ISSN 0913-50

# 学術情報センター紀要

## Research Bulletin of the National Center for Science Information Systems

第 10 号

1998 年 3 月

文部省 学術情報センター

# 学術情報センター紀要第10号

## 目　次

**調査研究分野**



**講演**

Research Bulletin
of
The National Center for Science Information Systems
Volume 10, March 1998

Table of Contents

## Research Trend Reseach Area



## Lecture Notes

学術情報センター紀要　第10号（1998）

巻頭言

# 学術論文のフォマット

山口東京理科大学教授・付属図書館長　清水　忠雄

数年前までの３年間、東京大学の図書館長を努め、いまは新設されてから間もない山口東京理科大学の図書館長を努めている。ごく粗い言い方をすれば、日本で一番大きな大学図書館から日本で一番小さな大学図書館に移ってきたことになる。規模の違いだけではなく、学術情報ネットワークの上では、送信・維持の立場から、受信・利用の立場へと質的転換もあった。立場が変わるとまた問題もいろいろ出てきて、あの頃ああしておけばよかったなど悔いても後の祭りである。

小規模図書館からすると学術情報ネットワークの存在は、有り難いものであることは疑いもない。しかし思ったほど使い勝手のよいものでないことも確かである。長い時間としかるべき費用が費やされたあとで結局情報として役立たないこともしばしば起こる。とどのつまりコンピューター画面上で即論文がみられればと願うこと切である。学術論文の全文データベース化の事業の一層の促進が期待される。

ところで学術論文の形式は実に美しい形にできあがっていると言える。基本的には一編毎に完結する体裁をもっている。まずはじめにその研究が行われた意図、すなわちその分野での研究の流れの中で、自己の研究がどう位置ずけられるかが明らかにされる。次に報告の本体ともいうべき何がなされたか、どういう結果が得られたかが記述される。つぎにその結果がもつ意味や価値、問題点が既存の成果などと比較されながら議論される。最後に問題の発展性や応用の可能性などが予測もまじえながら述べられる。最近はレター形式で簡略化されたスタイルもあるが、ストーリイとしては同じような流れを踏襲しているように見える。私は物理学とその周辺の分野の論文しかみていないが、おそらくほかの分野の論文も同じではないかと推察している。100年まえに書かれた論文をみてもほとんど同じ形式で書かれていることに驚かされる。

学術論文は学術情報の伝統と記録・保存の二つの重要な機能をもたされていることは言うまでもない。そして好むと好まぬとにかかわらず論文が個人の業績の評価に使われていることも現実である。さて、情報の伝達という面からみると、現在の論文の形式は大変〝紳士的〟でまずまず満足すべきものといってよいであろう。しかしやや専門外の人からみれば十分に意を尽くしているとは言えず、また専門家や同一分野の論文を多数読む人からみれば、冗長の感は拭えない。なかなかポイントがつかめずやきもきさせられた経験をもっている方も多いと思われる。冗長なストーリイのなかに言葉では伝達できないいわば言外の情報も存在するという事実を全く否定する訳ではないが。記録という機能からいってもほぼ同じことがいえよう。ということであれば、中途半端な性格でもあろうか。すくなくともいま電子メディヤがもっている性格、もつべきと要求されている機能と必ずしもよくマッチングがとれている形式とは言えないのではないか。

電子的ネットワークに最も適した学術情報の伝達形式はなになのか？それは恐らく冊子体で使われてきた形式とは抜本的に異なるものであろう。ホーム・ページといってもその形式や内容はひとつではないが、それに似たものなのかもしれない。最近ある種の学術情報の伝達に既にこれが使われているのはよく知られている。しかしその中でみられるものは従来の論文の形式を踏襲しているようだ。ホーム・ページという形式は電子媒体の中で生まれて来たものであるから、それなりの説得力はある。書き込み書き換えの自由さ、グレードに応じたアクセスの可能性など魅力もある。しかしこれが学術誌にとってかわるものとは思えない。学術論文がもつべき機能を十分に満たしていないからである。ところでこういうことに関する研究は学術情報センターなどで既に十分行われているに違いない。寡聞にして私が知らないだけである。まだこれでよいという段階には至っていないであろうが是非現状を知ってみたいと思っている。

学術情報センター紀要　第10号(1998)

研究論文

# メディアシー—使命と方向性

# Mediacy: what it is and where to go

学術情報センター　内藤　衛亮
Eisuke NAITO
National Center for Science Information Systems
学術情報センター　越塚　美加
Mika KOSHIZUKA
National Center for Science Information Systems
学術情報センター　井上　如
Hitoshi INOUE
National Center for Science Information Systems

**要旨**

「デジタル・リテラシー」をめぐる問題、課題及び共通の利害に関する枠組みを設定するために概念と文脈を展望した。デジタル化された知識へのアクセスの提供、情報リテラシーの強化などのために、初等・中等教育、高等教育、生涯教育分野で日本において実施されている政府の施策について展望する。最近の日本における最近の開発過程で認識された課題について、その共通性を先進国、発展途上国いずれとも共有すべく検討する。将来的な国際協力の可能性について検討する。

**ABSTRACT**

Concepts and context are reviewed in the light of "digital literacy" to set the framework of the problems, tasks, and common interest. Government actions taken in Japan are reviewed in terms of providing access to digital knowledge, creating information literacy such as in schools, higher education, and life-long education, etc. Tasks, which have been found during the recent development in Japan, are explored for common interest to be shared with advanced as well as developing countries. Possibilities shall be discussed for the future international cooperation.

[キーワード]　情報リテラシー、情報活用能力、デジタル情報、日本における政府方針、こねっとプラン、100校プロジェクト
[Keywords]　Information Literacy, digital information, Japanese government programs, Konet Plan, 100 Scool Project,

## はじめに

本稿は1997年3月10日から12日にかけてモナコ公国モンテカルロにおいてユネスコ主催により開催された「情報倫理—デジタル情報の倫理的、法的、社会的諸側面に関する国際会議」(INFO-ETHICS - International Congress on Ethical, Legal and Societal Aspects of Digital Information, Monte Carlo, Monaco, 10-12 March 1997)における招待講演(Hitoshi INOUE, Eisuke NAITO, Mika KOSHIZUKA. "MEDIACY: what it is and where to go")の日本語訳である[1,2]。

ユネスコ主催のこの会議は、デジタル技術の発展と世界規模のネットワークにおける応用が急速に進展していることから、社会のあらゆる側面での利用が可能となりつつあるという認識のもとに、この状況からすべての国が利するべきであり、その機会(挑戦)を見逃

すべきではないという前提にたって開催された。

この会議に先だって、1995年7月にユネスコ本部において "International Expert Meeting on Legal and Ethical Issues of Access to Electronic Information" が開催され、デジタル情報の倫理的、法的、社会的諸側面に関する国際会議の開催が勧告されたのである。この勧告のもとにユネスコ事業計画1996-1997の一環として加盟国の国内委員会に呼びかけて、ここに紹介する会議が開催されるに至ったのである。

この会議は、地球情報流通基盤(GII: Global Information Infrastracture)における普遍的なアクセスの重要性を確認し、情報社会(Information Society)[3]の進歩を実現し維持する方策を探ろうとするものであり、次のような目的が掲げられていた。

1　地球情報ハイウェイ上におけるデジタル・マルチメディア情報の製作、アクセス、配給・流通、蓄積保存および利用に関して、国際的な関心を呼ぶべき、主要な倫理的問題を明らかにする。
2　これらの問題に関して、各国が政策立案の際に考慮すべき必須の原則を明らかにする。
3　協力のための国際的長期計画提案を立案する。

会議では「デジタル情報へのアクセス」、「デジタル情報および記録の保存」、「マルチメディア環境に対する社会的準備」という三つのテーマのもとに、さらに数多くのトピックに分かれて発表と議論が進められた[4]。事前のプログラムでは基調講演3件を含めておよそ40件の発表が予定されていた。これらの論文のフルテキストは現在もユネスコのホームページに掲載されて参照できる[5]。

