知的サーチエンジンの開発に向けて

(Towards the Developing yet another Intelligent Search Engine)

－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－

－An idea for re－use of the knowledge under distributed environment－

春木　良且

0．　はじめに

近年、Internet に纏わる様々な技術が提起され、その応用が試行されてきている。90年代以降アメリカの政策転換により Internet の商業使用が解禁されてからは、特にディジタル通信技術を商取引の各フェーズに利用するいわゆるeコマースが、少しづつではあるが実績を上げてきている。また、通信技術を利用した商業活動全般を、ビジネスモデルと称して法律上の権利（無体財産権）を認める動きもある。

これらのいわゆるeビジネスを技術面から見ると、そこで使われている主要なアプリケーションとしては、まず WWW サービスが想定されている。一般に Internet と言えば、WWWサーバによる Web を意味するものと捉えられており、そうした意味から、WWW は Internet のキラーアプリケーションとも呼ばれている。(注1)

確かに、世界的な規模でマルチメディア情報を容易に提供できる情報通信手段はそれまで皆無であったため、その意味では WWW サービスが Internet の代名詞となるほど注目されるのも当然ではあるといえるかもしれない。

こうした状況において痛感するのは、後述の V.Bush による1940年代の問題提起に明らかなように、氾濫する大量な情報に対してどのように対応すべきか、といった課題である。日々刻々とその数字が変化しており、また接続形態もさまざまなので、その正確な数字を得ることも困難であるが、WWW サービスを提供するサーバ数は、若干古い数字であるが、1998年6月現在では全世界で約241万台と推定される。また、それらのサーバから提供される HTML ペー

ジ数は、1997年12月現在で3.2億と推定されている。もちろん現在はもっと大きな数になっているはずである。

問題は、それらのページの全てが、Internet に接続している端末であれば、理論上はどこからでも容易に参照することができるということである。要するに我々は、突如として世界規模のマルチメディア情報の洪水の中に放り出されてしまったと言えるのである。

いわゆる情報化社会における特有な現象として、「情報エントロピーの増大による情報の欠乏」がある。情報化社会以前においては(注2)(1)、情報は早く入手し、さらに独占するということが重視されていた。ワーテルロースクープとネイサン・ロスチャイルドの行動など、そうした情報化以前の社会における情報の特質を示す事例は様々に存在する。しかしながら情報メディアが一般化する現在においては、特定の情報を誰よりも早く入手し独占するということはできない。ある罪を犯した未成年者の写真が一瞬にしてネットの世界に伝わってしまったことなど、記憶に新しい。

多くの情報が容易に入手できるようになると、むしろ逆に真に必要な情報が、多くの情報に埋もれてしまい、実質的に我々は情報の欠乏状態と同じ状態に陥ってしまう。iモードなど携帯メールシステムが増大する SPAM やジャンクメールによって、実質的に機能マヒを起こし、管理システムを導入せざるを得なくなった例などは、まさにその情報エントロピーの増大が人々に情報の欠乏状態を作り出す例であろう。

「情報はいくらでも持つことができるが、ある閾値を境に手持ちの情報を増やすメリットは減衰し、われわれは情報過多が原因のストレスに悩んだり、情報的に混乱したり、かえって物事が理解できなくなったりする」という指摘もある(2)(3)。

情報エントロピーの増大をいかに抑止するかということは、いわゆる情報化社会における重要な政策的な課題であったと言って過言ではなかろう。「情報エントロピーの肥大抑止力として古代には宗教があり、近世では国家権力による情報管理があった。現代では高水準の教育に支えられた人々の情報選択能力と、政府機関の利益選択によって情報エントロピーに制限が加えられている。」という(4)の指摘は、Internet の登場以前のものであるが非常に興味深い。

以上から明らかなように、現代における増大する情報エントロピーを体現する存在である Internet 上の WWW サービスにおいて最も重要な課題は、Web ページの製作や Web サーバの構築などではない。無数に存在し、さらに拡張を続けている Web ページをどのように整理し、それらをどのように適切なユーザに対して適切な情報として提供するか、すなわちそうした情報を提供するサーチエンジンこそ、WWW を支える最も重要な技術である。しかしアメリカにおける状況と比較すると、日本では新しいサーチエンジンに関する研究は殆ど行われていない。現状の商用エンジンの比較などが散見できる程度である。

サーチエンジンは、Web 中心の現在の Internet 環境においては、OS 以上に重要な基本ソフト、すなわちインフラストラクチャである。以上のような問題意識をもとに、本稿では、サーチエンジンの概要と要素技術に関して述べた上で、筆者が現在仕様をまとめながら外部企業とともに開発を行っている知的サーチエンジン（memSerch）に関する方向性について述べる。

1．知識再利用システム memex とその影響

発表当時、その真価や意義などが理解されずに埋もれてしまったものが、時代を経て新たに評価されるといった例は、音楽や絵画など感性を重視する芸術の世界では枚挙の暇も無いほどに見られる。無理やり T.クーン的な言い方をすれば、アノマリーが生まれなかったためにナチュラルサイエンスにより駆逐されてしまったパラダイムの萌芽とでも言えるだろうか。

芸術の世界における重要な評価基準たる人々の感性は、時代を経るにつれタフなものとなって行く。卑近な例ではあるが、最近では男性が化粧したりピアスをしたりしても、誰も奇異には捕らえなくなってきている。これは我々の感性が耐性を身に付けたのであり、新たな感性によって過去を振り返ると当時奇異だったものが現在の目からするとさほど奇異には感じられないものなど様々に存在するだろう。

しかしこと技術においては、そうした例は珍しい。端的に言えば、工業技術は全て先行する技術に対するアンチテーゼとしての側面を

持つ。そのため後発の技術によって追い抜かれていくという、いわゆるエスカレーションと呼ばれる現象が生起する(5)。時間の流れから見ると、技術の世界では後発のモノは常に優れているのである。我々が工業製品を購入する場合、一抹の躊躇を感じるのは、こうした摂理が背景にあるのは否定できないだろう。このように技術は段階を経て発達を遂げていくものであるため、新しい技術はまず先行する技術に対する問題提起という形で現れてくる。工学の世界では、モノという対象が存在し、それに対する工業製品としての評価や問題意識が明確であるため、突拍子も無い技術や考え方が生まれにくいのである。これは逆に、未来予測など工業技術に感性が関わってくる場合には、こうした例が起こり易いとも言えるかもしれない。

そうした歴史に埋もれてしまった例の内の一つとして(注3)、本稿ではV.Bushの論文「As We May Think」とそこで提起されているmemexと呼ばれている情報システムについて取り上げる。

アメリカの近代技術史において見逃すことのできない重要な人物であるヴァネバー・ブッシュ（V.Bush・1890年～1974年）は、その業績や後世への影響に比べて、驚くほど個人に関する情報が少ない人物であり、一般にはあまり知られていない。それは彼がアメリカの戦時研究と深く関わっていたということと無関係ではあるまい。実際電子計算機の理論的なモデルを提起したアラン・チューリング（A.Turing）にしても、第二次大戦における連合国の暗号解読技術に関するプロジェクトであるコロッサスに携わっていたことから、1970年代に到るまでその業績の詳細が明らかにはされていなかった。戦後50年以上経て、各国において戦時中の様々な重要文書が公開されるようになり、多くの史実が明らかにされてきている。

V.Bushは、(7)などのバイオグラフによれば、戦前から米国マサチューセッツ工科大学(MIT/Massachusetts Institute of Technology)の副学長、カーネギー研究所長、カーネギー協会会長、米国航空学諮問委員会委員長などを歴任している。さらに、第33代大統領フランクリンD.ルーズベルトの科学顧問を勤め、トップ・サイエンティストとして戦時研究を統括する重要な立場にあった。「専門化した科学技術を平易な言葉で解説する人材」として絶大の信頼を得ていたようである(8)。

