Web上のハイパーリンクと文書の意味構造
—検索は情報爆発に有効なのか—

Semantic principles of Hyper-links
Does the Search engine really work for tomorrow?

春木　良且
Yoshikatsu HARUKI

1．問題提起

Webシステムは、その衝撃的な登場から早くも10年以上経過し、その間PCのキラーアプリケーションという位置づけを経て、現在では社会、経済活動においても欠くべからざる存在となっている。その有効性の源泉は、技術そのものと言うよりは、データをハイパーメディア形式で組織化したという点にあると言っても過言ではなかろう。

つまり、データの内容そのものではなく、データの組織化という、いわばメタデータ的な側面を持つため、時間の経過に従って、データがハイパーテキスト形式で増大してきており、その情報量は、質の問題を除けば、図書館などは遥かに凌駕するものであろうことは想像に難くない。

筆者は、こうしたWebの自然発生的な側面に着目し、そのデータ構造、知識構造について、様々な側面から分析、検証してきた[1]。既に存在しているモノを考察対象とするのは、工学的な観点から知識ベースを設計し検証するよりも、遥かに直接的に人間の知識の共有形態などを探ることができると思われる。もちろんその知見をもとに、工学に反映することを想定するのは言うまでもない。

1．1　情報爆発とその課題

昨今、公的機関や研究者などで頻繁に使われているキーワードに「情報爆発」がある。

カリフォルニア州立大学バークレー校の「How Much Information?」プロジェクトで、2000年前後より、人類が創出する情報量が爆発的に増加しているという統計データが明らかにされており、[2]、それによれば、全世界で新たに作成された情報は、1999年で2EB、2002年で3－5EB程度にまで達しているということである。この数字は、「2000年におけるディスクのWW総出荷容量3EBと比べてそれほど、大きくはずれていない」[3]とされている。また同プロジェクトでは、人類が誕生以来今日まで残してきた全ての情報の量を、2003年～2004年のわずか2年間に生み出された情報量が超えてしまったことが指摘されている。こうした情報量の急激な増加を、「information Explosion（情報爆発）」あるいは「information TSUNAMI（情報津波）」と呼んでいる。

いわゆる「Web2.0」と呼ばれている次世代型のネット技術が、そのトレンドと大きく関わっているのはあえて指摘するまでもないであろう。

この「情報爆発」現象は、単純に言えば、全世界の65億の人間が容易に情報の受発信ができるようになったことに起因するわけであり、豊富な情報が入手可能になっているという側面から、このこと自体は望ましい現象と言えるかもしれない。しかしながら、メディアや技術がドラスティックに変化をしたとしても、人間の情報摂取能力は大きくは変わってはいないだろう。つまり、この膨大な情報空間から情報を取り出すことに纏わる課題があるということは、自明である。

そもそもInternet上には、質の低い情報しか存在しないと、

旧来から言われていた。ネット上の掲示板を「便所の落書き」と評したジャーナリストもいたほどである。しかし［4］などでも指摘されているが、その指摘は現在では成り立たない。現在のネットは、公的機関や学術機関などから充実したデータベースが公開されているし、また書籍のオンライン販売サイトなどは、図書館以上に充実した文献リストともなっている。しかしながら、今の情報技術は、シャノンが体系化した、情報の意味内容を捨象して表現のみに着目した情報理論に基づいて情報を捉えざるを得ないため、爆発する情報のうち、どれが重要で必要なのか、誰も判断できなくなってきている。「個々人は自分の理解できる情報（つまりレベルの低い情報）だけに手を出し、もはやレベルの高い情報の存在には気づかない。」[4]という指摘は、極端ではあるものの、まさにその通りであろう。

正確な出典は不明であるが、一般に「知識労働者の活動時間の30%が探すという行為に費やされている」という指摘がある。(注:この指摘も、主にブログで言われている情報であるということが、皮肉ではあるが)

Web2.0の社会的な影響として、「Web接触時間が勝敗を決める」と言った意見が巷間多く言われているが、であるとするならば、「Web2.0社会で最強のプレイヤーになれるのはニート」という[5]指摘は、あながち極論でもなかろう。

よって、現在のWebアプリケーションでは、情報を「探す」ことに特化した、「検索サービス」にフォーカスが当たっている。例えば、「わが国が、著しく知的集約度の高い産業を維持し、高付加価値製品・サービスを創出し続けるといったコンピタンスを得るには、情報空間からのサーチの効率化が国力を左右する」[6]といった意見に見るように、現在では、検索サービス、特にそのコア部分であるサーチエンジンの開発は、国家的な問題という指

摘すら存在する。

その背後には、言うまでもなく、今や幅広いサービスを提供する巨大企業となった「Google」の存在がある。確かに「Google」の登場は、技術的にも社会的にも衝撃的であった。

キャッシュ機能やシンプルなインターフェース、様々な検索オプションなど、旧来のポータルサイトの１サービスとして位置づけられていたものと比較しても、確かに画期的ではあったが、何より重要なのは「ページランク」という検索結果の評価アルゴリズムである。