本稿の目的は、日本における情報リテラシー(情報活用能力)対応の各種活動について、特に文部省を中心とする活動を紹介することにより、欧米以外における現状の一端を明らかにし、かつ、国際的な動向のもとに、わが国における発展の方向を考察しようとするものであった。また、アジア地域における情報リテラシー振興における条件について指摘し、情報先進国(したがって情報所有国)と情報後進国とのあいだにある格差・断絶を認識することによって、情報共有・情報流通の円滑化の可能性について検討した。招待講演の性格から紙数、発表時間などの制限が与えられていたこと、また参考資料の多くは日本語のみのものであったことなどから、記述は簡潔にならざるを得ず、したがって、大いに独断的かつ断片的である。

## 1　デジタル・リテラシー：文脈と内容についての認識

### 1.1　リテラシー：定義

1990年は「国連識字年」(UN International Literacy Year)であった。この年から識字率を高める10年計画が開始されたのである。これに先だって、1989年1月には米国図書館協会(ALA: American Library Association)は次のような定義を公表している。

「情報リテラシーは情報時代における生き残りのための技能である。日常生活にあふれる情報に呑み込まれるかわりに、情報について認識し能力のある人々は、効果的に特定の問題を解決するためや意思決定のために、情報を発見し評価し利用する方法を知っている。この場合に、人々が選ぶ情報はコンピュータ、本、政府機関、フィルム、その他の無数の情報源にある。」[6]

「リテラシーの今日的意味」を追求して Behrens は1970年代から1990年代にかけての情報リテラシーに関する歴史的な文献展望を提示している。彼女は情報リテラシーの今日的定義が可能であるとしながらも、「この概念は個人が機能として存在する多様な発展段階の連続体を意味している」としている。彼女が展望の対象とした1970年代からの文献のなかでは、コンピュータ・リテラシーから一般的な情報処理の知識と技能へと概念の移行をあとづけることができた。彼女が指摘した傾向は、情報リテラシーがもっと広義のリテラシー連続体の一部分と見なされているというものであった。Behrens は次のように要約している。

> 「リテラシーは進化する概念と見られており、その意味は特定の社会の社会的・個人的要求に依存したものである。リテラシーは、個々の文化、社会、政治的な文脈において再考察されるべきものであり、その定義はその社会における拡張しつつある情報要求を考慮しなければならない。」[7]

リテラシーは McGarry が定義したように、定義するにはつかみどころのない概念である[8]。書くことと読むことという人間活動のためのトゥールと組み合わされた多くのリテラシーがある。これらの専門用語としてのリテラシーは、社会過程の個々の構成要素が織りなす結び目の見えない蜘蛛の巣の部分なのである。手は脳のトゥールであるが、脳は、また、使用するトゥールの広がりと範囲によって恒常的に影響を受けている[9]。リテラシーは、いかに粗雑なものであれ、書く技術の入手可能性に依存している。その意味で、

ここに経済的な基礎がある。リテラシーを定義し、測定するためには、文化的な文脈が必要である。理想的なリテラシーとは、McGarryがScribnerを引用して述べているように「共時的に適合可能であり、社会的には権限を付与するものであり、自らで向上するものであり、人間精神を向上するものである」、また、「自らの人生と他の人々の人生における意味を批判的に思考し、探索する能力」である。このように多くの種類のリテラシーが存在しており、それぞれに対応した専門性と専門的な関心を持つグループが存在している。

McGarryが示唆するように、リテラシーと教育はプラトンの時代から政治的な問題であった。機能的なリテラシーに、所与の文化的文脈における測定可能性と適合可能性、および人的投資における見返りの可能性を意味するものである。機能的リテラシー概念によって定義される技能とは、機能的な社会グループが必要とする水準と相対的な関係にある。McGarryは社会における情報の蓄積と伝送に関して次のような5段階を設定している[10]。

口承の段階(Oral stage)
文字の段階(Alphabet stage)
写本の段階(Manuscripts stage)
刊本の段階(Print stage)
電子の段階(Electronic stage)

これらの段階は非連続的な調和のなかで共存している。それぞれの段階において人々は知識を利用する知識と技能を必要としてきたのであり、そこで、リテラシーの意味も変化してきたのである。

インターネットとWWW(ウェブ)の出現と普及は、いずれ来るであろう日常生活のイメージを特徴付けるものである。情報リテラシーは、米国図書館協会によるバランスの良い定義はあるものの、通常の意味合いとしてはコンピュータ・ハードウェアにのみ限定する傾向がある。一方、ネットワーク情報資源の理解と製作は、市民生活における基本的要件の一つとなりつつある。ネットワーク情報資源の内容の多くはデジタル・データである。McGarryの示唆には反するのだが、ここに情報リテラシーとデジタル・リテラシーのあいだの区分が必要とされる理由があるのである。

マルチメディア・コンテンツは、データ・コンテンツとそれを処理するハードウェアを含むメカニズムの統合的実体である。この点では、コンピュータ・リテラシーという用語はマルチメディア・データを含むデータ処理の能力を含むべく拡張されるべきである。このような議論を経て、メディアシーというもう一つの新しい用語が、McGarryが定義する電子の段階における社会的要求を満たすべく登場したのである。しかしながら、この新しい用語は社会過程において孤立したものではなく、連続性のある蜘蛛の巣の一部分なのである。メディアシーという用語を使用する上での重点は、コンテンツとしてのマルチメディアおよび情報トゥールとしてのコンピュータおよびネットワークの両方を使いこなす能力を指しているところにある。

### 1.2　情報リテラシー：日本における定義

近年、日本の文部省が推進している、情報リテラシー教育に関連した教育課程改善は、臨時教育審議会の答申に示されており、その力点は次のようなところにある[11]。

1) 情報活用能力の育成：将来の高度情報社会に居きる児童生徒に必要な資質として「情報活用能力」(=情報及び情報手段を主体的に選択し活用していくための個人の基礎的な資質)を「読み、書き、算盤」と並ぶ基礎・基本として位置付け、学校教育においてその育成を図ること。
2) 情報手段の活用による学校教育の活性化：コンピュータ等の情報手段を主体的に活用することにより、教育方法の改善・充実を図るなど、学校教育の活性化に役立てること。
3) 情報モラルの確立：情報及び情報手段に関して、その重要性、価値、責任等についての基本認識(情報モラル)を確立すること。
4) 情報化の「光と影」への対応：情報化のプラス面を最大限に生かすとともに、情報化のもたらすマイナス面(情報への過度な依存、間接体験の肥大化、情報犯罪等)について、これを補うために教育上適切な配慮をすること。

日本の文部省が定義する情報活用能力(information literacy)は次の四つの要素で構成されている。[12]

1) 情報の判断、選択、整理、処理能力及び新たな情報の創造、伝達能力
2) 情報化社会の特質、情報化の社会や人間に対する影響の理解
3) 情報の重要性の認識、情報に対する責任感
4) 情報科学の基礎及び情報手段(特にコンピュータ)の特徴の理解、基本的な操作能力の修得

ここに示した用語定義の日本版すなわち「情報活用能力」は「情報と情報メディアを主体的に選択し使用

するための個人の能力」を意味しており、「読むこと、書くこと、数えること」と同等の位置付けにあり、この能力を促進することは「批判的な思考能力」や「情報および情報メディアの批判的評価力」を増強することと等価のものとされている。

### 1.3　アジア地域に対する関係

アジア太平洋および中東地域における情報スーパーハイウェイの背景には多様性がある[13]。この多様性はこれらの地域の次のようなさまざまな側面に存在している。

人口
経済発展
社会開発
文化
言語
宗教
教育システムとリテラシー
情報媒体の普及・拡散

これらの多様性を体現している人口は、世界人口、地理的範囲、エネルギー消費などのなかで相当規模の割合を占める。アジア諸国における情報利用(アクセス)もまた紙媒体を介する伝統的な様式からデジタル様式を介する高度な方法など多様化している。情報利用における格差はこれらの地域と先進国の間にも存在し、また、各国の都市部と農村部の間にも存在する。これらの特色が[メディアシーの]基本要因を構成しており、地球情報基盤構造(GII)のもとにおける情報スーパーハイウェイを地域的に開発を進めていく際の条件となる。

Torrijosは、国家情報基盤(NII)を開発するために、長期計画(strategies)の構想・立案において、次のような要因との関連において、劇的な移行が起きつつある傾向を指摘している[14]。