さらにブッシュは、アメリカの近代社会、経済を特徴付ける軍産学複合体による研究体制を提案し、1941年にはそれを推進する科学研究開発局(OSRD)の初代局長にも任命されている。その体制のもとで、第二次世界大戦中の「マンハッタン計画」や、戦後の「ARPANET」も実現されているのは言うまでもない。マンハッタン計画がディジタルコンピュータを生み出し、ARPANET が後の Internet のバックボーンとなって行った事を考えると、現代の最も重要な 2 つの技術の誕生に携わっている、とてつもない業績を上げている人物である。特に、1942年 8 月に開始された原子爆弾製造のための研究開発計画、通称マンハッタン計画は、アメリカ政府による初めての巨大科学プロジェクトであり、アポロ計画や SDI 構想など以降の国家的巨大科学プロジェクトの見本となったとも言われている。ちなみに、同計画において、日本に対する原爆投下の候補地を選定したのが、現在のディジタルコンピュータを実現した V.Newmann（フォン・ノイマン）その人であったことは、広く知られた事実である。余談ではあるが、MIT で教職に就いていた時代の教え子に、通称「シリコンバレーのゴッドファーザー」とも言われるフレディック・ターマンが居り、ブッシュ流の科学技術観が、ターマンを経由してシリコンバレー、すなわち民間企業にも根付いているとも言えるだろう。(注 4)

V.Bush はそうした多くの人間と資金が集まる大規模プロジェクトを統括する立場から、様々な論文やエッセイなどを通して、科学技術に対する啓蒙活動や提案などを行っている。例えば、1945年 6 月に、「科学：その終わりなきフロンティア(Science : The Endless Frontier)」という大統領宛レポートを小冊子として刊行している。これは連邦政府が科学的基礎研究に公的資金を投入することを提案する内容で、戦後のアメリカにおける研究振興政策の基本路線を示したものと言われている。それを受けて1950年に、NSF(全米科学財団)が設立されている。この NSF は、端的に言えば科学技術研究を政府が助成ししかるべき成果として公開をする、科学技術に対するコーディネータ、プロデューサ的な役割を果たす機関である。特に、民間企業ではなかなか実行ができない基礎研究への投資を一貫して行なっており、その存在はアメリカの技術開発において、現在

も大きな影響力を持っている。

例えば1985年には、スーパーコンピュータの有効利用のため、プリンストン大学、コーネル大学、カーネギーメロン大学、イリノイ大学、カリフォルニア大学にスーパーコンピュータ・センタを設立し、同時にこれらのスーパーコンピュータを接続する NSFnet（全米科学財団ネットワーク）を構築している。1989年には、その NSFnet が ARPAnet を吸収し、これが現在の Internet のバックボーン回線となっている。NSF は現在、全米の約150の大学と約50の企業が参加する次世代インターネットの研究プロジェクトである、いわゆる Internet2 に対して出資を行っており、現在もアメリカの科学技術政策に対して影響力を持ち続けているのである(10)。

同1945年に、V.Bush は「As We May Think」と題された一般向けの論文（あるいはエッセイ）を、「アトランティクマンスリー（The Atlantic Monthly）」6月号に発表している。また同論文は、同年の「ライフ」9月10日号に再録され、後述するように後の情報技術の方向性に対して重要な役割を果たしている。

ちなみに原雑誌である「アトランティクマンスリー」6月号自体は、時代的なこともあり、国会図書館にも所蔵されていないようである。日本には原雑誌自体が存在しないとも言われている。「The Atlantic Monthly」は、当時の代表的なオピニオン誌と言われ、現在も「The Atlantic」として刊行されている。

同社のWeb「http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm」には、「As We May Think」が電子化されて全文が掲載されており、同論文は現在いつでも読めるようになっている。

同論文は、多くの技術者を統括する立場から、いわゆる情報の洪水に対処するという問題意識によって書かれたものである。ちなみに同論文に付記されているアトランティック・マンスリーのコメントによれば、「（ブッシュは）科学の軍事利用に従事している米国の約6,000名にのぼる指導的な科学者の活動を統括してきた」とのことである。

当該論文では、その大量の情報を扱うために memex（メメックス）と称するある仮想的な情報処理機器を提起しているのであるが、

興味深いのはそこで明らかにされているその機器は、我々が現在想起するような情報処理機器、すなわちコンピュータではない。

「私的なファイルや蔵書を機械化したような、未来の個人用の装置について考えてみよう。呼び名が必要なので、適当にメメックスとでもしておこう。メメックスは、個人が自分の書籍や記録、通信を保存するための装置で、機械化されているので驚くべき速度で柔軟に検索することができる。人間の記憶を拡大し、補う個人的な装置といえよう。」(11)(12)

memex とは、端的に言えば人間の頭脳の動作をモデルにした、連想による情報検索システムである。具体的には、個人がそれぞれの蔵書、記録、手紙などをマイクロ・フィルム化し、人間の脳細胞が複雑な回路網で連結されているように蓄積し、そしてその中から必要な内容を自由かつ高速に検索し、スクリーンに映し出して閲覧できるような個人用の機器といったイメージである。

そこには、情報のディジタル化といった概念は含まれていない。実現性のある提案とするために、当時の技術レベルによりマイクロ・フィルムが想定されたとも言われているが、おそらくディジタルコンピュータの考え方自体が、当時の最先端の軍事機密であったということも大きな要因であると思われる。

また、ブッシュは元々微分解析機（アナログコンピュータ）の研究を行っていたこともあり、当時の先端的な情報の記録技術であるマイクロ・フィルムをもとにした検索システムに大きな可能性を感じていたようである。実際アメリカにおいては、1930年代頃から科学技術やマスメディアの発展に伴った情報量の急激な増加が起こりつつあったようであり、本に代わる情報の保存媒体としてマイクロ・フィルムに期待されていたという時代背景もあると言えよう。いずれにせよ、1945年の時点でコンピュータをも含んだ軍事研究を統括していた立場にあった人間が、ディジタル技術を知らなかったということは無かろう。

技術的に見れば、ディジタルコンピュータが理論的なベースにしている「チューリングの万能機械」は、人間の思考を（あくまでも計算という思考活動に限定されるのではあるが）機械に肩代わりさせるものとして研究されていたのに対し、memex は他者が作成し

た情報を効果的に利用するシステムであり、いわば知識の再利用システムなのである。その意味において、コンピュータとは決定的に使用者との位置関係が違っており、あくまで人間の知的活動の支援を行うものと言える。memex には、“memory extender”というニュアンスもあるが、模倣するという意味の単語 mimic がその語源であることでも明らかなように、情報を自ら作成するのではなく、人間の思考パターンを模倣する機械である。

尚、mimic とは、日本が戦後高度成長を遂げた時代において、欧米人が日本の工業製品を攻撃する時の常套句として用いられる言葉になって行くのは皮肉なことである。また1968年に作成された後の科学技術を大胆に予想した映画、「2001年宇宙の旅」の中で、狂っていくコンピュータ「hal」を、BBC のアナウンサーが mimic と表現するとの指摘(8)もある。

この memex のアイディアは、当時としてはかなり SF 的であり、論文自体がいわゆる工学的な論文の体裁を取っていなかったため、研究者にはあまり重視されなかったようである。そのためブッシュの論文は、皮肉なことにこの論文中で自ら例示している様々な学説と同じように、多くの論文や記事などの中に長く埋もれたままになってしまう運命を辿ったのである。しかし、非技術者や若い研究者にとっては大きな関心を呼び起こしたことも確かで、例えば、SF 小説にこの memex が取り上げられることもあった。著名な SF 小説家である R.ハインラインの「深淵」(1948)の中には、未来の人類の知識活動が描かれており、その脚注にブッシュの memex が紹介されている。こうした memex のイメージが、文化的なミームとして数十年後に現実化をしていくことになる。

