がある．
- 動的に変化している．

TREC などの情報検索の評価会議においても Web を対象としたタスクが取り上げられているが，用いるデータコレクションの内容や規模，検索課題，評価方法など，多くの課題が残されている [福島 00].

本章では，サーチエンジンの選択的利用と，ハイパーリンクなどによる参照関係を利用した検索システムを取り上げる．

### 4･1 選択的メタサーチエンジン

一般に一つのサーチエンジンのカバーできる Web ページは，全体の一部であり，また，複数のサーチエンジン間で重複する Web ページも一部にすぎない．よって，既存の検索エンジンをうまく組み合わせて利用できれば，より広い範囲の Web ページをカバーできる．このような考えのもとに，自分自身では Web ページのデータベースを持たずに，既存の複数のサーチエンジンを利用するメタサーチエンジンが増えてきており，MetaCrawler [Selberg 97]，Inquirus [Lawrence 98] などがある．

メタサーチエンジンは，自分自身で Web ページを収集せず，ユーザから与えられた検索質問をほかのいくつかのサーチエンジンに転送して，結果をまとめて表示する．しかし，登録されているサーチエンジンの特徴を考慮しないため，検索内容を得意としない検索エンジンが多くの非適合 Web ページを返すことで，メタサーチエンジンの検索結果の適合率が低くなる場合がある．この問題に対し，検索質問に応じて，サーチエンジンを選択的に利用する選択的メタサーチエンジンが研究されている．

ProFusion [Fan 99] は，あらかじめネットニュースのグループから単語を取り出し，各単語のカテゴリを決めて，それに対応したサーチエンジンを選択的に利用する．よって，登録されていない単語では，適切なサーチエンジンを選択できない．

SavvySearch [Howe 97] も，サーチエンジンを自動的に選択するメタサーチエンジンである．ユーザの検索質問や，以前利用したユーザが検索結果リストのページを閲覧したかどうかなどのユーザからのフィードバックを参考にして，各サーチエンジンを評価し，選択的にサーチエンジンを利用する．

MetaWeaver [Mori 00] は，実際に検索を行うサーチエンジンごとに，その専門分野を捉え，検索質問に応じて適切なサーチエンジンを自動的に選択して検索依頼する．さらに，類義語辞書を用いて，未検索の語に対するサーチエンジンの評価が可能という特徴をもつ．

### 4･2 Google

Google において採用されている PageRank アルゴリズムについては先に述べたとおりである．サーチエンジンとしての Google の特徴が PageRank による適切なランキング結果にあることはもちろんであるが，そのほかの特徴として以下のものが自身のサイトに記述されている [Google].

（1）入力した語句を含むページだけの表示.
（2）ページ内でのキーワードどうしの位置の考慮.
（3）各検索ごとに関連性の高いプレビューの表示.
（4）“I ’m Feeling Lucky ”によるページの自動表示.
（5）Web ページのキャッシュ.

（1）は，一見当然のことのように思えるが，多くのサーチエンジンが検索結果として得られる件数の多さを競っているのとは対照的であり，検索結果の質を重んじる立場が明確になっている．（2）は，入力キーワードがページ内で近接しているものを上位にランキングすることを示しており，（3）は入力されたキーワードと一致する

テキストの抜粋を表示することで，検索結果が有用かどうかを判断しやすくしている．(4) は，検索結果の最上位のページを自動的に表示するものであり，検索結果に対する Google の自信を表しているものといえる．

(5) は，サーチエンジンとしての機能とは直接関係のないものであるが，Google のページ収集能力の高さがもたらす特徴である．検索結果として得られたサイトにアクセスできない場合でもその内容を見つけることができるよう，Google ではキャッシュへのリンクが検索結果の URL と共に表示される．Google はおよそ 13 億の Web ページをアーカイブに保存しており，ディスクスペースは 4 〜 5 テラバイトを占めるという．この大規模なキャッシュは，いわばネットワーク・バックアップサービスとみなすことができ，Web サーバがクラッシュした際に，Google のキャッシュでファイルを回復することができたという報告もある [Kahney 01].