- 文化的複数主義(人種多様性)
- 科学的価値の開発を振興すべき必要性の増大(分析技能、解釈技能)
- 文化の積極的側面の支援
- 教育改革
- 女性蔑視を推進する文化的慣習の抑制
- サイバー・カルチャーとして、社会を情報プアと情報リッチに階層化
- 法制的、倫理的問題

メディアシー、デジタル・リテラシー、あるいは情報リテラシーさらには情報活用能力が何であれ、国家は、とりわけアジアにおいては、情報教育という面では西欧諸国に追いつこうとしているし、また、近隣諸国と競争している。しかし、フォント・スタイルを論外にしても、文字が100個以下にすぎない西欧諸国とのあいだには、文字使用という点で決定的な違いがある。例えば、日本の文字はおよそ1,500年以前に中国から渡来したものであるが、8,000以上の個数がある。中国における漢字を含めるならば、漢字の個数は6万を越える。この数はボールド[ゴチック]やイタリックといったフォント・スタイルを含めるとなれば、数倍になる。状況をさらに悪化させるものは、新しい文字が絶えず発生していることである。中国語、日本語、韓国語の文字[漢字]は、中国に根ざすものと言えるが、進化し、かたちを変え、新しい意味を追加するという生命力を持っている。

これらの国々に1970年代の早い時期にコンピュータが導入された段階で、各国の母国語処理はアプリケーション開発において必須の要件であった。洗練度の低い文字入力方法が、リテラシー問題の近代化にもう一つの障害を付け加えてしまい、現在も情報システム開発における主要な問題となっている。

キーボード操作を学習することの困難さを検討する以前の問題として、文字に関するリテラシーこそが、すべてのリテラシーにおけるもっとも基本となる問題となっているのである。この基本的かつ古典的リテラシーは中国語、日本語、韓国語以外のアラビア文字、インドの各種の文字など他の多くの文字においても共通する問題なのである。これらの非常に大きい個数からなる文字・言語は非西欧国にこそ多く存在し、また、これらの地域においてこそは、そもそものリテラシー問題が主要問題であったのである。この意味からこれらの国々におけるメディアシーというよりはコンピュータ・リテラシーは複線的な構造のもとにある[15]。

- 西欧のキーボード操作(aからzまで)に親しむ(利用者側)
- 西欧のアーキテクチャの上で、母国語を生成する方法を学習する(利用者側)
- 西欧で開発された製品を母国語環境に合わせて再設計する。これはさらに投資を必要とし、製品公表の遅れを生み出す(製造者側)
- 再設計が繰り返し行われるために、製品同士(同一製品の新旧)の間に矛盾が生じうる(製

造者側)

これらの問題が解決されたあかつきには、国としてのインテグリティに対する文化面での植民地化を防ぐことができるだけでなく、多くの側面において国としてのインテグリティを維持しているコンピュータ産業の植民地化をも防ぐことができる。

## 2　デジタル時代を志向した政府の対策：教育システムにおける日本の経験

### 2.1　文部省における政策開発

臨時教育審議会(National Council on Educational Reform)は首相の私的諮問機関として1984年8月に設置された。1987年8月までに、審議会は次の8件のテーマに関する4種類の報告書を連続して提出した[16]。

- 21世紀に必要とされる教育における基本的要件
- 生涯教育のための組織・制度の整備及び教育的背景に対する認識不足から生じた弊害の是正
- 高等教育の拡充および高等教育機関の個性化
- 初等・中等教育における拡充と多様化
- 教員資質の向上
- 国際化への対応
- 情報時代への対応
- 教育行政、教育財政の見直し

1987年に出版された最後の報告書において、審議会は教育改革における三つの基本的視点として次のような課題を提示した[17]。

- 個性を重視する原理
- 生涯教育制度への移行
- 国際化と情報メディアの普及を含む社会的変化への対応

一連の報告書に対応するため、1987年10月に内閣は小学校、中学校における情報リテラシー教育の現在の方向を定めるために、指針「教育改革に関する当面の具体化方策について」を公表した[18]。

1989年に文部省は情報教育を重視して小学校、中学校の教科課程を改訂した。この改訂は次のような観察を基礎として行われた[19]。

1) 情報化社会：産業・商業におけるコンピュータ応用、銀行・観光業、製造業、卸、小売り店などにおけるネットワーク応用はまもなく家庭にリンクされる。コンピュータおよびネットワークが処理する知識のかたちの発展(例えばマルチメディア)、家庭電気製品におけるコンピュータ応用(洗濯機、ビデオ、電話など)
2) 生活の情報化：放送(ラジオ、テレビ、新聞、衛星放送、ケーブルテレビ)、テレビ・ゲーム
3) 地域の情報化：図書館、博物館など生涯教育施設におけるコンピュータ応用は市民が個別に学習する機会を提供する
4) 情報化の特徴：i)デジタル化、ii)ネットワーク化(コンピュータと通信)、iii)データベースの構築と利用、iv)各種メディアの統合・接近(マルチメディア)
5) 学校教育における情報化に対応した教育の必要性：児童生徒はすでにコンピュータやコンピュータ応用に取り囲まれている。彼らに対する情報リテラシー教育は社会人となるために必要とされている

情報リテラシー教育との関係において、改訂されたカリキュラムの重点は先に挙げた(注11参照)。

### 2.2　日本における各種事業

日本における情報リテラシー(メディアシー)を推進する活動はすでに数多い。文部省は初等・中等教育にパーソナル・コンピュータ(PC)の配置を1987年に開始している。現在(1997年2月時点)の配置目標は次の通りである[20]。

| 学校の種類 | 学校総数 | 目標台数 |
|---|---|---|
| 小学校 | 23,977 | 22 PC (コンピュータ学級において児童2人当たり1台) |
| 中学校 | 10,498 | 42 PC (生徒1人当たり1台) |
| 高等学校 | 3,054 | 42 PC (生徒1人当たり1台) |
| 専門学校 | 902 | 8 PC (生徒1人当たり1台) |

この事業が10年を経過して、小学校のおよそ85%にはパーソナル・コンピュータの配置が完了している。この事業は文部省が遂行する政府施策のひとつである。この他にも公共部門、産業界が進めている多くの事業がある。以下では、そのうちの二つの事例を紹介する。それらは、こねっとプラン(産業界主導の事業)と100校プロジェクト(政府主導の事業)である。

#### 2.2.A　こねっとプラン

こねっとプランは、学校においてマルチメディア環境の開発を推進しようとする事業である(専門的職員

を支援要員として、第一にネットワーク・リンクを確立し、次にコンテンツ開発を普及させる)[21]。このプランは、大部分の小学校、中学校にはすでにPCがあるという事実に基づいている。このプランは1996年4月に日本電信電話株式会社(NTT)と文部省が共同で開始したものである。

このプランの目標は、ISDNネットワークの設置とインターネット・アクセスの振興を次のような手段によって進めようとするものである(1997年2月時点)。

1) NTTの職員100名が支援要員として全国的に配置される
2) 参加校(小学校、中学校、高等学校)への30万円の寄付
3) ホームページおよびメーリング・リストサービス
4) ホームページ作成に対する支援と振興

1997年2月の時点で、推進委員会には企業・個人合わせて30社/人が参加している。参加者(社)はインターネット・プロバイダ、電気通信業、主要なソフトウェア・ハウス、コンピュータ製造者などである。文部省は全国の学校に情報を提供する役割を果たす。このプロジェクトは、先導的な通信業者による人材支援や資金的援助など産業界主導である点に特徴がある。

### 2.2.B 100校プロジェクト

1993年に通商産業省と文部省は、国内100校にネットワーク・リンクを提供する事業を開始した[22]。参加希望学校を募集して、教職員の能力、PCの設置状況、そして企画の質という基準で選択した。1997年2月の時点で、111校(108校と3視聴覚センター)の参加がある。当初の応募は、1994年の時点では1,543校であった。情報処理振興事業協会(IPA)[23]を事務局として、財団法人コンピュータ教育開発センター(CEC)[24]が推進している。