注:「知識の整理、取り出しやすさは、昔も今も変わらず、もっともスピードを必要とする問題だ。新人類(ニューマン)については、完全に系統づけられた記憶がほとんど問題を解決し、記録の保存や読み書きのほとんどを、そして特に、読みなおして時間を無駄にする面倒を、ほとんど不必要にしている。... 新人類たちは、数限りない書類の山に埋もれることはない。かれらは決してメモをとらないのだ。」

本論文が情報科学の世界で再度顧みられるようになるのは、時代

が下って1980年代頃からである。具体的な切っ掛けは、その論文からの影響を示唆する様々な技術的成果が1980年代に具体化してくるからであるが、その影響は、私見では2つの技術的な方向性に区別できるものと考える。個人用の情報機器というハードウェア的な側面と、知識の整理と再利用、連想処理といった、ソフトウェアの側面がそれに当たる。

1-1. memexと個人用情報機器

前者の側面、すなわちmemexのハードウェアとしての影響は、PC（パーソナルコンピュータ）に繋がっている。元々パーソナルコンピュータのアイディアは、当時Xeroxの PARC（パロアルト研究所・Palo Alto Research Center）にいたアラン・ケイ（A.Kay）が、「コンピュータの専門家以外をユーザとして想定したデスクトップ・パーソナルコンピュータ」[13]という発想であるDynabook（Dynamic Personal Media）を元としている。

アラン・ケイは、そのDynabookのアイディアの後世への強い影響力から、“Pied Piper of the Computer”とも、ハイテク・ラスプーチンとも呼ばれている。しかし「パーソナルコンピュータ」という用語自体、1968年にARPAが主催した大学院生会議において、当時ユタ大学大学院の学生だったアラン・ケイ本人が初めて使ったものとされ、1980年代以降急激に花開くパソコン技術の方向性を決めたとして、無視できない人物である。

このあたりの事情はいわば伝説化しており、具体的によくはわからないのではあるが、ユタ大学に在学中に、アラン・ケイは図書館から「アトランティック・マンスリー」を借りて、前述のブッシュの論文を読んだとされている。また子供の頃に、R.ハインラインの「深淵」を読んで大きく影響されたという説もある[8]。

いずれにせよ、ダイナブックそしてパーソナルコンピュータという技術概念は、元来自動計算機たる当時のコンピュータの範疇を越えたものである。「パーソナル」という用語からも明らかなように、極めて個人的な作業の色彩が強い「思考」という知的行為を前提とした、いわば「思考のためのメディア」というべきものであり、その意味からも、memexの影響を大きく受けたものであると言えよ

う。

「コンピュータは、他のいかなるメディアー物理的には存在しえないメディアですら、ダイナミックにシミュレートできるメディアなのである。さまざまな道具として振る舞う事が出来るが、コンピュータそれ自体は道具ではない。コンピュータは最初のメタメディアであり、したがって、かつて見た事もない、そしていまだほとんど研究されていない、表現と描写の自由を持っている。それ以上に重要なのは、これは楽しいものであり、したがって、本質的にやるだけの価値があるものだということだ」(13)

パーソナルコンピュータへのmemexの影響は、さらにダグラス・エンゲルベルト（Douglas C. Engelbart）という人物にも遡ることができる。エンゲルベルト自身は、マウスの発明で知られているが、マウスというデバイスを発明したと言うよりは、現在 Windows など PC では標準になっているユーザインタフェースを包含した、いわゆるコンピュータ環境という概念を整理したと捉えた方が適切であろう。

第二次大戦中、アメリカ海軍のレーダ技術者を勤めていた彼は、日本降伏の直前にフィリピンレイテ島の赤十字図書館で、ブッシュの論文が再録されていたライフ９月10日号を手にする。戦後、スタンフォード研究所（SRI）在職中に、レーダ技術の経験から情報をスクリーンに表示させ、レバーやキーボードで操作するようなコンピュータ、すなわち memex とコンピュータを結びつけたようなシステムNLS(oN Line System)を現実化する。そのアイディアは、国防総省の ARPA（高等研究計画局）から評価され、1950年代から開発されていた防空システム（SAGE）に、敵機をレーダで捕捉し迎撃するサブシステムとして採用されることになった。1964年からNLS には巨額の援助が開始されることになるのであるであるが、その成果発表として、1968年12月９日にサンフランシスコで開催された Joint Computer Conference においてエンゲルベルトは、そのNLS のデモンストレーションを行った。このデモは、コンピュータスクリーンの画面に映し出された映像を対話的に操作するような、現在のパソコンの操作と相通じるるようなものであったが、当時としては画期的なものであり、そのデモ自体情報技術の世界では

伝説となっている。

そのデモの観客の一人に、当時大学院生だったアラン・ケイがいて、大きな影響を受けたことを後に明らかにしている。

“The way we had been thinking about it was sort of Doug Englebart's view that the mainframe was like a railroad, owned by an institution that decided what you could do and when you could do it. Englebart was trying to be like Henry Ford. A personal computer as it was thought of in the sixties was like an automobile. ”(14)

1-2. memex のソフトウェア技術への影響

ソフトウェア技術における memex の影響は、パーソナルコンピュータよりもある意味重要であるかもしれない。

前述の様に、memex の情報検索方式は、人間の頭脳の動作をモデルにした連想に基づくものであり、情報の関連性や、関連する情報の経路などに関する問題意識が明らかにされていた。

「この連想索引法の基本的なアイディアは、一つの事項が決まれば、もう一つの事項を即座にまた自由に選択できるようにするという点にある。これが memex の本質である。二つの事項を結びつける過程が重要である」(11)(12)

ここで言う「二つの事項を結びつける過程」を明示的に包含する情報メディアを提案したのが、テッド・ネルソン（Theodor Holm Nelson）である。

テッド・ネルソンは、コンピュータとそれによる社会変革を推進しているコンピュータ科学者且つ社会運動家であり、「コンピュータ・リブ」「ホーム・コンピュータ革命」などの著作がある。1960年代に、当時ハーバード大学大学院の社会学修士課程の大学院生だったテッド・ネルソンは、「ハイパーテキスト」という概念を提起し、さらにそれを実現推進するプロジェクトとして、1967年「ザナドゥ」を提唱している。(15)