ページランクとは、簡単に言えば、Google の基本的な仮定である「より多くリンクされているサイトはより良いサイト」[7] より算出される Web ページの評価であり、より多くリンクされていればより高いページランクを持つということである。このページランクとは、グラフ理論に基づいた厳密な数学モデルであり、情報の内容そのものに立ち入るものではない。しかしながら、そのアルゴリズムが、検索者の意図に適した精度の高い情報検索に効果を上げたという点が、特に技術の側面で高く注目された。

１．２　検索アルゴリズムの限界

しかし、最近この Google の検索結果が、非常に精度が落ちてきているとの印象がある。これは検索内容や分野、領域、目的などにより相違するし、多分に検索者の主観に依存するため、その事実を客観的に示すのは難しい。

例えば、本研究で検証のために用意したいくつかの検索用語のうち、キーワード「消費者問題」で Google 検索をした場合、2006 年 12 月の時点において、検索結果は「約 3,440,000 件」ある。

その中での上位 10 件は、各々表１のようになる。

タイトル	アドレス（ドメイン）	主体	内容
和歌山県消費生活センター～消費者問題とは？～	www.wcac.jp/f/f_01.html	和歌山県消費生活センター	事業案内・FAQ
クーリングオフ・ネットで悪徳商法解約代行/クーリングオフ無料相談	www.cooling-off.net/	法律事務所	サービス案内
金融消費者問題研究	homepage3.nifty.comfinancialconsu/	任意団体	団体案内
兵庫県弁護士会 消費者問題判例検索システム	www.hyogoben.or.jp/hanrei/	兵庫県弁護士会	判例 DB
広告＊消費者問題 Blog	blog.goo.ne.jp/nancy_9	個人	ブログ
JCIC :OPAC へようこそ	opac.kokusen.go.jp/opac/index.html	国民生活センター	文献検索
消費者問題・経済生活	www.chifuren.gr.jp/bunya/syouhisya.htm	全国地域婦人団体連絡協議会	サービス案内
消費者問題解決のすすめ	aokioffice.seesaa.net/	司法書士事務所	サービス案内
Yahoo! 知恵袋ー「消費者問題」に関する解決済みの質問	search.chiebukuro.yahoo.co.jp/listlist.php?dnum=2078297957	Yahoo!	用語･FAQ
Yahoo! 知恵袋ー「法律、消費者問題」に関する解決済みの質問	search.chiebukuro.yahoo.co.jp/listlist.php?dnum=2078297944	Yahoo!	

本来検索サービスを使うには、多くのキーを付加するとか検索オプションを使うなどして、検索者の意図を表現していかねばならない。しかし本キーワードは社会性を持ったものであり、常識的に考えても、公的機関や公益団体などのページが、本来期待される検索結果であると言えよう。

Web サイトやページの優劣や権威などに関しては、一概に言うことはできないが、社会的な問題に対しては、個人よりはその問題に近い所にいる団体や公的見解の方が、検索者の望む適切な検索結果であると言えよう。

しかし実際の検索結果として、１位にあるのは地方自治体によるものであり、また２位は個人運営サイトである。また３位は個

人的色彩の強い（注：あくまでも筆者の主観的判断である）任意団体のものである。

国家機関である「go.jp」は6位に出現しているが、文献検索のサービスであり、キーワードに対する情報源としては余り機能しない。

本来、同キーワードに関しては、特に社会的側面から、経済産業省（www.meti.go.jp）にある消費者政策に関するページや、公益法人「消費者関連専門家会議」（www.acap.or.jp）などが、重要な情報源である。しかしそれらは、検索結果の10位内には登場していない。

ちなみに、余り知られてはいないが、Googleの検索結果は始終変化しており、同一結果の再現は保証されるものではないことを付記しておく。

同じキーワード「消費者問題」に対するYahoo!での検索結果は、「約10,800,000件」であり、その結果の分析を表2に示す。

タイトル	アドレス（ドメイン）	主体	内容
WEBニッポン消費者新聞／日本消費者新聞社	www.jc-press.com	日本消費者新聞社	会社案内
消費者関連専門家会議	www.acap.or.jp	社団法人消費者関連専門家会議	サービス案内
兵庫県弁護士会 消費者問題判例検索システム	www.hyogoben.or.jp/hanrei	Googleと同	ブログ
消費者政策（METI／経済産業省）	www.meti.go.jp/policy/consumer	経済産業省	政策の広報
国民生活センターホームページ	www.kokusen.go.jp	国民生活センター	サービス案内
広告＊消費者問題Blog	blog.goo.ne.jp/nancy_9	Googleと同	ブログ
月刊消費者	www1.sphere.ne.jp/jca-home/getukann.html	財団法人日本消費者協会　普及室	出版物案内

無題・消費者問題の講話、明電舎見学	www7.plala.or.jp/haraguchi_h/syouhisya.htm	個人	ブログ
消費者問題	syousapo.at.infoseek.co.jp/sub13.html	個人	試験対策
(1)消費者問題	www.geocities.jp/tomokun_adv24/A-1-1.htm	個人	試験対策