1995年2月以来、ネットワーク・リンク、通信装置、サーバーおよびクライアント・マシンなどを含む施設・設備の配備が開始された。政府主導のこの事業には次のような重点が強調されている[25]。

a) 自主企画
b) 教育ソフトウェアの共同開発
c) 運用支援
d) メーリング・リスト、ニューズグループ
e) 技術サポート窓口
f) 専門研修会
g) 発表会、研究会
h) 広報資料等

100校プロジェクトが達成したものとして、次のような事項がある。

全国発芽マップ：小学校生徒のグループ学習プロジェクト
酸性雨調査：酸性雨のデータ収集
プロジェクト参加教師のメーリング・リスト
生徒・学生のためのニューズグループ

これら二つの事業は、他にも事業がある中で、日本におけるメディアシー振興のための活動を代表するものである。政府と産業界はコンピュータ、電気通信、マルチメディアの次の段階を目指して協力し、かつ競争しているのである。

## 3 当面の課題

1989年の米国図書館協会の会長勧告において、次のような6項目の勧告がなされている[26]。

1) われわれ全員が、情報を制度的に組織してきた方法、構造化された情報アクセスの方法、家庭、共同体、職場における情報の役割についての定義などをを再検討しなければならない。
2) 情報リテラシーを振興するために、情報リテラシーのための共同活動(Coalition for Information Literacy)をALAの指導のもとに、他の全国的な組織・機関と協調して設立すべきである。
3) 情報とその利用に関する研究とデモンストレーションの事業を開始すべきである。
4) 連邦政府の教育庁(Department of Education)、高等教育委員会(Commissions on Higher Education)および学術関係の行政機関(Academic Governing Boards)は、児童生徒が日常生活や学校において情報リテレイトになり得る環境条件(climate)を実現することに責任があるべきである。
5) 教員養成および教員のパフォーマンス期待値は、情報リテラシー問題を含めたものに改訂されるべきである。
6) ホワイトハウス図書館・情報サービス会議の各テーマに対する情報リテラシーの関連性の理解普及を推進すべきである。

この勧告は十年以上も前のものであり、最後の条項は米国に固有の性格のものであるが、それにしても勧告全体としては、情報リテラシーあるいはメディアシーを振興するためには、先進国においては世界的に適用可能の内容となっている。

公共部門あるいは民間部門のいずれかによるコンピュータおよび通信に対する投資がこれまでになされてきたが、マルチメディア処理、ネットワーク技術に関する専門家集団の養成は、日本における各種プロジェクトにおいて示されているように、メディアシーを志向する社会的動向にとっては緊急の課題である。

多くの発展途上国における緊急課題は、しかしながら、情報教育および情報産業という側面では、米国図書館協会の勧告や日本における各種プロジェクトとは趣きが異なるものであるかもしれない。それは社会的・文化的価値、国家目標などが異なる環境条件にあるからである。また、これらの国々では、情報流通基盤の国家的開発においてのみならず、個々の家庭における[コンピュータ利用]の準備や、特にマルチメディア技術の評価と利用の両面における専門家をはじめとする人材育成などにおいても、巨額の投資が必要とされている。この点において地域的・国際的協力を検討すべきである。

## 4　将来における国際協力の可能性

言語は古典的な情報の障壁であり、それには中国語・日本語・韓国語だけでなく、アラビア語やその他のアジア諸言語が含まれている。そして国際的、地域的あるいは国家的な努力が重ねられてきたのである。

過去10年間、そして次の10年間にあっては、言語障壁のほかに、情報流通基盤の開発が[開発途上国においては]主要関心事であった。ここでも巨額の投資が必要であり、国家的な長期開発が必要であり、途上国は追いつく過程にある。残された地域は、メディアシーの振興のみならず、道具なしに読み、簡単な棒(鉛筆)で書くという、本来のリテラシーそのものの振興に予算を配分するという古典的課題を担っている。ユネスコ国際識字年は、国家の開発状況とは関係なしに、メディアシー振興と合わせて、本来のリテラシー振興というそもそもの使命を依然として有している。国際識字年を再定義することは近い将来における課題を作り出すことになろう。

マルチメディア技術を共有するメカニズムを国際的に創設すべきである。各国の文化的主体性を維持しながら、各国語によるマルチメディア処理の専門家を養成する人材育成プランについても検討すべきである。

## 謝辞

原典執筆にあたっては、野末俊比古氏および通山正年氏からは資料提供ならびに背景説明をいただき、また、倉西美由紀氏からは発表に至る過程で種々の示唆をいただいた。さらにユネスコのInformation and Informatics Division Information Policies and Plans のMr. Victor M. Montviloffには再々の質問に丁寧に答えていただいた。記して謝意を表す。

## 参考文献及び注

[1] 本稿における意見は著者らの意見であり、引用・参照された組織・機関の意見ではない。

[2] 本稿の原典は次のURLで参照できる。http://www.unesco.org/webworld/infoethics/speech/inoue.htm

[3] 地球情報流通基盤(GII: Global Information Infrastracture)は米国のキーワード、情報社会(Information Society)は欧州連合(EU)のキーワードとみなされている。

[4] テーマA「デジタル情報へのアクセス」、
トピック1「情報ハイウェイへの普遍的アクセス」
トピック2「著作権、知的所有権、公正利用」
トピック3「多言語主義と文化の多様性」
トピック4「情報のセキュリティ、プライバシー、自由」
テーマB「デジタル情報および記録の保存」
トピック1「デジタル情報の保管」
トピック2「時期を隔てての情報の信頼性と説明能力」
トピック3「長期保存における法制的要件と実務」
テーマC「マルチメディア環境に対する社会的準備」
トピック1「デジタル・リテラシー(メディアシー)」
トピック2「メディアシー・パートナーシップ：文化セクター、学術セクター、公共セクターおよび民間セクター」
トピック3「地球情報流通基盤における責任」

[5] http://www.unesco.org/webworld/infoethics/infoethics.htm

[6] American Library Association Presidential Committee on Literacy. Final Report. Amer-

ican Library Association. Chicago, Ill. Jan. 1989. 21 p. (ED 315 074 IR 053 029).

[7] Behrens,Shirley J., A Conceptual Analysis and Historical Overview of Information Literacy. College & Research Libraries. Vol. 55, No.4 (July 1994), p.318.

[8] McGarry,Kevin J., "Definitions and meaning of Literacy." In : Keith Barker and Ray Lonsdale ed. "Skills for life ? - The meaning and value of Literacy". (*Proceedings of the Youth Libraries Group Conference, Mason Hall, University of Birmingham, September 1992*). London : Taylor Graham. 1994, pp. 3-17。McGarryは、わずか15ページのしかし刺激的な論文のなかで多様なリテラシーを紹介している。すなわち、アダルト・リテラシー、コンピュータ・リテラシー、文化リテラシー、電気・電子リテラシー、環境リテラシー、実験リテラシー、映画フィルムリテラシー、機能リテラシー、理念リテラシー、情報リテラシー、音楽リテラシー、印刷リテラシー、リテラシー保持力、学校リテラシー、テレビ・リテラシー、視覚(画像)リテラシー、執筆リテラシー、職業に関するリテラシー、心の狭い機能リテラシーなど。

[9] McGarry,Kevin., The Changing Context of Information - An introductory analysis : Second edition. Library Association Publishing, London, 1993. 197p.(Chapter3:pp.59-105).

[10] McGarry. op.cit. p. 93.

[11] 文部省 「情報教育に関する手引き」ぎょうせい 平成3年7月(MESC 1-9118) 230 p.(ISBN 4-324-02387-5)p. 17

[12] 同上

[13] NAITO Eisuke., "Organizational Measures required, at the national level, to ensure the convergence of telecommunications, broadcasting and computer networks, and the conditions for regional and international cooperation." *UNESCO Committee of Experts of Asia, the Pacific and the Middle East on Communication and Copyright in the Information Society*, New Delhi, India, 25 -29 November 1996.

[14] Torrijos,Delia,E., "Address." In : *Report on the experts 'donors' meeting on the development and training of information professionals in Asia and the Pacific*. 14-16 August 1996, Cuezon City, Philippines. pp. 25-27.

[15] NAITO,Eisuke,; SATO.Takayuki,K., "Data Book of Cultural Convention in Asian Countries : In pursuit of common data container." *SEARCC* '96, July 4-7, 1996, Bangkok.

