情報検索の動向

Trends in Information Searching

東京農工大学府中図書館　**岡谷　大**
YUTAKA OKAYA
Libary of Tokyo University of Agriculture & Technology

1. はじめに

周知のごとく2000年ついにアメリカのセレーラ社と日米欧の国際共同チームによって、ヒトゲノム（人間の全遺伝情報）の全解読がなされた。しかしゲノムのどの部分が意味を持ち、機能を持つかかが解明されたわけではなかった。そこで情報技術（IT）を駆使してゲノムから重要な遺伝子を探索するバイオインフォマテックス（生命情報工学）や、病気の起こる仕組みをDNAレベルで調べて新薬開発を目指すゲノム創薬、プロテオームと呼ばれるタンパク質研究などいわゆる「ポストゲノム」の研究が新たにスタートすることになった。これらの成果において情報やコンピュータの果たす役割が大きかったことは否定できない事実でもあった。

ところでこうした生物情報工学への見方として大きくは生物系と情報系があり見解の相違もあるようである。このことはバイオに限らず材料科学、高分子化学、特許・知財などに及ぶものと思われる。本稿では最新の情報検索を紹介し情報学の立場からいくつかのべたい。

まず最近の傾向つまり通信系（ネットワーク）と制御系（情報の表現や操作など）の現状にふれる。通信系ではインターネット（WWW）という巨大でヴァーチャルな情報空間を示す。さらに引用（citation）というネットワークを示す。後者の制御系では電子ジャーナルの出現などの情報環境の変化、全文検索、CASのScifinderなどの情報検索の使い勝手や情報の可視化などの情報の表現技術のいくつかを示す。

しかしこうした膨大な情報を本来の研究という面からみると、データベースが巨大になればなるほどノイズつまり研究にとって不適切な検索結果が増大してくることにもなる。同時に一方で研究における発想や創造性ということがより深い次元で問われてくる。つまり質のよい検索とはなにか、研究や発想と結びつく検索とはなにか、そのためにはどうすればよいのかといった情報検索の本質が鋭く問われてくる。そのひとつの解決の側面としてはコトバや分類の問題、つまりターミノロジー（概念・用語学）とオントロジー（存在論）があると思われる。ターミノロジーとオントロジーは密接に関係するが、それぞれにつきバイオインフォマテックス、高分子化学、電子化辞書その他いくつかの事例で検討したい。その議論をふまえてその先の論点として情報の面からの研究の評価、さらにコンピュータによる発想支援や創造性の可能性をとりあげたい。これらのトピックスが研究における情報検索利用上でのなにかの参考になれば幸いである。

2. 情報検索の最近の傾向

2.1　ネットワークと情報空間の拡大

最近WWW（インターネット網）とHTMLや、XML（ネットワークの言語）によって大規模でヴァーチャルな情報空間が構築されたことが注目される[1)]。またこれにはロボット型のgoogleと階層型のyahooなどの検索エンジンの発達がある。図1はサーバーやwebロボットによるロボット型サーチエンジンの仕組が、図2はハイパーリンクによるwebロボットの動作が示されている[2)]。

データベースとその活用についてふりかえると、データベースとはもともと1950年代にアメリカで生まれ、データ（情報の最小単位）とベース（軍事的な補給基地）の合成

語で、データの構造化を背景とする技術であった。さらに大容量のCD（compact disc）技術、推論エンジンと知識データで構成され、付加価値を生み出す知識ベース、その知識ベースの大型化である大規模電子化辞書やゲノムデータベースなどにみられる大規模知識ベース、そして最近のインターネットの普及とそこからの知識の取り出しと活用（知識マイニング）といった一連の流れがある。例えば材料関係、高分子化学、NLM（アメリカ国立医学図書館）の医学情報であるMedline、とりわけ文献情報の全文検索が可能となったPubmed、先述の国際的に進められたヒトゲノム・データベースなどがある。

一方でガーフィールド（Garfield,E）を嚆矢とする引用（citation）というネットワークによるデータベースがある[3]。最近「Web of science」というネットが構築された。これにより文献、著者、キーワード間の引用の頻度による引用-被引用関係の地図作製、つまりマッピング（可視化）が可能となった。またこれにより研究の未来予測もできるという。図3は筆者による引用分析の結果を示している。4人の農芸化学や獣医学などの研究者を対象に引用文献を調査し、2 step mapにより中核となる雑誌（例えばCではJ.EndocrinologyやEndocrinology）を見いだせる[4]。