なおYahoo!の検索の順位付け技術に関しては公開されていないが、Yahoo!のヘルプの「キャッシュを表示するサービスは、Yahoo!Inc.が管理、運営するロボット型全文検索サービスによって提供されています」[8]といった記述や、Yahoo!がデータベース中でキーワードがヒットしなかった場合、Googleにリレーして検索するといった点から推定するに、データベースであるディレクトリ中を語彙を用いた全文検索を行ってある程度のランク付けを行った後、さらにリンクを用いた評価を行っているものと思われる。これらは、本来はYahooAPIを利用して実験することができると思うが、今回は本質的なことではないため、そこまでの検証は行わなかった。

以上のように、非常に単純な実験結果ではあるが、明らかにGoogleの検索精度は、今回のようなキーワードにおいては劣っていると感じられる。

前述のように、Yahoo!は、本来ロボット型ではなく、ネットディレクトリであり、そこに登録されているWebページから検索する。そのため、公的団体などを上位に挙げることができるものと思われる。

しかし、Yahoo!の検索結果においても、それほど精度が高いと言うものではない。実際、1件目の「日本消費者新聞社」(www.jc-press.com)は、偶々キーワードを含んだ記事があるためであり、情報源としては、前述の経済産業省のものが遥かに重要であるにもかかわらず、5位にランク付けされている。またブログ

を含め、個人の運営によるサイトが多数を占めているため、検索結果に情報のレベルに関するばらつきがあることも事実である。

これら一連の検索サービスが精度を落としていると思われる最大の理由は、Weblog・ウェブログ（ブログ）にある。ブログの最大の特徴は、おそらく巷間においては余り理解されていないようではあるが、トラックバックにあると言っても過言ではなかろう。言うまでもなく、Webのリンク機能は、Ted Nelsonによるハイパーテキストを元にした片方向の参照であり、それが自由に情報の相互関係を生み出してきたということは間違いない。しかし、個人の記録であるブログにおいては、片方向のコミュニケーションだけではなく、ブログ相互による記事のやりとりが必要とされることも多々あるとされる。そうした要求に答えるのが、トラックバックである。通常ブログにはトラックバックURLが指定されており、それを用いてトラックバックピングを送ることによって、こうした双方向の情報の通知が成立する。

筆者個人は、それがまさにブログという媒体の限界であり、情報レベルの低さを端的に示すと思っている。元来ブログとは、しばしば日記とも言われるように、個人的な感情の吐露のようなものであり、他人の記事をネタにすることが多いということは、個人的なおしゃべりの連鎖という以上の意義を見出しにくい。またしばしば有名人のブログが注目を集め、「ブログの女王」などとの呼称を持つようなテレビタレントもいる。こうした有名人には、数多くのトラックバックが通知されることが想像される。ちょっと高級なファンレターのようなものと言えばよかろうか。

このトラックバック機能があることで、勢い多くのリンクがブログからブログへ貼られていることであろう。残念ながら、シャノン流の情報技術の元では、我々は情報の意味に基づいた価値を判断することができないことと同様に、リンクの意味、価値を判断

するための手法を持たない。どんなリンクでもリンクなのである。

Google によるページランクにおいては、多くのリンクを持つということが重要な前提であるため、ランキングの上位にブログが登場するという現象が起こっているのである。

今回の実験キーワードのうちの一つである「オントロジー」の検索実験においては、かなり驚く結果となってしまった。Google を使い、上記キーワードを検索した結果、約 18500 件を得ることができた。通常こういう大量の検索結果が出た場合、Google では検索オプション「site:」を用いてドメインを限定したり、あるいはこれは日本語しか用いることができないある種の裏技であるが、「とは」「について」などの言葉を補って、検索の精度を上げていくことをやっていく。

「site:ac.jp」オプションを使った場合、4440 件であり、「site:co.jp」オプションを使った場合は 2030 件であった。よってキーワードは学術用語として認知されていることがわかる。尚こういう場合、企業での認識や製品化状況などを知る目的で「co.jp」に限定して、技術用語や学術用語を検索することが多々あることは付記しておく。

次に、言葉の定義を検索するために、「オントロジーとは」とのキーワードで検索した場合、「4740 件」が得られた。さらに「オントロジーとは site:co.jp」の検索は、578 件を得ることができた。こうした一連の検索実験の結果から見る限り、検索サービスは適切な量に情報を絞り込んで提供しているように見える。しかし最後の実験結果において、ランクの 3 位から 28 位までを、同一人物のブログが独占しているのである。当該ブログの記事は、上位 100 件までの中に実に 44 件を数えることができる。またその他にもブログの記事は 15、6 件ほど発見できた。

特定の Web ページに対して優劣を言うことはできないが、少なくとも論文を書くなど学術的な目的にはこれらのブログを引用

はできない。誰のブログにせよ、それがきちんと校正され管理され、責任を持って提供されている情報であるという評価はし難い。つまり検索結果の上位の半数以上を、使えない情報が占めているというのが、現在の検索技術のある種の実態なのである。

学術用語はまだしも、馴染みやすい一般用語や社会的な事象、トピック、ニュースなどをテーマとした場合、そうした傾向がより高くなるであろうことは、想像に難くない。

「そしてゴミのような意見をブログで公表し、さらにジャンク情報は増殖していく。ネットはレベルの低い同類同士の交信を加速させ、情報のぬるま湯に浸った低レベル人間を大量に作り出してしまった。」これは筆者の意見ではなく、［4］からの引用である。ブログのみならず、掲示板や話題のSNS（ソーシャルネット）などの実体を見る限り、この意見をあながち否定することはできないだろう。