[16] Moura,Seiichiro.; Matsushita,Tomoko.; Nakamira,Masayuki.; Suezaki,Fujimi., Lifelong Learning in Japan: An Introduction. National Federation of Social Education in Japan (ISBN: 4-7937-0084-5) pp.65-72.

[17] 文部省内生涯学習・社会教育行政研究会編「生涯学習・社会教育行政必携」(平成8年版). 第一法規. 平成7年. 1466 p.(ISBN 4-474-00550-3)(「教育改革に関する第四次答申(最終答申)(抄)(昭和六二・八・七臨時教育審議会答申)」pp. 171-188)

[18] 同上 pp.189-191.

[19] 文部省 「情報教育に関する手引き」ぎょうせい pp. 3-6.

[20] http://www.monbu.go.jp/special/j961102.html 表「標準的な学校における整備水準」と表「教育用コンピュータ新整備計画の達成状況」から合成。

[21] http://www.hamajima.co.jp/tim/ko_net.html。こねっとプランの1998年1月時点のホームページは http://www.wnn.or.jp/wnn-s/index.htmlである。

[22] http://www.edu.ipa.go.jp

[23] http : //www.ipa.go.jp

[24] http://www.cec.or.jp

[25] http://somenosuke.edu.ipa.go.jp/100school/ayuumi/activity.html

[26] American Library Association Presidential Committee on Literacy. Final Report. American Library Association. Chicago, Ill. Jan. 1989. p. 14 (ED 315 074 IR 053 029).

学術情報センター紀要　第10号(1998)

研究論文

# 「情報利用学」の構築に向けた予備的考察
## －図書館情報学における情報利用行動研究と情報利用教育研究を中心に－

# Toward Construction of "Information Use Studies" : Studies on Information Use Behavior and Information Use Education in the Library and Information Science

学術情報センター　野末　俊比古
Toshihiko NOZUE
National Center for Science Information Systems
学術情報センター　越塚　美加
Mika KOSHIZUKA
National Center for Science Information Systems

**要旨**

情報環境の変化の中で情報利用をめぐる研究の必要性が増している。学術情報センター研究開発部では、1997年度から情報利用学研究部門が設置された。本稿では、図書館情報学における情報利用行動研究および情報利用教育研究の現状を概観し、問題点と課題を挙げ、研究領域としての「情報利用学」の確立の意義と必要性、およびそのための方向性を論じる[1]。

**ABSTRACT**

Studies and surveys on "information use" are getting more important in the information society. Information Use Research Section was established at R&D Department, National Center for Science Information Systems in April 1997. This article discusses studies on information use behavior and information use education in the library and information science from the viewpoint of "Information Use Studies."

[キーワード]　情報利用学、情報利用行動、利用者研究、情報利用教育、利用者教育、情報リテラシー
[Keywords]　Information Use Studies, Information Use Behavior, User Study, Information Use Education, User Education, Information Literacy

## 1　はじめに

学術情報センター研究開発部には、1997年度から学術情報研究系に「情報利用学研究部門」が新設された。「情報利用学」という名称の研究部門は、わが国の大学(大学院)の学部(研究科)、学科(専攻)、および大学共同利用機関の研究系、研究部門を通して初めてのものである。初めてということからもわかるように、少なくとも現在のところ、わが国において「情報利用学」という名称は必ずしも学問名としては定着しておらず、発展途上のものであるという認識が一般的であろう。このように考えると、「情報利用学」がいかなる研究領域(学問)であるのかを明らかにすることは、一つの使命であると考えられる。

そこで本稿では、今後確立をめざす「情報利用学」とは、どのような方向性や目的を持つ研究領域であるのか、その意義や必要性はどこにあるのか、どのような対象や方法を持つのかなどについて検討するための予備的考察を行う。「情報利用学」が扱う「情報利用」は、「情報」と名のつく他のいくつかの研究領域においてと同様に、理工系から人文社会系まで多様な学問領域にまたがる研究対象であり、いろいろな切り口からの分析が可能かつ必要なテーマである。よって、長期的には、関係する学問領域にわたり広く横断的な検討が必要であるが、今回は、そのための足掛かりとして

図書館情報学における議論に着目し、情報利用学が扱おうとしている課題を洗い出すことで、情報利用学のイメージの一端を描き出すところに主眼を置く。ただし、必要に応じて、かつ可能な範囲で図書館情報学以外の関連分野における状況にも言及したい。なお、図書館情報学の中でも、情報利用行動研究と情報利用教育研究を中心的に取り上げる。

本論文の構成は次のとおりである。まず第2章では、情報利用行動研究の現状について、情報利用環境の変化にも配慮しながら概観し、問題点と課題を述べる。第3章では、情報利用教育研究について、情報リテラシーをめぐる議論、ガイドライン策定等の動向にも触れながら、現状を概観し、問題点と課題を述べる。第4章では、総括として、学術情報センターでの位置づけを中心に、今後の「情報利用学」の方向性と意義に触れ、第5章では展望を述べる。

## 2　情報利用行動をめぐって

図書館や学術情報センターのような機関、あるいは様々な情報サービスやメディアの目的が、極論すれば、利用者の求める情報を提供することにあるとするならば、利用者がどういった情報を求めているのかについて多角的、総合的に把握することは、効果的かつ効率的なサービスのために不可欠であるということができる。図書館情報学においても、そういった観点から、利用者[2]の「情報利用」に関する研究が、主に「利用者研究」と呼ばれる分野で行われてきた。図書館情報学における利用者研究とは、“図書館や情報センターなどの機関や資料の利用者について、その特性、ニーズ、利用行動を明らかにするために行われる調査”であると定義され、その“対象には、直接の利用者に限らず潜在利用者や非利用者も含まれる”[3]とされる。利用者研究は、利用者調査と表現される場合があることからもわかるように、あるメディアをそれらの利用者がどのように使うかについて研究する分野である。なお、ある集団や個人がどのような資料や情報を用いるのか、また、どのように用いるのかといった、「人工物」に焦点を当てた「利用調査」について、ここでは「利用者調査」とは区別しておく。

### 2.1　情報利用行動研究の展開

利用者研究は、「効果的、効率的な図書館・情報サービスの設計や運営のために利用者の情報要求を的確に把握することが必要である」という認識に基づいて、特に第二次大戦後から、利用者の情報要求と利用の実態について調査研究が行われるようになったのが始めとされる[4]。以降、今日まで多くの研究・調査が行われているが、それらについては各所でレビュー、考察がなされているので、以下では大まかな理解にとどめ、2.3における問題点と課題の指摘にスペースを割くことにしたい。

初めて利用者研究の領域区分を行なったのは、1966年のMenzelであるとされる[4]。この頃(Menzelによれば、63年が利用者研究の「離陸」の年とされる)を一つの区切りとして、利用者研究は重要な研究領域として広く認識され、全盛の時期を迎える。ARIST (Annual Review of Information Science and Technology)は、66年の創刊以来、72年までは毎年、それ以降は数年おきにこの領域のレビュー論文を掲載しているが[5]、それらによれば、3～4年間で数百というペースで論文が発表されてきた。初期には科学技術分野を中心とした特定の機関や情報システムに関するケーススタディが多かったが、やがて、一方ではINFROSS[6]や米国心理学会[7]の調査などに代表される大規模調査が行われるようになり、一方では科学技術分野以外の分野、あるいは研究者以外の利用者にも調査対象を拡大し始めるなどの発展を見せた。

しかし、このように隆盛を迎えた利用者研究であるが、比較的早い時期から調査方法を中心として方法論に批判も出されていた。利用者研究は、情報サービス、システムなどの改善には役に立たない、という批判もなされた[8]。こうした動きを受けて、70年代後半には、旧来のやり方に批判的な意見を持つ研究者らによって、新しい視点や枠組みが打ち出され、後にパラダイムシフトが起こったと評される時期を迎える。古いアプローチと新しいアプローチを比較すれば、表1(文献[8]をもとに作成)のようにまとめられよう。