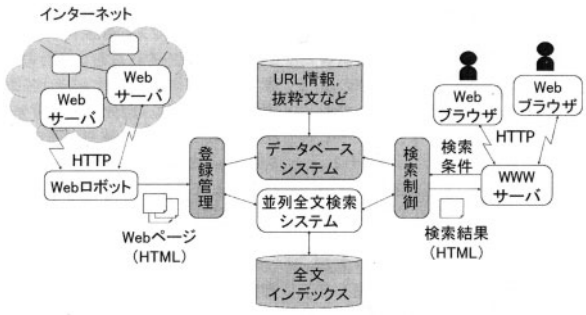

図1　ロボット型サーチエンジンの構成（出典　参考文献2,p369）

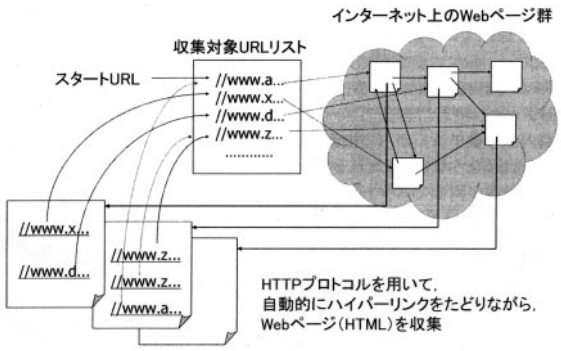

図2　ロボット型サーチエンジンの動作（出典　参考文献2,p369）

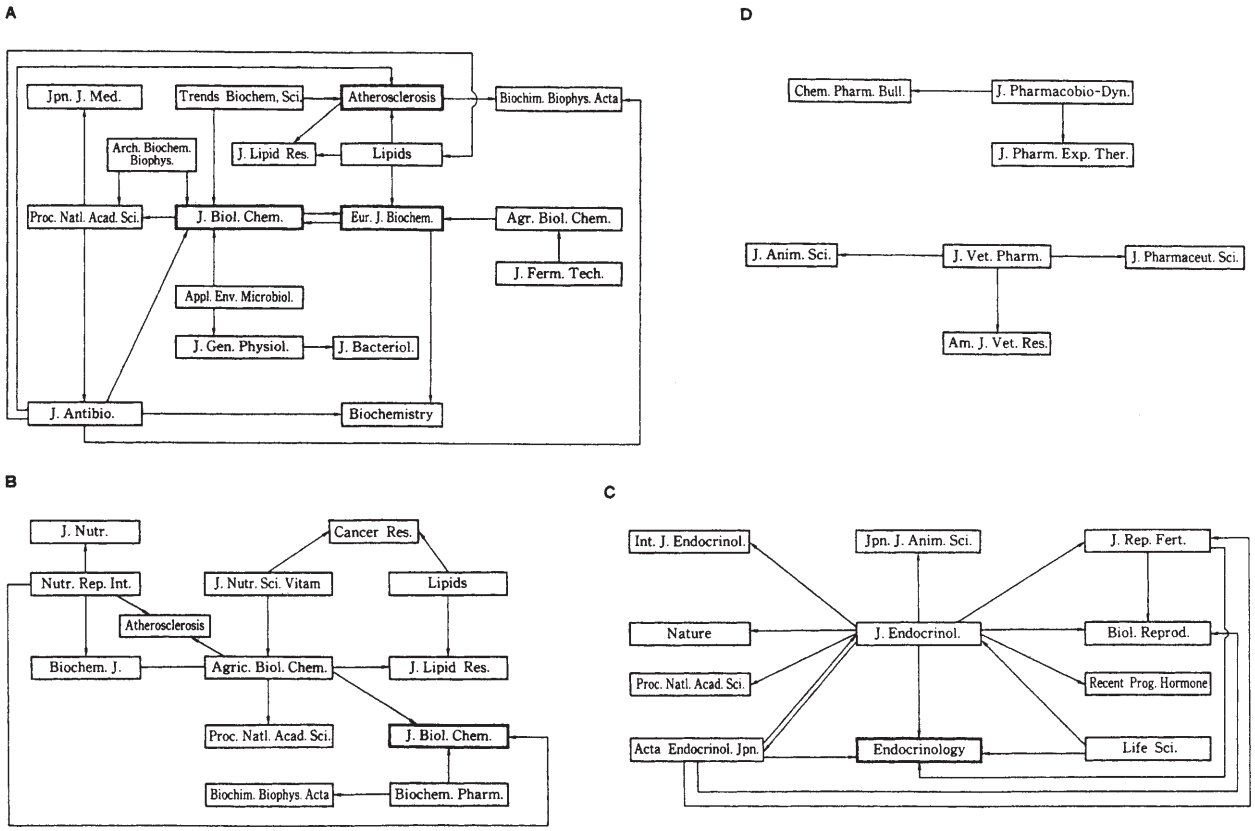

図3　引用データベース（出典　参考文献3,p122）

ただし実際には引用の本質的な見方や、引用結果の雑誌の規模などによる補正の問題、つまりインパクト･ファクター（impact factor）などををどう考えるかといった問題がある。こうしたファクターには実際はArticles（論文総数）、Immediacy Index（最新文献指数）、Cited Half-Life（被引用半減期）、Citing Half-Life（引用半減期）、Total Cites（被引用総数）などいくつかの種類がある。さらに4.1でも触れるがこうした引用の量的結果を安易に研究や研究者の質的な業績評価と結びつけてはならないといったことがある。

しかしインターネットにして引用ネットにしても、巨大な情報空間となればなるほど検索上のノイズが大きくなってくることは避けられない。結局正確を期すには検索結果の見直しをしなければならないことになる。また入力段階で前方一致や後方一致などのトランケーションなどの技術はあるものの基本的にはどうしてもコトバ（検索語）が検索の重要な鍵になってくる。このコトバや用語の学問であるターミノロジー学については3.1でのべるが、最近ネットにコトバの意味関係をもたせた「セマンテックweb」の研究が盛んである[5)]。