しかし、［4］に言う「ジャンク情報」の存在やブログ技術を否定することもできない。何より、ジャンクであっても情報なのである。

1．3　新たな検索アルゴリズムしか無いのか

元来、Internetというメディアは、その受発信者の規模から言えば、明らかに最大の「マス・メディア」であるが、テレビ、新聞などの他のマス・メディアとの違いは、中枢の存在にある［9］。中枢のあるメディアとは、そのメディアを流れる情報が、発信主体によって何らかの管理がなされているものである。

Internetには、技術的にもまた社会的にも情報を規制、管理する手段は存在しない。そのため、受発信者が容易に増大して行き、外部経済が極大化して行ったわけであるが、それは情報エントロピーの増大をも意味する。つまり、他のメディアとは異なり、受発信者の増加がそのまま利便性の増大を意味するわけではない。

情報の質の低下は、Internetの本質的な摂理とも言える。技術的な予測は付かないが、ジャンクな情報は、明らかに増大していく。[10]では、そういった前提で欠陥のある情報を情報によって修正する試みについて明らかにした。情報そのものの質的な問題について考察したわけであるが、ここでは、情報のいわば量的な欠陥を考察の対象にする。

問題は、現在の標準的なサーチエンジンであるGoogleのコア技術であるページランクアルゴリズムが、現在の「情報爆発」に対応出来ていないように思える所にある。さらにそれはGooleのみの問題ではなく、デフォルトのエンジンとしてGoogleを使っている検索サービスが多いという点にもある。要するに、検索のための有効な代替手段も存在していないのである。

筆者は、[11]において、語彙のマッチングによる検索システムが、必ずしも権威の高い検索結果を挙げるわけではないということを示した。これは実際に、最近Yahoo!でも「Jword」機能として、特定のキーワードとサイトを対応させたデータベースを構築することで対応していることからも、明らかな問題である。つまり、語彙、用語の出現頻度だけで、人間の知識構造を解き明かすことはできない。

こうした問題意識に対して、Googleはハイパーテキスト構造に着目してページランクという検索モデルを作り上げたわけであるが、以上に示したように、その技術も有効とは言い難い状況になりつつある。

また、[1]などで情報の表現ではなく意味そのものを扱うということについて、いくつかの試みを行った。その結果、文書構造を示すためのHTMLタグ情報から、ある程度文書を作成した側の意図を推定し、再構成することが可能であるということが明らかになった。

こうした問題に対する関連技術として、Webシステムの開発者であるTim Berners Leeらの提唱による、セマンティックWeb技術をあげることができる。セマンティックWebとは、ページにその構造や属性に関する「メタデータ」を付加することによって、Webページに意味情報を与えようとする試みである。

これはいわば「Webの再構築」を目指すものであり、未来におけるあるべき姿を模索する試みとして、W3Cなど組織的にアプローチがなされているが、現実化するためにはまだ時間が掛かる模様である。本来文書構造を示すタグ情報からすら、多くの意味情報が推定できるわけであり[1]、メタデータの付加という方法は、「情報を情報によって修正する」という試みとしても有効であろうことは、想像に難くない。

しかし問題は、今ある「情報爆発」であり、何よりも情報がここまで蓄積してしまったということは肯定しなければならない。すなわち「量」を「質」に転化させることはできないのだろうか。

以上を前提に考えると、議論は「新たな検索アルゴリズム」に向かうことになる。しかしこの問題は、検索エンジンの仕様のみにとどまらず、こうした情報エントロピーの増大にどう対処すべきかという課題、ひいては、情報化社会の進むべき方向の問題としても捉えるべきである。

Webを考える場合、その文書のみを考察の対象とするわけにはいかない。Webの文書は、ハイパーリンクにより、文書の製作者以外によって組織化されて行くという点が、最終的に情報エントロピーの増大につながっていくわけである。つまり複数の人間による共同行為によって生み出される情報を考察の対象とせねばならない。以降に、Webの情報の構造を理解することで、情報爆発現象への対処の糸口を考察する。

前述のように情報エントロピーの増大は止められないとするならば、我々には「検索」するしか対処法がないのであろうか、そ

もそも「情報とは探すものなのであろうか？」。

2．検証仮説とシステム仕様

以上の問題意識を元に、以下の仮説を立てた。これは、端的に言えばWebを利用してきた過程における、いわば経験則に基づくものであり、我々検索する側が暗黙的に捉えているWebの知識構造の一部の姿である。少なくとも、筆者はWeb検索をする場合に、以下のような想定の元で利用をしている。以降には、サーバベースの実験システムによって、これらの仮説を検証することにする。