このパラダイムシフトを通して、「何らかのメディアやサービスの利用者」に主に焦点を当てていた利用者研究は、「ある目的を達成しようとするときにどのような情報利用行動をとるか」という広い視点へと枠組みを広げてきた。新しい枠組みでの研究者としては、地方公共団体や企業における情報利用行動の研究を行い、情報探索行動という枠組みを提示したWilson[9]、意味付与アプローチを提唱したDervin[10]、情報探索の契機を自らの知識が当面の課題に対応できない「変則状態」にあるとするASK理論を提案したBelkin[11]などがよく知られている。彼らの個別の研究につ

表 1　古いアプローチと新しいアプローチの比較

| | 古いアプローチ | 新しいアプローチ |
|---|---|---|
| 情報の捉え方 | 客観的 | 主観的 |
| 利用者の存在 | 受動的 | 主体的 |
| 利用者の行動 | 断片的 | 総体的 |
| 着眼点 | (利用者の)<br>外的行動 | (利用者の)<br>認知過程 |
| 人間の捉え方 | 断片の寄せ集め<br>(「カオス」) | 組織化された<br>統一体 |
| 主な方法 | 計量的方法 | 質的方法 |

いては他に譲り、ここでは、“このような新しいアプローチは、今では一種の流行のようになっているが、具体的な成果は意外に乏し”く、また、“しっかりした理論を構築するには至っていない”[12]という点を強調しておきたい。

一方、パラダイムシフトと称される時期を一つの区切りとして、利用者研究への関心が低くなり始めた。例えば、毎年掲載されていたARISTのレビューが72年を最後に、以降74年、78年、86年、90年の4回しか掲載されていないことからもわかるように、図書館情報学における利用者研究への関心は、全体としてはやや下火になった。

しかしながら、このことは利用者研究(情報利用行動研究)が必要でなくなったことを意味するのではない。むしろ、具体的な成果や理論構築が不十分であり、今後の発展を待つ領域であるということは強調しておかねばならない。すなわち、具体的なサービス等への還元という意味では、個別具体の状況における利用という観点が必要である一方で、研究としては、それら個々具体の状況を超えて「情報利用」現象を説明し得る理論が必要であり、それはまだ整備されていない。最近出された情報利用行動に関するレビュー[13]でも、情報利用行動研究による成果は、「情報学」をはじめとする学際領域研究にあって重要な示唆を与えるものであると訴えられている。次節および次々節で述べるように、最近の情報環境の変容の中で、再び情報利用行動研究が重要になってきている。

今日再び注目を集めている利用者研究であるが、これには、全体としては下火だとされた80年代から90年代初期において、積極的な研究者らによって方法論上の議論が盛んに行われ、現在における注目と今後の飛躍に耐えうるだけの土台となる研究実績を築いてきたことが大きな影響を与えている。

この時期の方法論上の議論は、人間の様々な行動を扱ってきた行動科学が人間の心理的な動きをブラックボックスとして扱っていたのに対し、認知科学が人間の内的な動き、すなわち、心理的な動きも科学的な手法によって明らかにできるという前提で様々な研究に取り組んでいったことに大きな影響を受けている[14]。その結果、図書館情報学分野でも、例えば、データベース検索中のログを分析すると同時に、検索をしながら頭の中で考えていることを口に出して話してもらったその内容を転記し、分析の対象とするプロトコル分析の手法や観察法等、人間の内的な思考過程が表出した部分をできるだけ直接考察しようとする研究が多く導入されるようになった。これは必ずしも質問紙法等による回答の統計処理のような量的な手法から質的な手法への転換を意味するわけではないが、図書館情報学分野の中でも利用者研究、特に人間の行動を研究対象とする情報利用に関連する領域においては、質的な手法に対する関心が高まった[15]。

一方、社会学における議論の影響を受けて、観察によって、あるいは質問紙に対する回答として明らかになった人間の行動そのものはどのように扱うべきなのかという問題も浮上していた。すなわち、認知科学的な態度では、客観的に明らかにできるとされた人間の様々な行動は、本当にそのようなものとして扱えるのかどうか、また、それらの一般化は可能なのかといった問題である。

こうした動きの中で、特に現在につながる注目に値する成果を挙げた研究者は、Wilsonであり、Dervinであり、Ellisであるといえる。Wilsonは、「文脈の中の個人(person-in-context)」という考え方を提示し、情報利用行動はそれが生じる文脈と切り離しては考えられないと述べた[16]。Dervinは、「状況-ギャップ-利用モデル(situation-gap-use model)」を示し、「人間は外部に情報を求め、それを変化させる過程でその情報を既に持っている情報体系に意味づけ、情報ニーズを刻々と変化させていく」という意味付与アプローチを提唱した[17]。Ellisは、研究過程における情報検索行動について観察法によって事例を収集、分析し、情報検索過程を構成する行動のカテゴリー化を図り、それらの組み合わせによって検索が行われるというパターンを提示した[18]。Ellisの研究の特徴は、確固とした一般的なモデルを構築しようとしたのではなく、個々の情報探索事例から帰納的に緩やかなパターンを見出

そうという態度である。これらの研究に特徴的なのは、表出してきた情報利用行動そのものだけを取り出して検討しようとするのではなく、それらが遂行された様々なレベルの文脈の中で情報利用行動を捉えようとする点にある。

こうした研究を積み重ね、さらに認知科学や社会学の様々な方法論上の議論や成果を検討しつつ、個々の行動事例をそれらが遂行された文脈も含め様々な角度から検討し、そこから緩やかなパターンを見出そうとする方法論上の一つの方向性が生まれた。そして、現在、こうした質的な手法を取りいれた情報利用研究が盛んになりつつある。

情報利用研究が現在盛んになりつつあるもう一つの要因は、次節でも触れるが、人間の情報利用行動に強い影響を与える要因の一つである「情報環境」自体が、インターネット等の普及により、大きな変化を遂げていることにある。そのため、研究者のネットワーク情報資源の利用や電子メディアの学術的な価値に対する研究等が、様々な形で遂行されている。この流れに対しては、今後の図書館がどうするべきかという情報サービス提供側の検討も盛んに行われている[19]。

ごく最近の動向にも言及しておくならば、1996年にフィンランドで開催された国際会議ISIC(Information Seeking In Context)が挙げられよう。ISICは、第2回が98年にイギリスで開催されることが決まっている。利用者研究が再び活性化しているという裏づけの一つとなろう。

### 2.2 情報利用環境の変化と情報利用行動研究

前節では、利用者研究の枠組みを用い、情報利用行動研究をめぐる経緯について、大まかな時期的区分とともに辿ってきた。現在、情報利用行動研究(利用者研究)は、ISICの開催に象徴されるように、再び活性化が起きており、新たな時期を迎えたといえる。前節の最後でも触れたが、その背景には、情報利用環境の大きな変化がある。

われわれを取り巻く情報利用環境の著しい変化は、われわれの情報利用に様々な影響を与えている。情報利用環境の変化を過去数十年のスパンで見れば、出版量の増加や情報メディアの多様化、コンピュータの登場などといったトピックを挙げることができる。これらは、これまで、情報利用行動およびその研究にも一定の影響を与えてきたといえよう。

1990年代においては、やはりコンピュータ(電子)環境があらゆる場面へ普及、浸透し、ネットワーク環境が展開、普及したことに注目せねばならない。このことは図書館にあっても例外ではなく、例えばインターネットにおけるOPACの公開[20][21]、ホームページの開設と利用(例えば文献利用指導への応用[22][23]など)は、増加を続けている。このように、図書館一つをとってみても、扱うメディアや情報源は増加し、情報探索・入手のためのツール、経路も多様化している。図書館以外にも目を向ければ、インターネット情報資源に象徴されるように、メディア、情報源、ツール、経路の多様化は今さらいうまでもないだろう。

このようにみると、現在および今後における情報利用環境を考えたときに、情報利用との関係からその特徴を端的に表現するには、やはり「電子(ディジタル)化」「ネットワーク化」というキーワードを挙げるのが適当である。もちろんそれまでにも電子化・ネットワーク化は進行していたのだが、90年代にはそれが広く一般にまで普及、浸透したところが重要な点である。すなわち、電子・ネットワーク環境の中で、情報利用者である我々の情報利用も、職業、学習、娯楽などといった生活の諸場面において、かつてと同じではなくなってきている。当然、情報利用にあたっての行動も多様化してきている。あえて例を挙げるならば、目録・索引・書誌等で文献探索を行い、図書館に出向いて図書・雑誌を入手していた利用者が、今では自宅やオフィスにいたままで、ネットワーク上で偶然出会ったテキストからリンクをたどって探すともなく資料を探していくという、電子・ネットワーク環境下に特徴的な行動(これは一般にブラウジングと称することができ、非探索的な行動であり、旧来とは異なる特徴を持つ。これについては、次節を参照)が行われることも珍しくなくなってきている。この意味で、情報利用行動研究もかつてと同じままではありえず、新たな展開の時期を迎えているといえよう。