この基礎にあるのがメタデータ（著者、書名など）や次項でのべるオントロジー（存在論）である。つまりセマンテックwebはメタデータの共通化などにより、情報の定義を明確にし、そのことによって膨大なweb空間（コンテンツ）を知的な知識ベースとし、多面的な検索を可能にしているのである。セマンテックwebの技術により今後コンピュータ間およびコンピュータと人間とのさらなる協力関係が可能になると期待されている。

2.2　情報検索の環境変化と制御の進化

情報検索の環境の変化として、紙という従来の表現媒体に替えてコンピュータの画面でコンテンツをみる電子ジャーナルが出現し、Adobe社のAcrobat Readerで読むpdf（Portable Data Format）形式やコンピュータのメモリーの増大により、部分検索から全文検索が主要となってきた。いまや全文検索による電子ジャーナルからさらに電子ブックへと進化しているという。また言語と画像、動画などのミックスによるマルチメディアなどデジタルという点でさまざまなメディアがボーダレスに融合してくることとなった。この例としていわゆる情報高分子関連情報としてDNA塩基配列情報、mRNA塩基配列情報、タンパク質関連情報（アミノ酸配列、モチーフ検索、タンパク質高次構造データベース、プロテオーム解析）などがある。図4はNCBI（National Center for Biotechnology Laboratory）から検索したタンパク質の三次元（立体）表現で、この図が動的に回転していろいろなゲノム情報の側面がみえてくる。図5は遺伝子、タンパク質関連情報の関係図でDNAのGenBankやヨーロッパのEMBL（European Molecular Biology Laboratory）などのリンク関係が表現されている。[6)]