仮説①

「関連する情報は、必ず近いところにある」

多くの検索結果が得られたとしても、それらの情報群において重要度の高い、いわばキーとなるサイトやページは、それほど多くない。さらにそれらは相互に直接リンクが貼られていないとしても、いくつかのリンクを辿ることで、到達することが多々ある。また直接リンクを貼られているページに関しては、当然のことながら何らかの関連性を持つはずであり、その意味では、特定の情報は何らかの原理で組織化されているはずである。その原理はどこにあるのであろうか。

仮説②

「ステータスの高い情報からトップダウンで情報が組織化される」

各検索キーの種類などによって異なるではあろうが、ネットの中では、重要な情報や情報源、公的見解など、いわゆるステータスの高い情報から、より詳細な情報、個人的な色彩の強いページへと構成されているはずである。但し、一般に公的サイトからは、個々のサイトにはリンクが貼られてはおらず、逆リンクによって

その構造が構築されているはずである。つまりリンクそのものは、より抽象化に向かっているはずである。

仮説③
「リンクは何らかの概念関係を示している」

前記仮説と関連するが、ハイパーリンクによって、文書の相互関係が示される。それを語の関係と捉えてみると、語の意味関係が存在するはずである。そこでは、特化、汎化、関連、集約、部分全体など、様々な関係が考えられるが、一般にリンクによって示されるのは、どのような概念関係なのであろうか。Webやサイト、検索語の種類などによってそれらは相違するのであろうか。

2.1　システム仕様

これらの仮説や問題提起を検証するために、サーバ側で以下の仕様に基づき稼動する実験システムを作成し、実際のWebデータを収集し解析することを試みた。昨年度［1］で述べた、特定のWebページのタグ情報を解析する小規模ロボットのエンジンを元に、タグ「<a>」を用いてリンクを辿り、ページ毎にページ内に含まれている情報を収集するシステムを作成した。

実際のシステム開発は、株式会社リバティシステムが行った。開発言語としてはJavaを採用し、サーブレットとして構築したが、［1］でも述べたように、本学におけるシステム環境においては、サーバをグローバルアドレスの元で稼動させることができず、プロキシとの通信を行わねばならないといった制約があった。これは昨年度と状況が変わっていない。

本システムの基本的なアルゴリズムは以下のようになる。

［アルゴリズム］
・あるページ（1）からリンクを貼られたページを（1-1）とする。

さらにそこ(1-1)からリンクの貼られたページを(1-1-1)とする。桁数で(1)からのリンクの世代を示す。

・(1-1)が複数ある場合、つまり(1)から複数のリンクがある場合、(1)のHTML上への<A>の出現順に、(1-1)、(1-2)、(1-3)と通し番号を与える。この通し番号は、ページの識別番号である。

・(1-1)から複数のリンクがある場合、(1-1-1)、(1-1-2)、(1-1-3)となり、さらに(1-2)から複数のリンクがある場合、(1-2-1)、(1-2-2)、(1-2-3)と名前付けをし、リンクされている文書を解析する。

例えばリスト1は、文書「http://econ.keio.ac.jp/」のHTMLソースの一部である。ここには、タグ「<A>」が二箇所ある(各々、①、②で示す)。リスト1の文書を(1)とすると、①からのリンク文書「http://econ.keio.ac.jp/index-jp.html」が(1-1)、②からのリンク「http://econ.keio.ac.jp/index-e.html」が(1-2)となる。リスト2は、文書(1-1)のハイパーリンク部分の一部を出現順に並べたものである。③は(1-1-1)、④は(1-1-2)となる。

[リスト1]

```
<HTML lang=en>
<HEAD><TITLE>Faculty of Economics, Graduate School of Economics, Keio University</TITLE>
</HEAD><BODY>
<IMG height=145 alt="Keio University" src="Faculty of Economics, Graduate School of Economics, Keio University.files/keio.gif" width=145><BR>
<IMG height=120 alt="Faculty of Economics, Graduate School of Economics, Keio University" src="Faculty of Economics,
```

Graduate School of Economics, Keio University.files/titlebar.jpg" width=750>
① <A href="http://econ.keio.ac.jp/index-jp.html">
<IMG height=25 alt="Japanese page" src="Faculty of Economics, Graduate School of Economics, Keio University.files/btj.gif" width=100 border=0></A>
② <A href="http://econ.keio.ac.jp/index-e.html"><IMG height=25 alt="English page" src="Faculty of Economics, Graduate School of Economics, Keio University.files/bte.gif" width=100 border=0></A>