### 4・3 ResearchIndex

学術論文などの文献を対象とした検索システムとして ResearchIndex（CiteSeer）[Giles 98] を紹介する．これは PostScript などのファイル形式で Web 上に公開されている論文をロボットによって収集し，論文中の記述やその引用関係から関連論文を見いだしたり，引用の文脈を明らかにするものである．このシステムは以下の手順で処理を行う．

(1) ネットニュースやメーリングリストでのアナウンスや，オンラインでアクセス可能なジャーナルの最新号などから，ロボットが論文ファイルを収集する．
(2) 収集された論文ファイルを解析し，ヒューリスティックに基づいて URL，タイトル，要約，引用，引用の文脈，本文などを抽出する．
(3) 引用されている論文の表記を正規化し，論文間の引用関係を明確にする．
(4) 単語ベクトル，文字列間の距離，引用関係における TFIDF に相当する CCIDF（Common Citation × Inverse Document Frequency）などの組合せによって論文間の関連性を見いだす．

このシステムは http://www.researchindex.com/において公開されており，以下のような特徴があげられる．

- Web 上に公開されたものであれば自動的に処理することができるため，新しい論文を扱うことができる．
- インデクシングに人手を必要とせず，完全に自動で処理を行う．
- 引用の文脈を明らかにすることで，論文の重要度を正確に見積もることができる．

## 5. Web における知的情報統合と管理

Web 上には膨大な情報や知識が蓄積されており，地球規模で巨大な知識ベースを形成しているといっても過言ではない．このような知識ベースは中央集権的に設計されたものではなく，いわば自律分散的に維持管理されているといえる．したがって関連する情報源が分散して存在することが多々あり，これらを統合して利用できるようにすればその付加価値を格段に高めることができる．例えば，Web 上には監督や出演者の情報を提供する映画サイト[*2]，劇場での上映作品の情報を提供する劇場サイト[*3]，映画の批評情報を提供する批評サイト[*4] が存在する．これらのサイトから得られる情報を統合することができれば，単独のサイトでは検索不可能な，「三ツ星以上の評価がされたスピルバーグ製作の映画で，現在大阪梅田で上映中のもの」といった検索要求に応えることが可能になる．

現在のところ，このような情報統合は人間の利用者がブラウザを用いて手作業で行うことは可能ではあるが，それを自動化しようとする試みが知的情報統合（intelligent information integration）である．

### 5・1 情報の抽出

Web 情報統合の第一ステップは，統合に必要な情報が書かれている Web ページを探し出し，その情報を抽出することである．しかしながら現在の Web ページの多くは HTML により記述されており，その内容を人間がブラウザを介して理解することは容易であっても，コンピュータで機械的に処理することは容易ではない．すなわち，HTML で提供されるタグは視覚的な構造を表現することができるが，意味的な構造を表現することができないからである．したがって映画に関する Web ページから監督や俳優の名前を機械的に抽出することも容易ではない．

しかし，Web を介して提供されるデータベースではページやデータの配置には規則性がある場合が多いので，リンクの構造やタグの位置関係を利用することにより情報の抽出が可能になる．このような情報抽出モジュールはラッパ（wrapper）と呼ばれている．ラッパを開発するための技術としては，情報抽出のためのテンプレート [Hsu 97] やプログラミング言語 [Florescu 98, Kistler 98, 北村 98] によるものがある．また GUI ベースでラッパの自動合成を行う試み [Knoblock 98] もある．

ただし，最近では XML の普及に合わせて Semantic

---

*2 http://www2u.biglobe.ne.jp/ste-man/
*3 http://eee.eplus.co.jp/cinema/
*4 http://www.lares.dti.ne.jp/charlie/