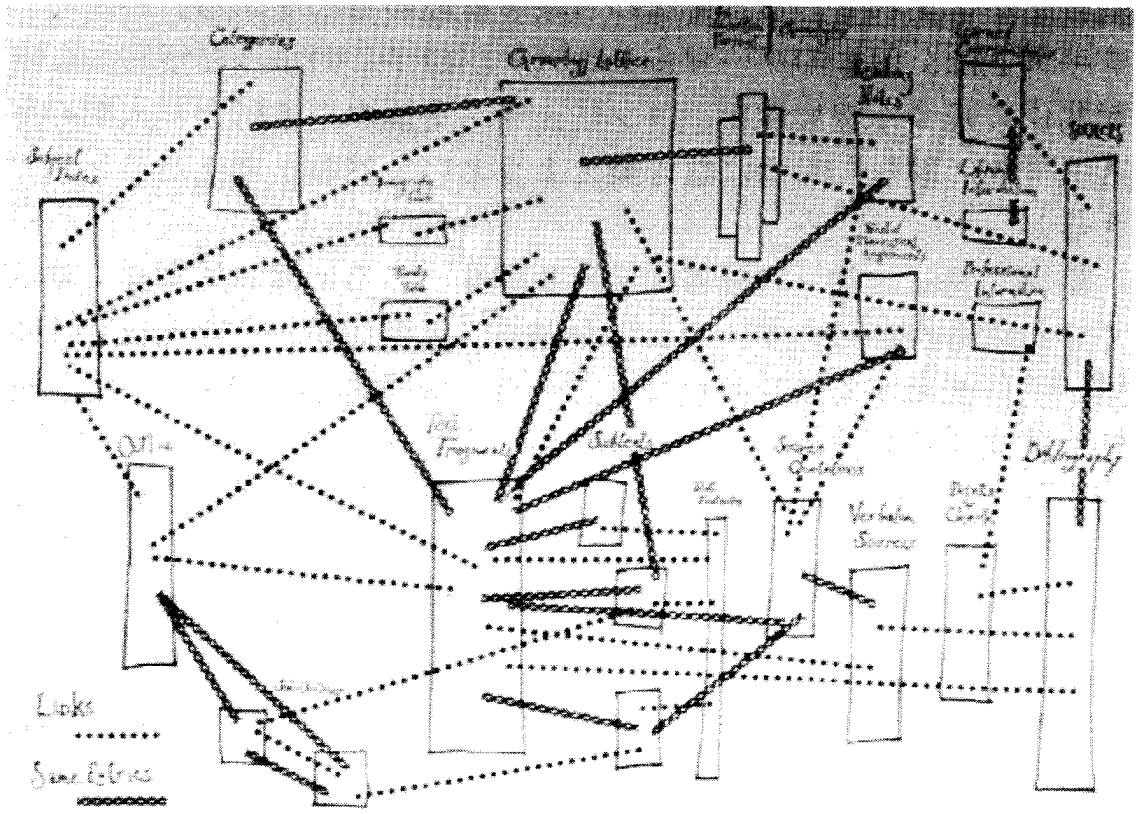
「連続性は必要ではない。思索の構造自体、連続ではない。アイディアがからみ合ったシステムだ。...アイディアというものは、必ずしもどれが先でどれが後と決まっているものではない。それに、アイディアを表現するためにひとつながりの列にする作業は、任意

性が多く複雑なプロセスである」(12)

人間の知的作業に対するこうした問題意識は、思考作業が他の知識情報と絡み合った構造をしているという捉え方で明らかなように、ブッシュによる「知識の再利用」を基にしたものである。実際、テッド・ネルソン自身“As We Will Think.”というタイトルで、そのハイパーテキストシステムのアイディアを、1972年に発表してもいる。(16)

ハイパーテキストとは、技術的に言えばノードとリンクから構成されている有向グラフ構造のテキストである（図1)。従来の紙媒体に表現された（ノーマル）テキストは、1次元の線形構造をしており、時系列に従って情報の伝達が行われる。ページという概念や、目次や索引を使った検索は、こうした線形構造を支援するものとして機能している。しかしこうした時系列構造は、あくまで紙という情報メディアの持つ（物理的な）制約に基づいたものでしかない。我々人間の思考は、ブッシュの指摘に見るまでも無く、並列的であったり、または発散したり、逆に収束したりする。要するに、非線形構造を持っているのであり、情報メディアは（この場合、文書を扱うメディアである)、人間の思考の持つ非線形性を表現、あるいは支援するものとして機能しなければならない、というのが大まかなテッド・ネルソンの考えである。こうした問題意識に基づいて、非線形性を有向グラフによって表現したものが、ハイパーテキストということになる。ハイパー（Hyper）とは、「過度」「超…」といった意味の接頭語である。大規模なマーケットを、スーパーならぬハイパーマーケットと呼ぶこともある。ハイパーテキストは、日本語では超文書と呼ばれる事もあるが、それでは若干意味がわかりにくいのは否定できない。

図1はテッド・ネルソン本人自筆によるハイパーテキストの結合構造をイメージしたもので、1965年に書かれたものを(17)より転載する。



ここで注意したいのは、ノーマルテキストは物理面と論理面での多層的な階層構造になっており、時系列の操作、すなわち線形性が重要な指導原理となっているという点である。我々が紙媒体でノーマルテキストを扱う場合には、まず上から下（あるいは右から左）といった「順序」に従い、情報を表現し解釈する。さらに紙媒体は複数の「ページ」概念によって統合されるが、そこには明示的に順序概念が与えられる。これらは、紙媒体に存在する物理的な「順序」概念であるが、それらはさらに、論理構造としていわゆる章立てによって整理される。大項目から中項目、そして小項目という章立ては、順序構造に包摂関係を導入するものであると解釈することができる。我々は章立てによって、順序関係に包摂という意味情報を付加して解釈する。この章立ては、あくまで紙の物理的な連なりに、著者が任意で意味を与えたものであり、その意味から論理的な順序概念である。その包摂関係は、あくまでも順序関係を前提としており、それを覆すようなものではない。このように、テキストの章立ては、ブッシュが指摘した「図書館のインデックス方式」と同じように、本来階層構造になっており、そうした情報の整理方式には時

間概念が含まれている、すなわち線形構造のものである。ハイパーテキストは、旧来のノーマルテキストが持つこのような重層的な階層関係に基づく線形性を、コンピュータ支援のもとにより人間の思考モデルに即したものに置き換えようという試みと考えることができる。

結局問題は、我々の思考は元来並列性、あるいは非線形性を持っているにもかかわらず、それらを表現する媒体が、その非線形性を明示的に表現する術を持たないということである。Memex は、端的に言えばその非線形性を、図書館に代表されるような「文献の整理」「情報の整理」に適用することで、知識の再利用を目論んだものであった。同じように、ハイパーテキストの概念も、非線形性を線形構造である紙媒体にどのように実現するかといった試みであった。情報機器をメディアとして捉えるこうした発想は、前述のパーソナルコンピュータにも通じるものであり、その意味からもブッシュの memex からの影響が見て取れる。

2．ハイパーテキストシステムとしての WWW

90年代に入り、Internet という地球規模の情報通信が突然我々市民に開放された。それは前述のように、1989年に NSFnet が ARPAnet を吸収したことに端を発し、1990年には連邦ネットワーク評議会が、インターネットへの加入団体となるには支援する政府機関が必要であるという制限を撤廃することで、誰でもコンピュータと通信回線さえあればインターネットに接続可能となったことで実現された(18)。

2－1．WWWの発想

1989年に、当時ジュネーブの CERN（欧州原子核研究機構）に在籍していたティム・バーナーズリー（Tim Berners-Lee）が、CERN 内の科学者同士での研究論文の交換を目的にした、「グローバル・ハイパーテキスト・プロジェクト」を提案している。CERN とは、ヨーロッパの12カ国が共同で設立した研究組織であり、巨大加速器を中心にした「研究都市」と言ってもよいくらいの大きな規模である(19)。そこでは、数千人にのぼる研究者や参加者間が研究

を行なっており、それらの間で、いかに情報を流通させ共有させるかが大きな問題であった。これは、ブッシュの置かれていた状況と同じであり、それがバーナーズリーの提案の背景にあるのは興味深い。

この提案は、後に WWW（World Wide Web）システムとして広く知られるようになり、Internet のキラーアプリケーションとも言われるようになっていくことになる。以降、1993年のマーク・アンドリューセンによる Web ブラウザ「Mosaic」の開発、ネットスケープ社による Netscape Navigator の発表、マイクロソフト社の Internet Explorer の発表などによるブラウザの高機能化と、Windows95 の登場によるパソコンの普及が、WWW の一般化を促していくことになる。1994年には、マサチューセッツ工科大学に着任したティム・バーナーズリーにより、WWW システムの標準化団体として、W3C（The World Wide Web Consortium）が創設されている。それにより、WWW システムは、一挙に標準化が加速し、今や経済活動や文化芸術活動にも欠くべからざる情報メディアとして、世界的に認知されているのは言うまでも無い。