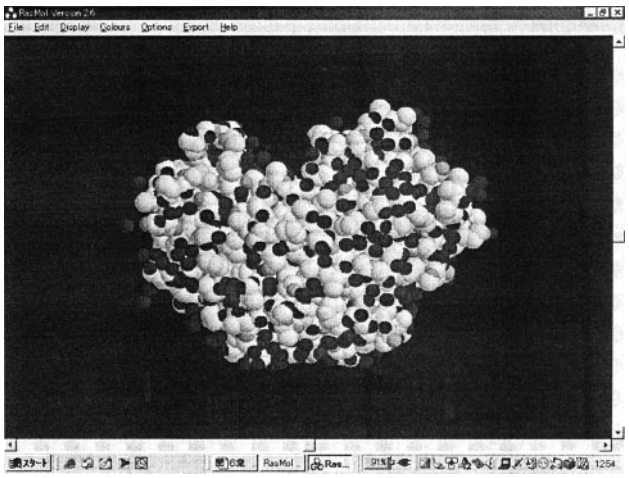

図4　RasMolのタンパク質表示画面（出典　参考文献6, p168）

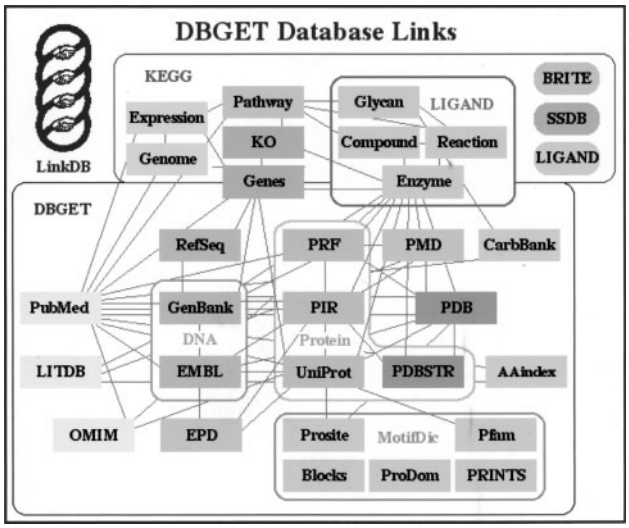

図5　遺伝子、タンパク質関連情報（出典　参考文献6,p171）

また1でのべたインターネットによる情報空間の拡大の例として 最近ではとくに先述の化学情報（CAS）のScifinderが挙げられる。この画面操作によって使い勝手がずいぶんと向上した。例えばこれまで面倒であった分子の構造検索などが簡単に画面入力で検索できるようになった。

次には計算ソフトとコンテンツ（マルチメディア）との結合による情報の「可視化」の技術が注目される。図6は特許情報におけるエイズワクチンの研究者別の時間変化による特許の傾向の三次元可視図である[7]。この図からいろいろなことが読みとれる。例えばヴィジュアルに研究の経過がみれるほか、時系列的に数年先の研究の予測も可能となる。