Web [Decker 00] と呼ばれる機械可読な Web ページに関する標準化の試みも進められており，今後は情報抽出の困難さは軽減されることになるであろう．

### 5･2 情 報 の 融 合

Web 情報統合の第二ステップはラッパにより抽出された情報を融合[*5] することである．異種情報源の統合はこれまでもデータベースの分野で盛んに研究が行われてきた [Ozsu 91]．しかし自律分散的に発展する Web 情報源ではあらかじめ固定的なスキーマを前提とすることができないという問題がある．そこで，このような動的な情報源の融合を扱うアプローチとしてはメディエータ（mediator）[Wiederhold 92] によるものが一般的である．すなわちそれぞれの情報源をラッパを用いて共通化し，メディエータを介して利用者やアプリケーションに融合された情報を提供する．メディエータは専門領域に応じて複数存在することも可能であり，その組合せに応じてさまざまな情報の融合が可能になる．

メディエータを用いた情報統合システムとしては TSIM-MIS [Chawathe 94] が代表的であり，OEM（Object Exchange Model）という共通データ表現を用いて，情報源におけるデータ構造の違いを吸収している．さらに Ariadne[Knoblock 98]では統一的なオントロジを用いて情報源の意味的な違いにも対処している．

### 5･3 情報アクセスプラニング

Web 情報統合ではさらに以下のような特徴をもつ情報源も考慮する必要がある．

- 通信路やサーバの混雑より情報源へのアクセス時間は一様ではない．
- 有料情報源へのアクセスには費用的なコストがかかる．
- 類似の情報を扱う情報源が多数ある．またその中には誤った情報も含まれる．
- 頻繁に更新される情報源があり，最新の情報を入手するためには，再検索が必要になる．
- 情報源の検索能力に違いがある．例えば，映画データベースの場合，監督名や俳優名からの検索が不可能で，映画タイトルからの検索のみが可能な情報源がある．

すなわち統合可能な多数の情報源の中から，利用者の検索要求，時間や経費などのコスト的な制約，要求される解の質などを満たす情報源を選択して情報統合を行う必要がある．これに対する試みとしては，BIG [Lesser 00] ではプランナとスケジューラを利用して，資源制約下における情報統合の問題を扱っている．また，Ariadne [Knoblock 98] には情報源の検索能力を考慮した情報統合プランナが組み込まれている．さらに頻繁に更新される動的な情報源に対して，解の質と情報の新鮮度を考慮して，適切なアクセスプラニングを行う試み [Kitamura 00] もある．

### 5･4 適応 Web サイト

Web サイトは，ユーザが欲しい Web ページにたどりつくのが難しい構造になっていたとしても，そのリンク構造を変更することは容易ではない．適応 Web サイト（adaptive Web site）[Perkowitz 00] は，ユーザのアクセスパターンから学習することで，リンク構造や表現を自動的に変更できる Web サイトである．

#### § 1 探索としての適応

適応 Web サイトでは，Web サイトの自動的改良を，Web サイトの構造の探索とみなす．その探索空間で，状態は，Web サイト自身であり，探している Web ページを見つけられる頻度である再現率と，見つけるのに要するユーザの負荷を評価関数として探索を行う．

また，初期状態は，現在の Web サイト，目標状態は，質が最大限に向上した Web サイトである．よって，状態遷移は，Web サイトの修正を意味する．具体的には，リンクやページの追加・削除などがある．

以上が，適応 Web サイトの概念的な枠組みであるが，以下に具体的な実装例を紹介する．

#### § 2 インデックスページの自動生成

Web サイトの自動的改良の一つとして，あるトピックの関連する相互にリンクされていない Web ページのリンク集であるインデックスページの自動生成が研究された．

まず，PageGather というクラスタ発見の学習アルゴリズムでインデックスページ生成を実現している．クラスタ発見とは，従来のクラスタリングのようにすべての対象を一つのクラスタに分類するのではなく，オーバーラップとどのクラスタにも属さないことを許した少数のクラスタを見つけることを意味する．PageGather は，まずページの同じアクセスでどの二つのページが閲覧されているかを調べ，共起行列をつくり，その上で連結性にもとづくグラフ抽出を行う．そして，発見されたクラスタをランク付けして，インデックスページのひな型として Web マスターに提示する．

さらに，人間の直感的な概念に対応するようなクラスタを見つけるために，概念記述言語を入力とし，それに合うようなクラスタをみつける概念クラスタ発見を定義し，それを説くことにより，人間の理解に合うインデックスページの生成を実現している．