Internet と WWW を巡る様々な事項の概要については、列挙の暇も無いほど多くの文献や Web ページで解説されているので、ここではこれ以上触れないが、ティム・バーナーズリー自身、テッド・ネルソンのハイパーテキスト（Xanadu）からの影響を明言しており、Xanadu プロジェクト側も以下のように、様々な影響を及ぼしていることを認めている。

「Project Xanadu was the explicit inspiration for the World Wide Web (see Tim Berners-Lee's original proposal for the World Wide Web), for Lotus Notes (as freely acknowledged by its creator, Ray Ozzie) and for HyperCard (acknowledged by its developer, Bill Atkinson); as well as less-well-known systems, including Microcosm and Hyperwave.」(15)

それらの事実は、Web データを記述する SGML 流のマークアップ言語を、HTML（Hyper Text Markup Language）と名付け、Web サーバとクライアント間でのデータの転送プロトコルを、HTTP（Hyper Text Transmission Protocol）としているところか

らも明らかである。何よりも WWW を生み出したプロジェクト自体、「グローバル・ハイパーテキスト・プロジェクト」と命名されていたという点にも注目される。

このように、WWW システムは一般には地球規模の Internet をベースとしたハイパーテキストと捉えられている。特にモザイク以降のブラウザは、画像から始まり、音声や動画などのいわゆるマルチメディアデータをも情報として扱うことが可能となってきているため、そのようなマルチメディアを包含したハイパーテキストを、特にハイパーメディアと呼ぶことがある。尚ハイパーメディアの用語は、後にテッド・ネルソンが自らの造語であることを[12]で認めている。

確かに実装技術の側面からは、WWW は Internet ベースのハイパーテキストシステムである。しかし WWW システムも、memex が問題提起しハイパーテキストが試みた、非線形的な人間の思考に対する、線形構造により整理された知識の適用という課題を含んでいるということに注意しなければならないだろう。

2-2. IP アドレスと階層

Internet は、その名前の通り、インターナショナルなネットワークであり、巷間言われているように、全体を統括、管理する組織が存在しない、中枢の無いシステムである。ユーザがアドホック的に接続をすることが可能であるため自由度が高く、パソコン通信などとは比較にならないほど急激にその規模を拡張していった。

そうした自由度の高い分散性を実現しているのが、プロトコル TCP/IP である。TCP/IP とは（Transmission Control Protocol/Internet Protocol）の略で、元来アメリカの国防総省（DOD）の主導で設立された DARPA（Defense Advanced Research Project Agency）による、国防ネットワークの研究に基づいている。1981 年頃から UNIX（4.1BSD）に実装され、デファクトスタンダードプロトコルとして、全世界で認知されている。

TCP/IP のうち、特に IP レイヤーでは、「IP アドレス」という概念により、ネットワーク上の全リソースを一意的に識別する。尚、この IP アドレスは ICANN（Internet Corporation for Assigned

Names and Numbers）という民間の非営利法人とその下部組織に当たる NIC（Network Information Center）という組織で一元管理されている。NIC はいわば事務調整機関であり、ネットワーク自体の運用を管理しているわけではない。

その詳細はここでは本旨から外れてしまうため省略するが、IP アドレスは32ビットのデータ列、ビットパターンである。要するに、2の32乗個の識別子がこれによって与えられている。

2の32乗 ＝ 4,294,967,296

つまり最大43億台のコンピュータをつなぐことができるが、Internet 上の全ての資源は等価な識別名を持っているため、Internet はフラットな構造を持ったネットワークであると言える。その意味においては、Internet 上の全資源は、元来構造化されていない乱雑な情報群であったと言っても過言ではない。

フラットな構造を持った Internet に構造を与え、現在のように非常に理解しやすいネットワークへ変えたのは、DNS（Domain Name System）の発明である。技術的には DNS こそが、地球規模のネットワークを効率的に運用しているいわば要であり、Internet に対してネットワークの外部性の極大化を実現した技術と評価できる。

DNS のアイディアは、南カリフォルニア大学情報科学研究所の P.モッカペトリスにより提起され、1983年にその技術を採用したサーバが Internet に登場する。1983年は、同様に TCP/IP がインターネットに正式に採用された年度でもあり、同年を「現在の大規模なインターネットを可能にする技術的な基盤が整った年」[20]とも評することがあるほど、DNS 技術は重要である。

DNS の機能を端的に言えば、ネットワークに接続したコンピュータ群から、効率的に特定のものを見つけ出すためのデータベース（あるいはデータベースの索引）である。TCP/IP プロトコルにおいては、IP アドレスに基づきホストを特定しなければならないが、それを効率的に行なうために、DNS では意味の無いビット列である IP アドレスにある基準を与え、階層化して整理している。意味情報を与えられた IP アドレスの表現形式を、ドメイン形式と呼ぶ。ドメイン形式は、一般的に馴染み深い Internet アドレスのことであ

る。

ドメイン形式は「.」記号をデリミタとして区切られて記述される。「com」や「jp」など最後尾のコードを TLD（Top Level Domain）という。TLD には gTLD(generic Top Level Domain)と、国別の ccTLD(country code Top Level Domain)がある。また TLD 以下、サブドメインと呼ばれる識別名が並んで記述される。特に、コンピュータの識別に使われるサブドメインは「ホスト名」と呼ばれることがある。

ここで注意したいのは、ドメイン形式は、実世界の住所表記と同様に、大分類である TLD から小分類であるホスト名まで、階層化されて整理されているということである。

例えば本学 DNS サーバ

dns－sv.ferris.ac.jp

は

202.248.252.130

である。同様に Web サーバである

www.ferris.ac.jp

は

210.131.76.20

である。(注5)

DNS は、こうしたドメイン名を IP アドレスに変換したり、その逆に IP アドレスからドメイン名を探す逆の機能を持っている。特に、現在の爆発的に拡大しさらに刻々と変化しつつある Internet 環境では、全世界の最新 IP アドレス情報を全ユーザが個々に管理するのは実質的に不可能である。そのため、こうした階層型のドメインを利用して全世界の IP アドレスは整理されており、IP アドレスとドメイン形式の対応を図るために、階層的なサーバ構造が利用されている。つまり、管理するアドレス情報の範囲とサーバが限定されているのである。特に全世界の DNS システムのトップに位置するサーバをルートサーバと呼び、全世界では13台が稼動している。いわば全世界の Internet は13台のコンピュータによって支えられているとも言えるかもしれない。

そうした階層構造を利用したアドレス情報を得るためのアルゴリズムは、以下のようになる。

① 全てのコンピュータは、ルートサーバの IP アドレス情報を持つ
② ルートサーバは次の階層のサブドメインを管理するサーバの IP アドレス情報を持つ
③ サブドメインを管理するサーバは、そのサブドメインを管理するサーバの IP アドレス情報を持つ

以下、②から③が再帰的に繰り返されて、目的のコンピュータのアドレスを得ることが可能となる。

以上述べたように、純プロトコル的にはフラットなネットワークである Internet は、DNS により論理的に階層構造化され、さらに物理的にも階層的に情報が分散されて整理されているということがわかる。その意味からすれば、Internet は線形構造によって整理されたデータの集合体である。極論をすれば、Internet も図書館や紙媒体などと同じように、線形構造により整理された、いわばオールドメディア的な構造を持った媒体なのである。