### 2.3 情報利用行動研究の問題点と課題

情報利用行動研究は、現在でも、図書館や情報センター、データベースのプロバイダやディストリビュータなどの様々な情報サービスの運営、およびサービスの計画・展開のため、あるいは、情報検索のインタフェースなどの様々なシステム開発のため、といった応用の観点からも重要な研究領域である。しかし、2.1で挙げたように、これまでにもいくつかの問題が指摘されてきたし、2.2で述べたような現在の情報利用環境

の観点からもいくつかの問題が指摘できる。以下に主なものを列挙してみる。なお、ここでは情報利用行動研究の問題点を記述し、今後「情報利用学」が扱う課題を明らかにすることが主目的であるので、必ずしも理論的であることにこだわらず、レベルの異なる項目も単純に並置した。

(1) 適切な調査方法の検討と開発

情報利用行動の調査にあたって用いられる手法は、観察法、面接法、質問紙法など、社会調査の方法をそのまま用いたものが多い。これに対しては、2.1で触れたように、情報利用行動にとって適切な調査方法とはいえない、という批判がなされてきた。むろん、調査手法に関する検討、開発も進められてきたが[24]、必ずしも充分とはいえない。質問紙法や面接法などには、それぞれ長所、短所があるが、利用行動の調査にあたっては、そうした特性を踏まえ、適切な方法を選択、あるいは併用していくことが必要となる。ところが実際には、調査の諸条件も考えあわせて常に最適な方法を用いることができているわけではない。

例えば郵送法による質問紙法に代表されるように、これまでは「利用者」を集団として捉え、いわば「最大公約数」的な利用者像を想定する傾向があった。そのため、多様な特性を持った個人の行動は埋没してしまった側面がある。このことに対する批判が1970年代後半になされたことは2.1で述べたとおりであるが、その後は「個人」が重視され、現在では、さらに「集団」と「個人」とを的確に両立したかたちで調査、記述していく方向が望まれている。

このように、現在までの情報利用行動研究は、方法論上の一つの限界点を持っているということもできるだろう。しかしながら、認知科学やコミュニケーション科学など、近年急速な発達を遂げている関連諸分野における成果を援用するなど、これまでの方法論上の障壁を乗り越える可能性を持つ動きも盛んになってきた。これは、2.1でも述べたとおりである。このような分野横断的手法の確立、いうならば総合性は、情報利用学を学問の一分野として確立するためには重要な要素となるものである。

(2) 非探索的行動への対応

「ブラウジング」という行動が、特に人文社会系の場合、研究活動等において重要な役割を果たしていることが明らかにされるにつれ[25]、「ブラックボックス」の多いブラウジングのような情報利用行動についても記述・説明できるモデル構築やそのための調査が重要視されるようになってきた。これまでの情報利用行動研究は、科学技術分野における研究者の研究上の行動という、とりわけ目的のはっきりした情報「探索」行動を扱うところから始まったわけだが(このこともまた、パラダイムシフト期の批判の対象ではあった)、ブラウジングのような「非探索」行動への研究は相対的に遅れているといわざるをえない。

2.2で述べたような現在の情報利用環境では、WWWブラウザに代表されるように、ブラウジングにこそその特性を有するツールが普及し、またそうしたツールを通して流通する情報も増加しており、非探索行動の研究はますます重要になっている。こうした傾向に対応した研究の強化が今後緊急に必要である。

(3) 新しいメディアへの対応

利用者研究においては、特に従来フォーマルな情報源とされていた冊子体を中心とする文献(印刷メディア)という固定的なメディアの利用が主流であり、それらは、比較的観察・記録が容易であった。しかし、2.2で触れたように、現在では、新しいメディアを通して流通する有用な情報が増加しており、また実際に研究者が研究を進める上でもネットワーク情報資源は不可欠なものとなっている。すなわち、調査研究にあたって、ネットワーク情報資源などの流動的なメディアをも想定した記録・観察を行わなければならなくなった。

一方、従来インフォーマルな情報源とされていたメディアの利用行動も多様化している。例えば、電子メールやメーリングリストなどのネットワークメディアを使うことが一般的になりつつあり、そうした環境での情報利用の研究などもいっそう進められるべきである。さらに、従来のメディアとは異なり、物理的な実体のないそれら電子情報資源の利用調査・研究手法についても検討する必要がある。この方面に対する研究は、今のところ、個別の文脈に特化した研究が主流であるが、それは一般化して扱うことの難しさの表れであるともいえよう。なお、(2)(3)は、現在の電子・ネットワーク環境における情報利用行動の多様化への対応を課題とするものであるといえよう。

(4) 情報利用行動における必要技能・知識の検討

「情報化」時代においては、個人に求められる情報利用に必要な技能・知識、つまり、いわゆる情報リテラシー(これについては、次章で詳しく触れる)が高度化かつ多様化しているといわれる。これを自然の経験のうちに、あるいは自らの学習のみですべて身につけるのが容易でないことは、ほとんどの人が経験済みで

あろう。人、組織、システム等、何らかの他者からの的確な支援があることが望ましいし、また実際にそういった支援によって情報リテラシーが高められていることも多い。

情報利用行動を遂行するにあたって必要とされる技能や知識の基盤にあるのが情報リテラシーであると考えるならば、例えば、OPACその他のデータベースの検索過程を研究する場合に、事前に与えられたインストラクションが被験者の検索行動に影響を与えることはよく知られている。情報利用行動が同種のインストラクション、すなわち、情報リテラシー教育によってどのような影響を受けて形成されていくかについての研究は、特に初等・中等教育における一般的な情報リテラシーの形成においても、また、高等教育における専門の学問分野に則した情報リテラシーの形成においても重要な示唆を与えるものと思われる。

## 3　情報利用教育をめぐって

### 3.1　情報利用行動と情報利用教育

2.2で触れたように、情報利用環境の変化によって、われわれが用いる情報源やメディアは多様化し(それにともない、われわれの情報利用行動も多様化し)たが、このことは、当然ながら、情報を利用する際に要求される技能や知識が多様化、高度化していることをも意味している。

情報利用に必要な技能・知識の多様化や高度化にともない、その習得・向上、およびそのための取り組みが現代情報社会の課題として訴えられていることは周知のとおりである。いま、情報利用技能・知識の習得および向上を支援する諸活動、情報利用技能・知識の育成をめざす諸活動を情報利用教育と呼ぶことにすれば、情報利用教育は、現在の情報利用環境にあっては、極めて重大なテーマであるといえる。実際に、学校・職場をはじめとする様々な場面において、情報利用教育は必須のものとなりつつあり、情報利用教育の体制・方法等が社会的な問題になっている。初等・中等教育においては、学習指導要領の例を出すまでもなく、いわゆる情報教育が非常に重視されており、高等教育においては情報関連科目必須化の動きなどに象徴されるように、これも非常に重視されている。企業における情報関連研修等でも同様に重視されている。図書館でも、文献利用指導(bibliographic instruction)というかたちで、長期間にわたって情報利用教育への取り組みが行われてきた。

情報利用行動の多様化・高度化にともなって、情報利用技能・知識も多様化・高度化し、情報利用教育の必要性や重要性が増してきたわけであり、情報利用行動と情報利用技能・知識、情報利用教育は、強いつながりを持っているといえる。情報利用技能・知識を基礎として情報利用行動がとられ、また利用行動の結果や経験を踏まえたフィードバックにより、利用技能・知識も習得・向上されていく。さらに、情報利用技能・知識は、他者による指導や評価というかたちでも習得・向上されていく。