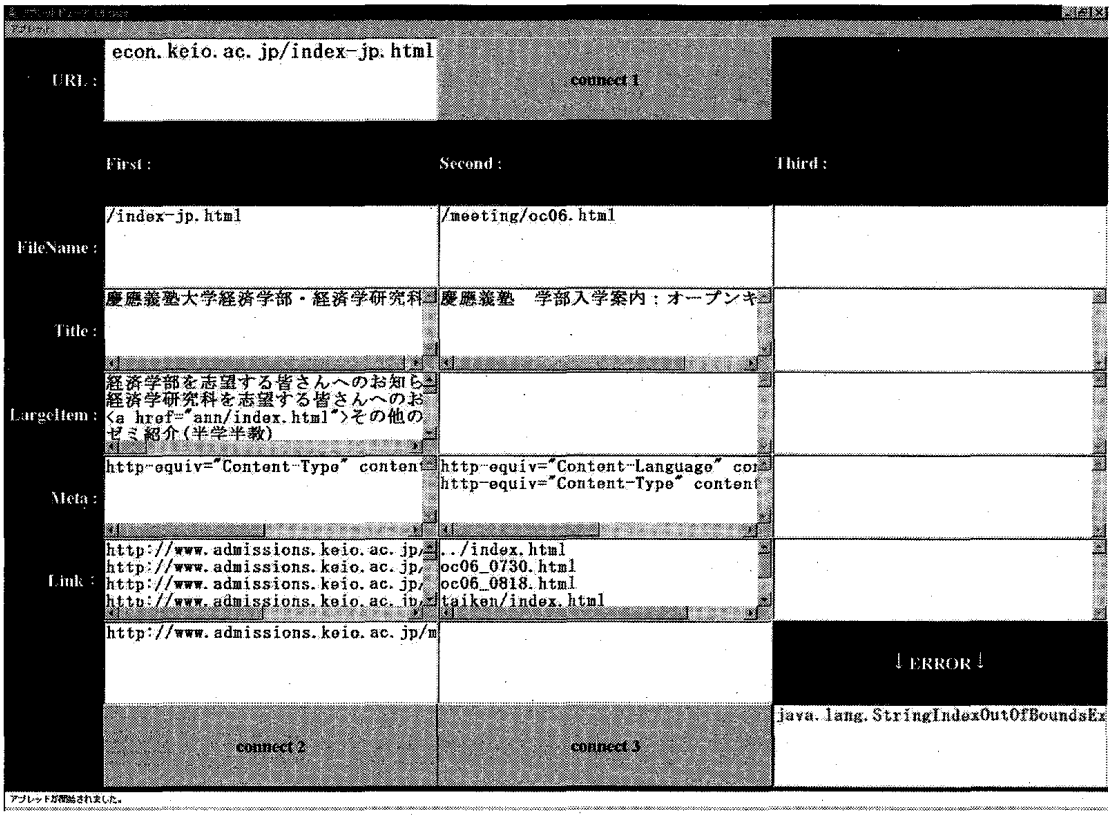
Copyright(c) 1996-, Faculty of Economics, Graduate School of Economics, Keio University </BODY></HTML>

[リスト2]
③ <A class="bar" HREF="/info/dean/dean-fj.html"> 学部長ご挨拶 </A>
④ <A class="bar" HREF="/info/dean/dean-gj.html"> 研究科委員長ご挨拶 </A>
 学部案内
<A class="bar" HREF="/info/staff-list.html"> 専任教員（三田）</A>
<A class="bar" HREF="/info/hiyoshi/profile/index.html"> 専任教員（日吉）</A>
 問い合わせ先

このアルゴリズムを元に、①文書アドレス、②タイトル(<Title>)、③メタタグ（<Meta >)、といった文書情報と、④リンク先アドレス（<A HREF= >)、⑤リンクのアンカー、などのリンク情報、さらに⑥ヘッダ（<H>)、⑦強調（<B>,<String>etc..)

など、コンテンツ情報を抽出した。尚これらの抽出タグの意味や機能、及び解析例は、［１］に示す。

図１に、実験システムのユーザインタフェースのイメージを一部示す。



［１］では任意のページの解析を主眼としたため、解析を開始するページを特に定めなかった。本研究においては、Webにおける情報の構造、関連性を分析するために、いくつかの検索キーワードを定めて検索を行い、その検索結果を起点ページとして解析実験を行った。

具体的にはリンクを辿って文書を取り出し、その文書のタグを元に、文書に含まれている情報を抽出して、その傾向を探った。

但し、サイトのトップページでは主に当該サイト内にリンクが貼られており、同一人、あるいは同一組織に所属する人間によって作られたページが続いている場合が多い。また同じドメイン内

ページへのリンクも同様の傾向があるため、それらはリンクとは考えなかった。

まず実験をしてみて、非常に驚いたのが、ハイパーリンクという手段によって結び付けられたWebページ群の情報量の多さである。

実際にWebを辿っているときにも直感的に感じてはいたが、当初の設計仕様で想定していたリンクデータのデータベースが瞬時に一杯になってしまい、またリンク数のそれこそ指数関数的な量の増加は、しばしばサーブレットのプロセス負荷が高いものとなってシステムの効率低下を来たした。

キーワードの問題もあるが、本検索では「ac.jp」ドメインの論文が多くヒットした。これらの文書は主に学術関係者によるものであり、専門用語や固有名詞、参考文献などに、適切にリンクが貼られているものが多く見られた。こうしたリンクの使い方は、元々のハイパーリンクの発想である、Ted Nelsonの全人類の文書を繋ぐという考え方を体現したものであろう。