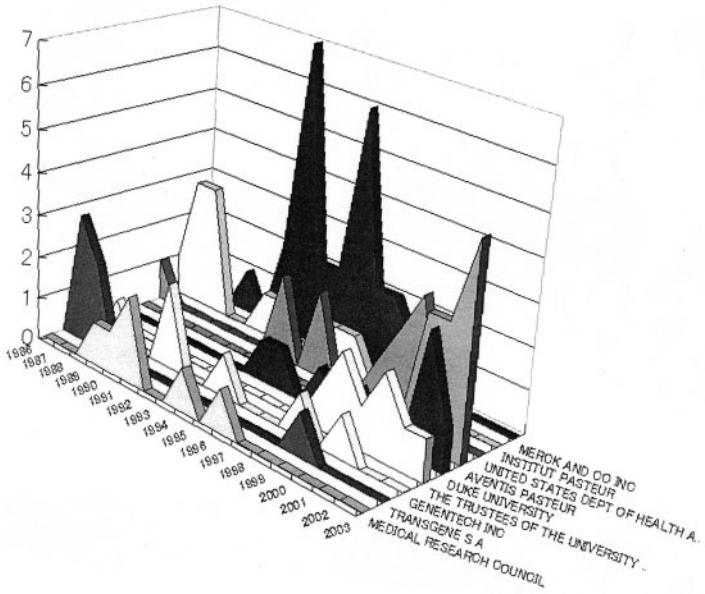

図6　特許情報の三次元表現（出典　参考文献7,p57）

情報の可視化の例としては、このほか一定の数式を用いた検索結果の適合順のランキングや、クラスタリング（層別分類）などがある。クラスタリングの例では先述のgoogleを超える新たな検索ツールとして、検索結果をすばやく分類し、各カテゴリーをクリックして検索結果を絞り込める検索エンジン「Vivisimo」が注目されている。さらにフィルタリング（特定の情報のみを通す）、連想検索（国立情報研究所のGETA）などがある。GETAは図書情報ナビゲーターに用いられている汎用連想検索エンジンである。この技術では一千万件規模の大規模文書データベースを対象に、文書間や単語間の関連性を高精度かつ高速に計算することが可能である。しかしこれらの技術の根底にあるのは次に述べるしっかりとした分類やオントロジーの構築である。

3. ターミノロジーとオントロジー

これまでの論述からわかることは結局検索における最重要の問題はコトバ（自然言語と統制言語）と分類である。どんなに検索技術、コンピュータ技術が進歩してもどのみちコトバで入力し検索しなくてはならないことには変わりはないのである。また現在の段階ではコトバ（概念）相互の意味関係をコンピュータに教えてやる必要があることはいうまでもない。

3.1　検索語とターミノロジー

まずコトバに関しては上述の自然言語（フリーワード）か統制言語（コントロールワード、例えばシソーラスの用語「ディスクリプター」など）かの問題がある。自然言語は操作が簡単であるがノイズが多い、反対に統制言語は操作が面倒だが適合的な検索、とくに専門的な検索に適しているなどtrade-offの関係にある。統制言語に関しては例えば同義語、類義語、略語、新語、外来語、合成語などが問題となる。2.1でのべたデータベースという用語もデータとベースの合成語でかつその当時では新語でもあった。後述のJST（科学技術振興事業団）のPoLyInfoというシステムでは、高分子をその構成単位に基づいて認識･同定するための化合物レジストリシステムである高分子辞書を作成･運用している。具体的にはポリマーの構造情報を構成単位化学構造に基づいた「高分子辞書書式」の形でもっている。またそのサブシステムの一つとして、IUPAC（国際純正応用化学連合）高分子命名法に準拠した高分子の名称を自動発生させる機能をもたせている。またバイオインフォマテックスにおける遺伝子の命名などは国や分野によって異なるといわれておりこの分野で最近研究が立ち上がっているという。ところでこういった用語や概念に関する学問が筆者の研究しているターミノロジー学である[8]。オーストリア人オイゲン・ビュスター（Eugen Wuester,1898-1977）によって創始され論理学、存在論、言語学、情報学などを背景の学問としている。実際には良質の用語集作成や用語のデータバンク（ターミノロジーデータバンク）構築が中心であるが

多くは国家的、さらにはISOやUnescoと関係した国際的な事業となっている。しかしその基本は次にのべるように概念（concept）であり、用語の表記の前に概念分析がなされなければならない。

3.2 分類

コトバ（用語）の根底にあるのが分類、つまり対象や動作などの関係（オントロジー、存在論）のことである。個物（対象）は認識論で扱う人間の認識とは独立である。個物には個物の認識とは無関係なそれ自身の関係、変化がある。オントロジーではこの関係や変化として例えば個物の全体･部分関係、部分相互の関係、時間的継承などを特に注目してその関係を明確にしている。こうしたオントロジーとはもともと哲学の一分野だが最近これと情報工学が結びついて「オントロジー工学」なる分野が研究されている。これは溝口理一郎の提唱する新しい学問であり、知識処理、人工知能の新しいパラダイムであるという[9]。溝口はオントロジーに関してその語義を、哲学から分類学（タクソノミー）、人工物を含めたものまで広範に考察している。そして溝口はタスクオントロジーを”診断、設計、制御などのエキスパートシステムが対象としてきたタスクに固有の問題解決過程を既述するために必要な基本概念（語彙）の体系的理論”と定義する。またターミノロジーを「用語論」と規定し、「用語論」では「概念に関する合意を得た後に、その概念に付与すべきラベルを決定する必要があるが、ラベルを何にするかを論じること、言い方を換えれば、概念の呼び名を議論の対象とする」ことと定義している。さらに溝口はオントロジー工学において、オントロジーの利用のなかでターミノロジーの標準化、概念の意味内容（基本意味素）の標準化にふれている。