線形構造による知識の整理には、多くの利点があるのは言うまでもない。DNS を中心とした一連の技術が生まれていないとするならば、我々はフラットな構造を持ったネットワークの中で必要とする情報の糸口さえ見つけることすら困難な状況にあったのは間違いが無い。しかし、我々の思考自体は、DNS のように階層的なものでも無いのである。

そうした観点から見ると、本来 WWW システムは、線形構造により整理された知識を用いて、我々の非線形構造的な思考作業への支援を実現する機能を持つべきものであり、ネットワーク環境とディジタル技術を前提とした memex なのだ、というのは言いすぎであろうか。だとするならば、我々は WWW を利用することで、知識の再利用を図ることができねばならない。しかし前述のように、大量の WWW サーバが Internet に接続している現在は、WWW のハイパーリンクにより自由に情報の再構成や整理ができたとしても、決して思考の支援には成り得ない。ブッシュが問題意識を持っ

ていた情報の洪水といった問題が、現在の WWW にはつきまとうのである。

そこで以降には、拡張を続けている WWW サーバを整理し、それらを適切なユーザに対して適切な情報として提供する、いわば WWW を現代の memex とするために必須なサービスである WWW の検索技術に関して述べる。

3. サーチエンジンの現状と問題提起

現在 Internet 上では、商用、実験を含め多くの数の WWW の検索サービスが運用されている。一般的な Internet の概説では、WWW システムの検索サービスは、大きく 2 種類に分類される。その一つはネットディレクトリと呼ばれる分類表を用いたシステムであり、もう一つは通称サーチエンジンと呼ばれるシステムである。

ネットディレクトリとは、各ページを特定の分類法に基づき階層構造で整理したもので、yahoo に代表される。WWW サーバの階層構造に対して、さらにコンテンツ内容に基づいた階層構造を与えるものである。初期の段階では、こうしたカテゴリーリストに基づいた分類を用いたものが主流であった。

後者のサーチエンジンとは、一般に WWW ページの検索を行うシステムを言い、本稿で考察の対象とする。サーチエンジンという用語自体比較的緩やかな意味合いで使われており、ほかにも検索エンジン、サーチツール、サーチサービスなど様々な語が用いられている。基本的には、ユーザからの要求を受けて、その要求に合致する WWW ページのアドレスを返すサービスを言う。

特にロボットと呼ばれる WWW コンテンツを自動収拾するプログラムモジュールを含んだサーチエンジンは、1994年頃から出現しており、「Infoseek http://infoseek.go.com/」「Lycos http://www.lycos.com/」「WebCrowler http://web.webcrawler.com/」などが当時の代表的なサービスであった。これらのサービスは、以降 WWW ページの急増に合わせて利用者も増大し、企業の統合などを経ながらも現在でもサービスが継続している。

(21)などで、サーチエンジンの機能比較が行なわれているが、そこでも述べられているように、初期のサーチエンジンでは従来の情

報検索分野において研究されている検索技術がシステムの中心になっている。また傾向としては、「見出し」を対象とした検索から「全文検索」へと技術が移ってきているのも特徴的である。1997年頃からサーチエンジン同士の競合が激しくなり、各サーチエンジンはニュースやフリーメールなど様々な付加サービスを備えていわゆるポータルサイトとしてのトータルな機能を備えるようになって行った。その傾向は、1999年の「google」の登場まで続くことになる。その間、自然言語処理「Infoseek」、拡張的な検索式「AltaVista」、検索結果の自動分類「NorthernLight」、最新情報の収集「FreshEye」など、様々な新規技術が導入されてきたが、ポータルサイトとしての様々な付加サービスに埋没してしまったことは否定できない。

この時期のサーチエンジンを第一世代とするならば、明らかに「google http://www.google.com/」の登場は革新的であった。1999年9月にサービスを開始した「google」は、ページデザインがサーチのみであることからも明らかなように、検索に特化したものであったが、第一世代サーチエンジンとは技術的にも大きな違いを持っている。それは端的に言えば、WWW の特性と情報検索手法を関連付けた、いわば WWW に特化した固有のアルゴリズムを備えているという点にある。

google は、Larry PageとSergey Brin が米スタンフォード大学大学院在籍中に開発した検索エンジンに基づいたサービスである。元々スタンフォード大学コンピュータサイエンス学科の研究プロジェクトとしてスタートし、ベンチャーキャピタルから総額2,500万ドルの投資を受けて、1998年9月に起業している。2001年3月末時点で8,000台の PC を使い、それらを用いた計算機クラスタによる分散システムにより、7,000万件／日の検索を1件あたりほぼ0.5秒で処理している(22)。また、Netscape Communications は Google がまだ試験段階だった頃から提携しており、Yahoo も2000年6月にサーチエンジンサービス（デフォルト・サーチエンジン）を Inktomi から Google に切り替えている(23)。

前述のように、Google は、ポータルサイト指向を廃したシンプルな画面構成やキャッシュシステム、計算機クラスタ処理など様々な点にその特徴が見られるが、検索結果の的確さがこれほどまでに

注目を集めている理由であり、それを実現する技術が、「PageRank」アルゴリズムと呼ばれている google 固有のページ重要度への自動判定技術である。「PageRank」は、(24)などで詳細が明らかにされている。また(25)では、google 側の公的な見解としてこの「PageRank」について解説されている。

これらによれば、注目すべきは、単純な語句の出現頻度ではなく、ハイパーテキスト構造を取っている WWW の特性に着目しているという点である。「PageRank」では、「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」[23]という、一般ユーザの暗黙の期待を、数値モデルとして明確化している。この「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」という関係は、再帰構造を取っていることに着目し、そのリンク構造を計算機上で数値化して適切な検索結果が算出される。すなわち、WWW のリンク構造をグラフ理論の応用として捉え、線形代数の数値解析的手法で実現されているのである。言ってしまえば、一般的な数学モデルの応用でしか過ぎないのではあるが、google が戻す検索結果は、確かに第一世代サーチエンジンと比べると、遥かに適合性が高い。

工学的な研究領域では、その技術がいかに優れた結果を出すか否かが、評価のポイントである。理論的に勝っていたとしても、それによって作られた工学的な産物（すなわち製品）が、なんらかの有効性を持たねば、その理論は工学上は劣ったものとみなされてしまう。乱暴な言い方をすれば、工学の世界に絶対的な真理は無いのである。そういうある意味ドラスティックな領域に、純理論的な数学モデルを提示し、その有効性を証明したところに、google の価値があると評価する。

3-1．第二世代サーチエンジンの限界

このように google の登場と成功により、ハイパーテキスト構造のリンクを利用した、評価手法が有効であることは明らかになってきた。しかし逆に、まだまだ問題点や考察すべき点も明らかになってきている。それらに対する問題意識が、本研究のもともとの出発点でもある。

元来ハイパーテキストシステムや WWW は、全て思考の支援システムというところから出発している。特にパソコンが情報環境の主流を占める現在においては、コンピュータを使った情報活動は、極めて個的な色彩を強く持つようになってきている。一人で CPU を独占するパーソナルコンピュータは、操作者を個別化する。その意味においては、コンピュータの操作は個々の知的作業とより緊密に結びつかざるを得ないと言える。前述のように、WWW は階層的に構造化された Intetrnet 環境に、非線形性を実現する可能性を持ったものであるが、情報量の余りの多さから、サーチエンジンによりそうした人間の思考活動を支援すべく期待されている。しかし第二世代までのサーチエンジンは、情報の発信者による WWW ページの非線形性（すなわちリンク）にのみ着目し、情報を受信する側の思考については全く考慮していない。サーチエンジンを使う目的性に関してはまったく考慮されていないのである。端的に言えば、無目的にネットサーフするようなユーザも、研究論文を書こうとしているユーザも、就職活動で情報を探しているユーザも、Internet 通販を利用しようとしているユーザも、全て等しく扱われてしまう。要するに検索要求にもユーザの個別性があり、その個別性は何らかの形でサーチエンジンが吸収していく必要があると思われるのである。