## 6. ま と め

本稿では，人工知能の応用という側面から，Web における情報システム，特に情報収集，検索，統合，ハイ

---

*5 http://www2u.biglobe.ne.jp/ste-man/

パーテキストの構造抽出について紹介した．とはいっても，この分野において，人工知能の素直な応用で実用的価値のあるテーマもまだ十分に開拓されているとは言えないと考えられ，今後やるべきことは，まだ数多く残されているだろう．また，社会的インパクトのある本当の応用研究も，特に国内においてまだまだ不十分である．多くの研究者，開発者の参入を願いたい．

## ◇ 参 考 文 献 ◇


[Broder 00] Broder, A. et al.: Graph Structure in the Web, *Proc. of the 9th WWW Conf.*（2000）

[Chakrabarti 99] Chakrabarti, S. et al.: Mining the Web 's Link Structure, *IEEE Computer*, Vol. 32, No. 8, pp. 60-67（1999）

[Chakrabarti 98] Chakrabarti, S. et al.: Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text, *Proc. of the 7th WWW Conf.*（1998）

[Chawathe 94] Chawathe, S. et al.: The TSIMMIS Project: Integration of Heterogeneous Information Sources, *Proc. of IPSJ Conf.* pp. 7-18（1994）

[Clever 99] Clever Project: Hypersearching the Web, Scientific American, `http://www.sciam.com/1999/0699issue/0699raghavan.html`（1999）

[Decker 00] Decker, S. et al.: The Semantic Web: The Roles of XML and RDF, *IEEE Internet Computing*, Vol. 4, No. 5, pp. 63-74（2000）

[Fan 99] Fan, Y. and Gauch, S.: Adaptive Agents for Information Gathering from Multiple, Distributed Information Source, in *Proc. of 1999 AAAI Symposium on Intelligent Agents in Cyberspace*, pp. 40-46（1999）

[福島 00] 福島: WWW 情報検索技術と評価の問題, 情報処理, Vol. 41, No. 8, pp. 913-916（2000）

[Florescu 98] Florescu, D., Levy, A., and Mendelzon, A.: Database Techniques for the World-Wide Web: A Survey, *SIGMOD Record*, Vol. 27, No. 3, pp. 59-74（1998）

[Gibson 98] Gibson, D., Kleinberg, J. and Raghavan, P.: Inferring Web Communities from Link Topology, *Proc. of the 9th Conf. on Hypertext and Hypermedia*（1998）

[Giles 98] Giles, C., Bollacker, K. and Lawrence, S.: Cite-Seer: An Automatic Citation Indexing System, *Proc. of the 3rd ACM Conf. on Digital Libraries*, pp. 89-98（1998）

[Google] Google, Google の人気の秘密, `http://www.google.com/intl/ja/why_use.html`

[Henzinger 01] Henzinger, M.: Hyperlink Analysis for the Web, *IEEE Internet Computing*, Vol. 5, No. 1, pp. 45-50（2001）

[Howe 97] Howe, A. E. and Dreilinger, D.: Savvy Search: a metasearch engine that learns which search engines to query, *AI Magazine*, Vol. 18, No. 2, pp. 19-25（1997）

[Hsu 97] Hsu, J. Y. and Yih, W.: Template-Based Information Mining from HTML Documents, *AAAI-97*, pp.256-262（1997）

[Kahney 01] Kahney, L.: Cache at the End of His Rainbow, `http://www.wired.com/news/business/0,1367,41065,00.html`（2001）

[Kistler 98] Kistler, T. and Marais, H.: WebL - A Programming Language for the Web, *Proc. of 7th WWW Conference*（1998）

[北村 98] 北村, 野崎, 辰巳: スクリプトに基づく WWW 情報統合支援システムとゲノムデータベースへの応用, 電子情報通信学会論文誌, Vol. J81-D-I, No. 5, pp. 451-459（1998）

[Kitamura 00] Kitamura, Y., Noda, T., and Tatsumi, T.: A Dynamic Access Planning Method for Information Mediator, Cooperative Information Agents IV, Lecture Notes in Artificial Intelligence, Vol. 1860, pp. 39-50（2000）