ここで以下いくつかのデータベースにおけるオントロジーの事例を紹介する。

1）高分子データベース

図7はJST作成の高分子データベースである。[10]図では、共通の性質をもつ高分子のデータ内容を「クラス」としてグループ化し、さらにクラス間の関係を記号により示している。この関係には1対1，1対多、多対多があるが、多対応側のクラスに●記号が付されている。菱形は全体ー部分関係を、三角形は階層関係を示している。全体としてモノマー側から化合物としての視点で見た高分子と、材料としての視点から見た高分子とに明確に整理され複雑な関連性を明確に捉えている。

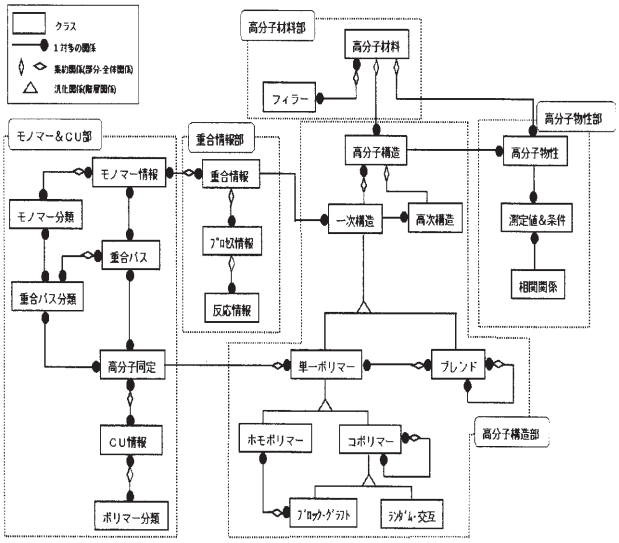

図7　高分子データベースのオントロジー（出典　参考文献10,p34）

2）バイオインフォマテックス

図8は主に大腸菌と酵母に関する代謝パスウエイと遺伝子制御のデータベースであるaMAZEの場合である[11]。ここで実線は概念の上位下位関係を、破線はポイント先を示している。データベースの構成要素が生物学的実体と相互作用で構成されており、グラフで表現すると、生物学的実体がノードとなり、相互作用がノードをつなぐエッジとなる。生物学的実体は化合物、遺伝子、RNA、タンパク質に細分され、相互作用は変換、制御、間接相

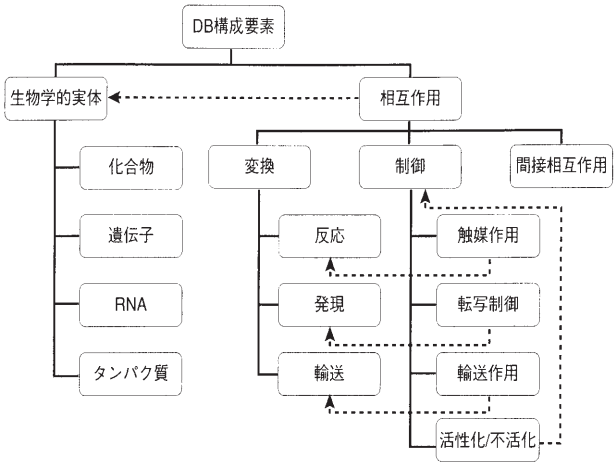

図8　バイオインフォマテックスのオントロジー（出典　参考文献11,p1864）

互作用に細分される。例えば発現エッジは転写制御という制御エッジで制御されるという知識を表現している。

2）大規模電子化辞書

筆者も参加した大規模知識ベースとしてのEDR（電子化辞書）を紹介したい[12]。これは日本版のターミノロジーデータバンクであり、通産省（現経産省）主導の、コンピュータ主要8社による「大規模知識ベース」のプロジェクトであった。EDRは単なるターミノロジーの機械処理以上に、コンピュータによる文章解析、翻訳作成、知識獲得、自動索引作成といった多くの機能実現を目指していた。プロジェクトそのものは終了したが、成果として以下にのべる「概念辞書」などがある。概念辞書は一般の辞書とは異なり、概念を体系化し、それを記述した辞書である。図9は概念辞書などの電子化辞書の構造を、図10はいくつかの関係子を、図11はEDRの概念既述などが示されている。図9の事象間関係子や意味的関係子にオントロジーの全体、部分、継承関係、事象の時間的関係などが表現されている。これによって知識の正確な表現がなされ知識の獲得や機械翻訳などが柔軟になされることなった。