それを示す、端的な例をここで問題提起として挙げる。これは筆者が実際にある企業の人間と仕事を行なっていて直面した事例である。

前述のように、WWW はいわゆる eビジネスのベースとして、欠かせないものとなってきている。また商業活動で使われる頻度が高まってくるにつれ、特に安価な広告宣伝の手段としても多く使われてきている。特にバナー広告（幟）と呼ばれる小さなシンボルをコンテンツの中に埋め込み、それによってユーザを集め顧客を開発する手法が、いわゆるプッシュ型マーケティングの一環として行なわれている。

例えばＡという商品開発を行なった企業が、その商品を Internet 経由でプロモーションすべくバナー広告を扱っている広告代理店に依頼をした場合、広告代理店はサーチエンジンを使って、そのバナー

を掲載するに適したWebページを発見して広告の掲載依頼を行なっている(注6)。しかし、その問題に関しては、語句の一致とリンクのランキングでは、必要な検索はほぼ不可能である。卑近な例ではあるが、商品Aが例えば野球関連のノベルティ商品だとする。その場合検索キーワードとしては、当然第一候補として「野球」の語句が使われるはずである。しかし実際のWebデータの統計を取ると、奇妙な現象に気がつく。表1に示すのは、日本のプロ野球チーム（セ・リーグのみ）の公式サイトのトップページに含まれている「野球」の語句の数である（2002年2月のデータ）。

Http://giants.yomiuri.co.jp/	読売ジャイアンツ	2
Http://www.baystars.net/	横浜ベイスターズ	0
Http://www.yakult-swallows.co.jp/	ヤクルトスワローズ	0
Http://www.carp.co.jp/	広島東洋カープ	0
Http://www.hanshin.co.jp/tigers/	阪神タイガース	0
Http://www.dragons.co.jp/	中日ドラゴンズ	0

表1　検索データ

ここで明らかなように、野球に関する情報を保持している各チームのコンテンツには、「野球」という語句が殆ど含まれてはいない。これは、情報を提供する側が、「各チーム名はプロ野球チームを示す名前である」という、いわゆる暗黙的な前提で情報を構成しているからである。若干トリビアルな例ではあるが、こうしたレベルのサーチは現在の技術では不可能である。これを実現するには、サーチ側に情報の発信側の暗黙的な前提、すなわちある種の知識を備えなければならない。

他にも、利用側から期待される様々な柔軟性を持ったサーチパターンや機能がある。そこで、第三世代サーチエンジンとして、知識処理を包含したサーチエンジンを提案する。

こうした知識に対する研究や問題意識は、人工知能として、長い間研究がなされてきており、その成果は情報科学やソフトウェア工学の分野に適宜吸収され、情報技術の発達を根底から支えている。第二世代サーチエンジンが、純数理的な手法で大きな成果を上げたのと同様に、サーチ技術に知識処理手法を導入することで、さらに

高い成果を上げることが可能であると確信するものである。

この場合、知識情報としては、①情報発信側の知識と②検索側の知識が含まれる。前者は、特定の情報の背後に有る暗黙知や領域知識（ドメイン知識）などが含まれ、後者には特定のサーチをする意図などタスク知識などが含まれると考えられるだろう。ここで言うタスクとは問題解決作業のことであり、タスク知識とは問題解決のための方法や構成といった知識、ドメイン知識とはその問題解決を適用させる対象に関する知識である。一般に専門知識は目的と強く関連するタスク(task)知識と、利用の方法からは独立した比較的客観的なドメイン(domain)知識とから構成される[26]。

ここでの仮説としては、発信側はその Web ページがどうサーチされ、どう利用されるかとは独立の、いわばドメイン知識によりページを構成しているが、受信側、サーチ側はタスク知識に従って問題解決を図っており、サーチエンジンは両者の知識が交差する場であると考えるものである。

3-2. サーチエンジンの要素技術と方向性

サーチエンジンは、ユーザサイドから見るとかなり複雑なシステムのように見えるが、技術的にはそれほど高度なものではなく、また難しいものでもない。前述のように、現在の商用サーチエンジンの多くが、大学において研究開発されたものがベースとなっており、また必要な要素モジュールやデータセットなども、その多くがソースを含めオープン化されている。

一般的なサーチエンジンの構成モジュールは、次のような論理構造を取る。

1　検索対象の収集
2　蓄積と索引付け
3　検索インタフェース

以下にこれらに関する技術の概観と、問題意識に関して述べる。

○　検索対象の収集

1は、検索対象となる HTML ファイル自身を特定する作業モジュールである。サーチエンジンでは、ロボット、あるいはスパイダー、

クローラーなどと呼ばれる自動巡回プログラムによって、HTMLファイルが特定され収集される。元来ロボットプログラムは、ネットワークのステータスやロギングなどを収集するものとして用いられてきた、いわゆるワームプログラム（蚯蚓）の一種であるが、サーチエンジンのロボットはこれを自動収集に転用したものである。
ロボットのアルゴリズムは、以下のようになる。

① 特定のHTMLデータに含まれるリンク（<A>）を発見する
② アンカー先のURLを既知URLリストと比較し未知のURLを発見する
③ 未知のURLに対してHTTP通信を行ないHTMLデータを取得する
④ 既知URLリストに加える

HTTPにはクライアント側からファイルの一覧を得るプロトコルが含まれてはいない[27]。そのため、収集対象のURLの特定が、サーチエンジンの結果を大きく左右することになる。（注7）

現在多くのロボットプログラムが使われているが、それらがパスワードやクレジットカード情報、機密情報、などを含めた重要なデータを収集してしまうケースが増えている。実際googleでは、「Adobe PostScript（PDF）」、「Lotus 1-2-3（123）」、「Excel（XLS）」、「PowerPoint（PPT）」、「Word（DOC）」「リッチテキスト（RTF）」などのファイルを限定してサーチすることができるため、名簿の流出が頻繁に起こっているのも否定できない。また、ロボットのクロウリングは、WWWサーバに負荷を掛けてしまうため、メタタグ（<META>）を用いたサーチの排除や、ロボット排除規約（Robot Exclusion Protocol）に従ったロボットの設計などが求められている。

このように、ロボットの巡回がサーチエンジンが対象とするWebデータを決定付けるため、重要な論理モジュールである。サーバ内の全ページを最初に巡回する縦型探索（Depth-first）と、リンクを重視する横型探索（Breath-first）が考えられるが、どのサーチエンジンでも、探索経路や優先度などは考慮されていない。今後、より効率的な探索手法が必要となるであろう。

○　データの蓄積と索引付け

収集された URL 情報は、索引付けされ、さらにデータベースに蓄積されて検索の対象となる。そのデータベースの規模、特にインデックスのサイズ自体が、特に関連した検索結果を選ぶ上では最も影響力を持つとの(28)の指摘がある。 google の公式アナウンスでは、そのインデックスサイズは16億以上とのことである。しかし、URL の移動による内容の重複や、変更、削除されたページも少なからず含まれており、有効なインデックスの実数には誤差があると思われる。

最近の傾向としては、Web ページの履歴を扱うアーカイブが現れてきていることが上げられる。Google ではキャッシュ機能により、削除された Web ページを google のサーバから探し出すことができる。また2001年10月から、「Wayback Machine」で、1996年以降ではあるが、特定のサイトを過去にさかのぼって保管しているアーカイブサービスも開始してきている。「Wayback Machine」では100億を超える Web ページを保管しているとのことである。(29)