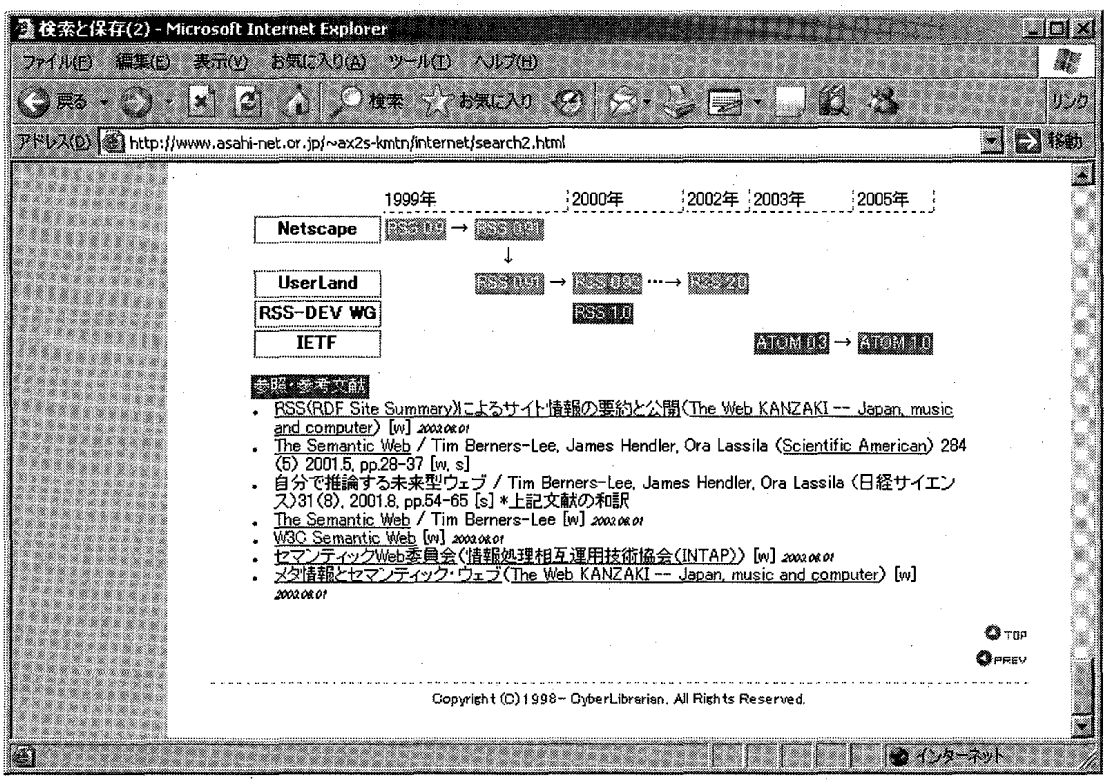
この情報量の急激な増大は、数学的なモデルとしては、当然の結果ではあるが、現実にそれを目の当たりにすると、「情報爆発」という事実を強く実感する。

2.2 解析結果

これらの抽出結果及び解析例を以下に示す。

例えば、検索キー「オントロジーとは」で抽出した「http://www.asahi-net.or.jp/~ax2s-kmtn/internet/search2.html」図2を起点とした場合は、以下のようになった。当該ページは、研究者個人によるサイトであり、Googleではランク10位、Yahoo!では同サイト、同一主体によるページが、17位、28位にあるが、当該ページ自体は100位以内には登場しなかった。また当該ペー

ジへのリンクは4件、内2件は同一サイトからのリンクである。



当該アンカータグ「<a　href=」は、79個ある。うち、59個は自ドメイン内ページへのリンクである。つまり他サイトへのリンクは20であり、この数は一見少なく見えるが、筆者の実験結果では、当該Webの外部へのリンクはかなり多いほうである。今回の実験では、主に技術、学術系のキーワードを使ったため、ハイパーリンクに関しては、その機能や趣旨の理解が正当なWebページが多かったと思われる。母数が少ないので断言はできないが、それでも外部サイトへのリンクは各ページ平均で、5から6程度であろう。本来であれば、ページ単位ではなく、文字数などの情報量を単位にすべきとは思われる。

各リンクデータを分析すると、Webの標準化団体である、「www.w3.org」が7で最も多く、国内の技術組織「www.net.intap.or.jp」や関連団体などを含めると、過半数を占める。当該サイトと同様な研究者へのリンクは4つある。

次に、研究者サイトを中心に、検索キーワードそのものや関連語の分析を行った。その結果、検索語（英語を含む）を直接含むサイトは、7つあった。研究者サイトの内の一つ「http://www.kanzaki.com/docs/sw/」を例にすると、

<meta name="keywords" content="メタデータ,HTTPヘッダ,XML名前空間,RDF,Dublin Core,RSS,Semantic Web,オントロジー" />
<h2>語彙とオントロジー</h2>

といった形で、metaタグや「<H」タグなどが用いられており、強い関連性があることがわかる。また「<a」タグでも以下のように使われているが、自Web内ページへのリンクであることが注目される。

ウェブに存在するものとその関係の定義：ウェブ・オントロジー言語OWL
日本語ウェブ・オントロジーの試み：EDICT + WordNet

またさらに、何らかの強調を行っているタグ「Strong,Font,B,H」などを抜き出し、文書の粗いサマライズを試みて行った。以下がその一部である。

<h1>メタ情報とセマンティック・ウェブ</h1>
<h2>リソースとメタデータの表現</h2>
<h2>語彙とオントロジー</h2>
<h2>メタデータの応用と提供</h2>