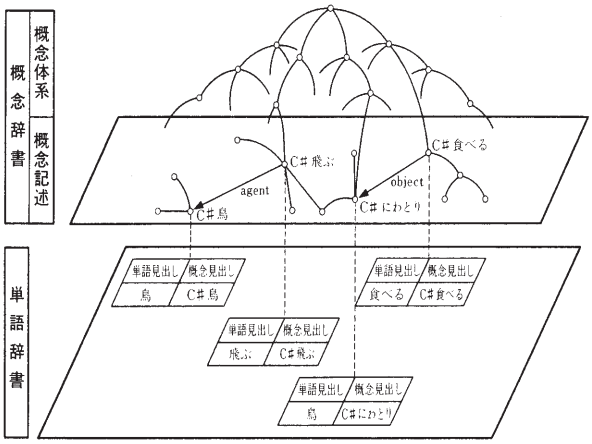

図9　電子化辞書の構造（出典　参考文献12,p110）

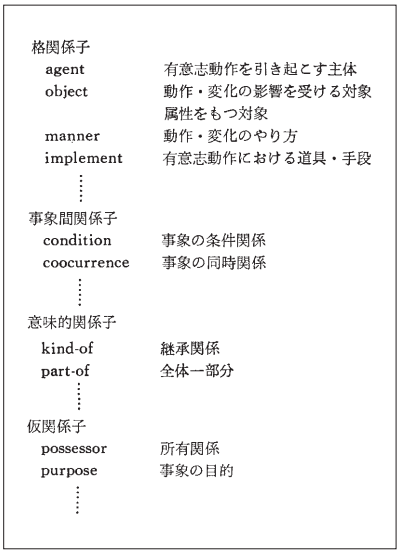

図10　電子化辞書の関係子（出典　参考文献12,p115）

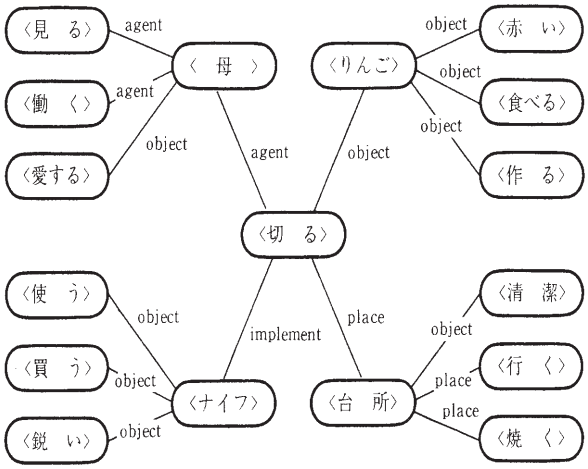

図11　電子化辞書の概念記述（出典　参考文献12,p116）

4. 情報検索と研究評価と未来のコンピュータ

これまではコンピュータにはコトバの意味関係などは理解できないものとし、専ら人間の側からコンピュータにそうした関係をプログラムで教えてきた。しかし最近はどうやらコンピュータも賢くなって人間に近づいてきたようである。そこでここでは研究と検索の関係や、情報検索と研究の評価などの関係、さらにこれからのコンピュータの可能性について若干のべたい。

4.1　研究と情報検索

これまでインターネットの拡大や情報の制御の技術を紹介し、おもに情報検索の最近の進歩をのべてきた。たしかにこれにより従来とはくらべものにならないくらいの利便性をわれわれは手に入れることが可能となった。しかしすでに述べたように量的な増大に反して増大するノイ

ズを回避する質のよい検索が求められている。また研究という視点からは必要な情報はあまり増えていないという指摘もある。またこれは先述の引用分析のガーフィールド自身も、インパクトファクターが本来の目的である「雑誌の評価指標」としてではなく、個々の著者の影響力を図るものとして使用されるのは望ましくないと何度も言っている。つまり引用分析の量的結果を安易に研究や研究者の質的評価に使ってはならないということである。ここには「引用」という情報における質と量の問題が潜んでいる。

4.2 コンピュータは隣人?