当初 google のキャッシュ機能が著作権を侵しているとの議論があったが、今後もこうしたサービスは WWW が社会、経済において重要なメディアになって行くに連れ、充実したものとなっていくことが予想される。こうした状況を前提とすると、サーチエンジン側で固有のデータを持つことがそれほど重要な機能では無くなっていくと思われる。メタサーチ的に、他のアーカイブを利用するサーチエンジンといった方向性も考えられる。

○　検索インタフェース

3 の検索インタフェースは、クライアントの入力に従い、蓄積されたデータベースを対象にして検索結果を返すモジュールである。前述のように、現在ではいかに大量の結果を出すかではなく、いかに適切な結果を出すかに力点が移ってきている。そのため、前述のように、google では「PageRank」アルゴリズムによりランク付けを行ない、検索結果を導き出している。他にもコンテンツのユーザへの適合性を判断する判断基準がいくつか明らかにされている。

「PageRank」は引用されている度合いを重要な要素としたもの

であるが、他にも検索語の出現頻度、ファイルのバージョン（新鮮さ)、検索や参照の頻度（人気)、情報の属性（ページのタイプ)、情報元の信頼度などが考えられる。しかし WWW の利用が多様化をしてきている現在においては、これらの要素のうちのどれもが決定的なものではなく、総合的に判断する必要があるように思われる。その意味から「検索語は主に問題解決ドメインを絞り込み、ページタイプは主に問題解決タスクを絞り込む」という(28)の指摘は興味深く、前述した「情報発信側の知識と検索側の知識の交差」の問題と併せて考察したいと考えている。

4. まとめ

2章で述べた現在の第二世代サーチエンジンの欠点、あるいは問題意識をもとに、現在第三世代の知的サーチエンジン（memSearch）の仕様をまとめながら、各サブモジュールに関して試行錯誤を行なっている。その仕様や検討結果の詳細に関しては、稿を改めて報告する。

memex から始まった人間の思考への支援の試みは、ハイパーテキスト、そして WWW へと大きく花開くことになった。直感的な情報検索が可能、ナビゲーション経路が多様、情報の拡張が容易といったハイパーテキストの利点は(30)、まさに WWW と Internet が拡張していった理由でもあろう。しかし皮肉なことに、それらの利点が新たに我々に情報エントロピーの増大という、問題を突き付けたのも事実である。

その意味からはサーチエンジンは、まさしく Internet のインフラであり、WWW により提供される情報が増大すればするほど、そして WWW の実務的な役割が重要になればなるほど、その機能もそれに対応して進化しなければならないであろう。ことに現在日本の情報産業はコンピュータのインフラたる OS に関しては固有な技術を実質的に失っており、技術的にアメリカの後塵を拝しているのは否定できない。サーチエンジンのこうした Internet 環境における重要性を考えると、わが国固有の言語、文化環境のもとにある技術が存在する意義は大きいと思われる。

本稿を書くにあたり Internet の向こう側にいる研究者や企業のエ

ンジニア諸氏、および研究の機会を与えてくれているフェリス女学院大学に感謝します。また、株式会社ウェッブアイの森川、戸本両氏からは多大な示唆を頂きました。併せて感謝します。

注記

（1）言うまでも無く Internet におけるアプリケーションは、ファイル転送、ファイル共有、資源共有など様々なものがあり、本来ネットワークとその応用を考えるには、それらを含めて考察せねばならない。

（2）それがいつの社会なのかに関しては、さまざまな議論があり、参考文献（1）にまとめる。

（3）参考文献（6）ではライプニッツの考案による二進法もそうした例として指摘されている。

（4）1939年に、スタンフォード大学の教授であったフレディリック・ターマンの支援によって、ウイリアム・ヒューレットとデビッド・パッカードが起業して作られた会社が、ヒューレットパッカード社である。同社は、シリコンバレー初のハイテク企業として知られている。ターマンは、HP のインキュベーター兼インベスターということになる。ちなみに文献（9）によれば、投資額は538ドルであったそうである。

（5）IPアドレスが連続していないのは、現時点（2002年 2 月）で WWW サーバをアウトソーシングしているためである。

（6）代表的なバナー広告システムには、パワークリックなどがある。
http://www.pclick.net/company/index.html

（7）index.html ファイルがサーバ上に存在しない場合に表示されるファイルリストは、Web サーバ側によって送られるものであり、クライアント側からの要求があるためではない。

参考文献

（1）情報技術とグローバリゼーション，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，No2，2000
（2）Data Smog, David Shenk, HarperEdge, 1998
（3）米国・ヨーロッパの旅から，公文俊平，情報文明論研究会レジメ(2000/10/20)，2000
（4）見えない洪水－ケースD，糸川英夫，ソニー出版，1979
（5）科学論入門，佐々木力，岩波新書，1997
（6）300年後に実現した思索－デジタル技術の驚異，黒崎政男，朝日新聞　文化欄(2000/11/9)，2000
（7）Events in the Life of Vannevar Bush,http://www.cs.brown.edu/research/graphics/html/info/timeline.html
（8）ハイパーメディア・ギャラクシー，浜野保樹，福武書店，1988
（9）シリコン・ヴァレー物語－受けつがれる起業家精神，枝川公一，中央公論新社，1999
(10) National Science Foundation (NSF) - Home Page, http://www.nsf.gov/
(11) As We May Think, http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm
(12) リテラリーマシン－ハイパーテキスト原論，テッド・ネルソン，/竹内郁雄/斉藤康己監訳，ハイテクノロジー・コミュニケーションズ訳，アスキー出版局，1994
(13) アラン・ケイ，アラン・ケイ，浜野保樹監訳，鶴岡雄二訳，アスキー出版局，1992
(14) Dynabook Revisited with Alan Kay, B.Ryan, Byte. vol 16, February, 1991
(15) Project Xanadu History(lo-res), http://xanadu.com/xuhistory.html
(16) Proceedings of Online 72 Conference,T.Nelson, Brunel University, Uxbridge,1972
(17) A File Structure for the Complex, the Changing and the indeterminate, T.Nelson, Proceedings of the ACM National Conference,1965
(18) インターネットの光と影，情報教育学研究会(IEC)・情報倫理教育研究グループ編著，北大路書房，2000
(19) CERN in 2 minutes,http://public.web.cern.ch/Public/whatiscern.html
(20) ポール・モッカペトリス，http://www.chienowa.co.jp/frame1/ijinden2/Paul_Mockapetris.html
(21) 情報検索システムとしてみたサーチエンジン，久野高志/安形輝/上田修一，第49回日本図書館情報学会研究大会発表要綱，2001
(22) サーチエンジン Google，山名早人/近藤秀和，情報処理学会誌，Vol. 42 No.08，2001

(23) Google の秘密,馬場 肇,http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html
(24) Efficient Computation of PageRank,T. Haveliwala,Technical Report, 1999
(25) Google の人気の秘密, http://www.google.com/intl/ja/why_use.html
(26) オントロジー工学序説,溝口理一郎/池田満,人工知能学会誌,Vol. 12 No.4,1997
(27) WWW サーチエンジンの作り方,原田昌紀,情報処理学会誌,Vol. 41 No.11,2000
(28) WWW 情報検索技術と評価の問題,福島俊一, 情報処理学会誌,Vol. 41 No.08,2000
(29) The Internet Archive: Building an 'Internet Library, http://www.archive.org/index.html
(30) 電子出版－紙の本から電子の本へ,斉藤孝,日本経済新聞社,1993