[Kleinberg 98] Kleinberg, J. et al.: The Web as a Graph: Measurements, Models, and Methods, *Proc. of CO-COON '99*, pp. 1-17（1999）

[Knoblock 98] Knoblock, C. A., et al.: Modeling Web Sources for Information Integration, *AAAI-98*, pp.211-218（1998）

[Kosala 00] Kosala, R. and Blockeel, H.: Web Mining Research: A Survey, *ACM SIGKDD Explorations*, Vol. 2, No. 1, pp. 1-15（2000）

[Kumar 99] Kumar, R. et al.: Trawling the Web for Emerging Cyber-Communities, *Proc. of the 8th WWW Conf.*（1999）

[Lawrence 98] Lawrence, S. and Giles, C.: Inquirus, the NECI meta search engine, in *Proc. of the 7th International World Wide Web Conference*（1998）

[Lesser 00] Lesser, V. et al.: BIG: An Agent for Resource-Bounded Information Gathering and Decision Making, *Artificial Intelligence*, Vol. 118, No. 1-2, pp. 197-244（2000）

[Menczer 97] Menczer, F.: ARACHNID: Adaptive Retrieval Agents Choosing Heuristic Neighborhoods for Information Discovery, in *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 227-235（1997）

[Menczer 99] Menczer, F. and Monge, A. E.: Scalable Web Search by Adaptive Online Agents: An InforSpiders Case Study, in *Intelligent Information Agents*, pp. 323-347, Springer（1999）

[Mori 00] Mori, M. and Yamada, S.: Adjusting to Specialties of Search Engines Using MetaWeaver, in *WebNet 2000*, pp. 408-412（2000）

[村田 01] 村田: 参照の共起性に基づく Web コミュニティの発見, 人工知能学会論文誌, Vol. 16, No. 3, pp. 316-323（2001）

[Ozsu 91] Ozsu, M. T. and Valduriez, P.: *Principles of Distributed Database Systems*, Prentice Hall（1991）

[大澤 99] 大澤, Benson, N. E., 谷内田: KeyGraph: 単語共起グラフの分割・統合によるキーワード抽出, 電子通信学会誌論文誌, Vol. J82-D-I, No. 2, pp. 391-400（1999）

[Page 98] Page, L. Brin, S. Motwani R. and Winograd T.: The PageRank Citation Ranking: Bringing Order to the Web, Online manuscript, `http://www-db.stanford.edu/~backrub/pageranksub.ps`（1998）

[Perkowitz 00] Perkowitz, M. and Etzioni, O.: Towards adaptive Web site: Conceptual framework and case study, *Artificial Intelligence*, Vol. 118, pp. 245-275（2000）

[Selberg 97] Selberg, E. and Etzioni, O.: The MetaCrawler architecture for resource aggregation on the Web, in *IEEE Expert*, Vol. January-February, pp. 11-14, IEEE（1997）

[Spertus 97] Spertus, E.: ParaSite: Mining Structural Information on the Web, *Proc. of the 6th WWW Conf.*（1997）

[山田 99] 山田, 大澤: 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, 人工知能学会誌, Vol. 14, No. 6, pp. 1125-1133（1999）

[Wiederhold 92] Wiederhold, G.: Mediators in the Architecture of Future Information Systems, *IEEE Computer*, Vol. 25, No. 3, pp. 38-49（1992）


2001年5月9日受理

## 著 者 紹 介


**山田 誠二**（正会員）は，前掲（Vol. 16, No. 1, p. 157）参照.

**村田 剛志**（正会員）は，前掲（Vol. 16, No. 3, p. 444）参照.

**北村 泰彦**（正会員）

1983年大阪大学基礎工学部情報工学科卒業．1988年同大学院博士課程修了．工学博士．同年，大阪市立大学工学部電気工学科助手．現在，同大学院工学研究科情報工学専攻助教授．分散人工知能，ヒューリスティック探索，WWW 情報統合の研究に従事．IEEE，AAAI，ACM，電子情報通信学会，情報処理学会，日本ソフトウェア科学会などの会員.