研究におけるコンピュータの役割を考えてみると、一つにはコンピュータは研究のための道具という考えがある。しかしいまや研究においてコンピュータは上述のさまざまの機能を備え研究上必要不可欠なものとなったことも否定できない事実である。

さらに近年は、コンピュータにおいて脳や生体のアナロジーやシミュレーションによる研究も盛んである。例えば筆者も脳の仕組みをシミュレートしたSAVVYというブール型ニューラルネットワークシステム[13]、一点交差法によるGA(遺伝的アルゴリズム)、免疫ネットワークシステム[13]などのコンピュータの自己組織化を試みた。そこでは例えば「巡回セールスマン」問題への適用などでコンピュータによる「学習」やある種の問題解決などの実感を得た。しかしそういったコンピュータの振る舞いがただちにコンピュータ自身による問題解決や意志決定、発想とよべるかどうかは疑問である。さらに最近の動向として4種類のDNAの塩基の機能をアナロジーしたDNAコンピュータなどのバイオコンピュータの可能性がある[15]。そしてこれからのコンピュータではコンピュータ自身があたかも人間のように自己組織的に発想することも可能になるともいわれている。 そうなると人間とコンピュータとの関係や、人間の発想や創造性とはなにかが改めて問われてくる。こういった方面の研究としては、例えば各種の創造技法と結びついた発想支援システムがある。われわれは市川亀久彌「等価変換法」に基づくベクトル型特許・発明発想支援システムを構築中である[16]。そして近い将来こうした研究の蓄積によって情報の概念や情報学概念の大きな転換がもたらされる予感がする。色々な意味で今後の動向に注目し期待したい。

5. おわりに

本稿ではまずネットワークや情報の可視化などの最近の傾向を紹介した。つぎにこれらの根底にある問題点としてバイオインフォマテックスや高分子化学、電子化辞書などにおけるターミノロジーやオントロジーを検討した。最後にこれからのコンピュータの可能性を考察した。このことを通して研究と情報検索との関係、コンピュータにおける発想や創造性について考察し展望した。

参考文献

1) 長塚　隆、WWWデータベースと情報文化、第12回情報文化学会全国大会予稿集,2004

2) 福島俊一、webサーチエンジンの基本技術と最新動向(上)、情報管理,46(6),2003、363-372

3) Garfield、E.“Citation Indexing. Its theory and applications to science, technology and humanities”,Wiley,1979

4) 岡谷　大、引用文献分析とターム調査からみた農学と薬学ー農芸化学と獣医学の事例、薬学図書館、**35**(4), 248-255, 1990

5) “特集セマンテックweb”、情報処理、**43**(7), 705-750, 2002

6) 竹内道雄、タンパク質と遺伝情報、in『科学技術情報検索の実際』、162-171、東京農工大学、2003

7) 化学情報協会、『STN　統計解析・ビジュアル化機能』、2004

8) 岡谷　大、尾関周二、『ターミノロジー学の理論と応用ー情報学、工学、図書館学ー』、東京大学出版会、2003

9) 溝口理一郎、タスクオントロジーとオントロジー工学、in:『新工学知1　技術知の位相』東京大学出版会,107-127,1997

10) 前田知子他、JST高分子データベースPoLyInfoの研究(1)、情報管理、**43**(1)、30-35、2000

11) “特集生命のシステム的理解に向けたバイオインフォマテックス”,実験医学**20**(13)、2004

12) 横井俊夫、『日本語の情報化』、共立出版、2000

13) 岡谷　大、ニューラルネットと情報検索、第29回情報科学研究集会発表論文集、57-63,1992

14) 岡谷　大、免疫ネットワークと創造性、第7回発想支援シンポジウム論文集、計測自動制御学会,1995

15) 野島　博、DNAコンピュータの仕組み、現代化学、(4), 46-50, 2002

16) 岡谷　大他、トータルな特許・発明評価および発想支援システムも構築ー視点の変化と技術の展開ー、第1回情報プロフェショナルシンポジウム、2004