ここで抽出された用語は、元々の検索語である「オントロジー」の関連用語や周辺概念である。その概念間の距離を測るための科学的な手法を今回の実験では用意しなかったため、最終的にどういった原理によってWebページが組織化されているか、数的な評価はできなかった。

様々な検索語を用意し、多くのWebページに対して分析を行ってみたが、結論的に言えば、実験結果に対して、統計的な処理が完全にはできず、傾向を定式化することはできなかった。

2．3　Web上のリンクに見られた傾向

発見できたのは、以下のような傾向である。

①　自サイト内のリンクの場合

自サイト内の別なページにリンクが張られる場合、一般的に元の検索語から大きく外れることは少ない。またリンク付けは、特定の概念や用語などを詳細化する役割で使われることが多く、リンクの世代が深くなっても、何らかの関連性を持った情報が存在する。ただし、トップページからのリンクの世代距離が近くなるほど抽象度が高くなる傾向があるようである。

②　他サイトへのリンクの場合

他サイトへリンク付けがある場合、具体化、詳細化する例は少ない。より抽象度の高い情報にリンク付けされるのが一般的である。特に、情報の権威付けや正当性の保証、根拠などの役割のものが多く見られた。他サイトへのリンク先に、検索語がそのまま含まれる例は、おそらく語彙の抽象度の問題もあるとは思われるが、比較的少ない。これも正確なデータではないが、おそらく2世代以上のリンクで、相当概念間の距離が深まるようである。

さらにもう一点、ブログに関する解析結果に関しても指摘しておく。ブログは元々個人によるものであるため、ここでは具体例を示さないが、文書構造として特徴的なのは、ブログデザインにもよるが、自 Web 内へのリンクが非常に多いという点である。平均すると、200 以上はあるようである。また他サイトへのリンクはそれに比べると遥かに少なく、ほぼリンクは一世代で検索語から離れてしまっている。おそらくこういった点が、検索エンジンの精度を落としている理由の一つでもあろうと思われる。

3．まとめ

元々検索したキーワードが、何世代のリンクで離れていくか、リンクは抽象化に向かうか、あるいは具象化に向かうのかといった興味から始めた実験だったが、数的な統計処理だけでは傾向を全て明らかにすることはできなかった。

結論として、まず Web の情報は、ページ単位では捉えることができない。ページはあくまでも媒体の構造であり、情報や知識の構造とは違っている。自サイトと他サイトのリンクで垣間見えたように、リンクには明らかに旧来の紙、ワープロといった媒体では明確化できなかった原理が働いている。それは、自然発生的な知識の構造であると言ってよいであろう。確かに情報は爆発しているのかもしれないが、決してカオス的な、あるいは無秩序なものではなく、量に埋もれて見出しにくいが、ある種の自己組織化が起こっているように思える。

但し、情報爆発現象が以降も継続するとするならば、「探す」といった方法論では、もう対応できないのは自明である。検索者が得たいのは、「語」そのものではなく「情報」なのである。データに意味を与えることが情報処理とするならば、雑多な Web データに、意味を与えながら情報の再構成をするといった方法論もありそうである。それは単純な「検索」ではなく、与えられた情報

を用いた、情報の新たな創造、おそらく類似設計的なものではあろうが、によってWebから「情報」を得るといった手法となるであろう。次年度以降は、そうした方向性について探ってみたいと考える。

本研究を行うにあたり、株式会社リバティシステムの、阪田、大林、石橋の諸氏に協力をいただきました。またゼノンコンサルティング株式会社の福田氏には、常に筆者の研究に有益な示唆を与えてもらっています。併せて感謝いたします。

【参考文献】

[1]タグ情報を用いたWeb文書解析システムの試作とその知見、春木良且、国際交流研究、フェリス女学院大学国際交流学部紀要No.8、2006
[2]HOW MUCH INFORMATION 2003?、http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm
[3]情報融合炉：情報爆発時代のIT基盤＝情報爆発と情報大航海プロジェクトについて、データベースとWeb情報システムに関するシンポジウム（DBWeb 2006）―情報爆発時代に向けて―資料、2006
[4]低度情報化社会 Ultra Law-level Information Society、コモエスタ坂本、光文社、2006
[5]Web2.0が殺すもの、宮脇 睦、洋泉社、2006
[6](仮) 大胆な事業再編とイノベーションを通じたダイナミックな産業構造改革の実現，経済産業省産業構造審議会新成長政策部会中間とりまとめ案、産業経済省、2002
[7]Google キラーの登場、http://japan.internet.com/busnews/archive/、Webビジネス
[8]Yahoo! ヘルプ、http://help.yahoo.co.jp、Yahoo!
[9]情報って何だろう、春木良且、岩波書店、2004
[10]ネット上の掲示板による口コミ現象とその対処―消費者と企業の非対称関係に関する一考察―、春木良且、国際交流研究、フェリス女学院大学国際交流学部紀要No.7、2005
[11]知的サーチエンジンの開発に向けて―分散環境における知識の構造とその再利用に関する一考察―、春木良且、国際交流研究、フェリス女学院大学国際交流学部紀要No.4、2002