このように情報利用教育は情報利用行動と非常に深い関連を持ちながら、実はこれまで個別に研究されることが多かった。図書館情報学では、「利用者教育」[26]と呼ばれる分野において、上述の文献利用指導を含め、情報利用教育に関わるテーマが扱われてきた[27]。しかし、例えば、情報利用行動の重要な一つである情報探索行動(過程)について分析、モデル化を行い、情報利用能力の育成プログラムの作成に活かそうとする動きなどがあるものの(この点では、情報利用教育は情報探索行動ないしは情報利用行動の応用の一つとして位置づけることもできよう)[28]、利用者研究と利用者教育は、相互に充分な好影響をもたらすまでの関連を持つには至っていない。例えば、ブラウジングなど「探索」以外の情報利用行動において必要となる技能・知識、およびそれと利用教育との関係について考察した研究はほとんど見あたらない。

### 3.2　情報リテラシーの支援としての情報利用教育

さて、情報利用教育に関する研究は、前節で述べたように、情報利用行動研究との関わりからも、また、実践のための枠組みづくりという意味でも重要である。これまで図書館情報学をはじめ、様々な分野で研究、あるいは実践が進められてきている。前節まで漠然と「情報利用技能・知識」などと呼んできたものは、特に1980年代後半以降、わが国では90年代に入ったあたりから、「情報リテラシー」と呼ばれることが多くなった。現在では、情報利用教育について論じるときには、この概念は不可避なものとなりつつある。以下では、情報リテラシーを一つのキーワードとして検討を進めていく。すなわち、本節では、情報利用教育を情報リテラシーの育成という視点からとらえて論じていく。まず情報リテラシーとは何であるのかについて触れ(3.2.1)、わが国での考え方について、教育行政の動向においてとらえ(3.2.2)、ついで、情報リテラシー

の育成に対する図書館界としての取り組みとして、ガイドライン策定の動向について概観する(3.2.3)。

### 3.2.1 情報リテラシーの定義

図書館情報学、教育学など様々な分野で、その必要性が強く訴えられている情報リテラシーであるが、従来および現在もいくつかの捉え方がなされている。図書館情報学ないし図書館界では、米国図書館協会(ALA)の報告書から“情報の必要性を認識し、必要とされる情報を効果的に発見し、評価し、利用する能力”という部分が定義として用いられることが多い[29]。このALAの報告書は、図書館界だけでなく高等教育の分野にも影響を与えるなど、一定の評価を得ることができた。報告書で示された方向性がある程度適切であったためであるといえよう。

ここでは、ALAの報告書から3年後に出された、全米情報リテラシーフォーラムの報告からの定義を挙げておきたい[30]。すなわち、情報リテラシーとは、“様々な情報源から情報にアクセスし、評価し、利用する能力”であり、情報リテラシーの具体的な属性として、次の諸点が挙げられる。

- 情報ニーズを認識する
- 正確で完全な情報が知的意思決定の基礎になることを認識する
- 情報ニーズに基づいて質問を定式化する
- 利用可能な情報源を同定する
- 効果的な探索戦略を立てる
- コンピュータなどの技術を利用した情報源にアクセスする
- 情報を評価する
- 実際の適用のために情報を組織化する
- 既存の知識体系の中に新しい情報を統合する
- 批判的思考と問題解決において情報を利用する

この定義は、米国を中心とする図書館関係団体、教育関係団体などのトップに対し、デルファイ法を用いて行われた調査研究から導き出されたものである。現在出されているものの中でも、最も共通理解を得た情報リテラシーの定義の一つとしてみなしてよい。

さて、上記の定義を見ると、情報リテラシー自体は、ごく抽象的な定義になるものであるが、その属性として挙げられている十の項目が注目される。これらは、大雑把にいえば、一連の情報利用(探索)行動に相当していると見なすことができる。すなわち、この点において、情報リテラシーを「効果的な情報利用(探索)行動を行うため技能・知識」と再定義できる可能性が示唆される。

### 3.2.2 「情報活用能力」の定義

わが国では、情報リテラシーの育成の問題は、教育分野で特に盛んに論じられ、実践も様々なされてきた。教育分野においては、情報リテラシーは情報活用能力と呼ばれてきた。情報活用能力と情報リテラシーが同一の概念であるかどうかは、別途議論を要するところであるが、少なくとも教育界においては通常、概ね同一のものと考えられている。本項では、わが国における情報利用教育についての概略をつかむため、初等中等教育において情報活用能力がどのように捉えられてきたのか、教育行政の動向に基づいて見ておく。

周知のように、わが国で教育界に情報活用能力という言葉が登場したのは、1986年の臨時教育審議会の第二次答申である。それが“包括的かつ具体的に”規定されたのは、情報化協力者会議から教育課程審議会に提出された「情報化社会に対応する初等中等教育の教育内容の在り方」においてである[31]。すなわち、情報活用能力とは、次の四つから構成されているとされた。

(1) 情報の判断、選択、整理、処理能力及び新たな情報の創造、伝達能力
(2) 情報化社会の特質、情報化の社会や人間に対する影響の理解
(3) 情報の重要性の認識、情報に対する責任感
(4) 情報科学の基礎及び情報手段(特にコンピュータ)の特徴の理解、基本的な操作能力の習得

上に挙げたような能力を育成するにあたって、わが国では、様々な施策が進められた。教育内容については、89年に出された学習指導要領において、各教科目等の内容と情報活用能力との関係が具体化された。これを契機にして、わが国ではいわゆる情報教育(情報リテラシー教育)が本格化したということができる。

しかし、定義の策定から約10年が経ち、その間に学校においても情報基盤整備が進み、多様な情報リテラシー教育の可能性が広がった。これに伴い新たな展開があった。すなわち、97年10月に「体系的な情報教育の実施に向けて」という報告が提出されたのである。

この報告で、情報活用能力を次のように焦点化し、情報教育の目標として位置づけることとしている[32]。

(1) 課題や目的に応じて情報手段を適切に活用することを含めて、必要な情報を主体的に収集・判断・表現・処理・創造し、受け手の状況などを踏まえて発信・伝達できる能力(以下、「情報活用の実践力」と略称する。)
(2) 情報活用の基礎となる情報手段の特性の理解と、情報を適切に扱ったり、自らの情報活用を評価・改善するための基礎的な理論や方法の理解(以下、「情報の科学的な理解」と略称する。)
(3) 社会生活の中で情報や情報技術が果たしている役割や及ぼしている影響を理解し、情報モラルの必要性や情報に対する責任について考え、望ましい情報社会の創造に参画しようとする態度(以下、「情報社会に参画する態度」と略称する。)

この新しい「情報活用能力(情報リテラシー)」観を踏まえ、現在、次期学習指導要領に向けた検討が進んでいる[33]。学校教育において、情報活用能力の育成はいっそう中心的な事柄になると予想される。

#### 3.2.3 図書館における取り組み：協会等によるガイドライン策定の動向

情報リテラシーの育成(情報利用教育)という問題に対して、図書館界(図書館情報学)でも積極的な取り組みがなされてきている。すなわち、先に触れたように、利用者教育の分野で論議、実践が進められてきた。1980年代の利用者教育研究・実践を一言で総括するならば、「単なる〈図書館〉の利用だけでなく、図書館を超えた〈情報〉の利用までを視野に入れた指導が必要である」というシフトであると表現できる。すなわち、(個々の)「図書館」の使い方(例えば目録・索引など)にとどまらず、より広く「情報」の「使い方」を指導する、という理念の発達は、80年代の利用者教育の最たる特徴であるということができる。

Reference Services Review には毎年、1年間の利用者教育関連の研究文献リストが掲載される[34]。それによれば、毎年、百数十から二百数十の論文が発表されている。それらのレビューは、他に譲り、本稿では、そういった利用者教育の研究・実践の成果の一つとして、図書館関係団体等のガイドライン等策定の動向について取り上げておく。

わが国の場合は、欧米ほどではないにしろ、教育機関に属する図書館を中心に、利用者の情報リテラシー習得に対する支援に図書館が関わっていくという意識が芽生えている。図書館界全体としても、研究・実践の成果を反映させ、図書館の利用技能の指導を中心としたガイドライン策定の動きが盛んになっている。現在、このようなガイドラインとしては、(1)全国学校図書館協議会(全国 SLA)によるもの、(2)国立国会図書館(NDL)の研究班によるもの、(3)日本図書館協会(JLA)図書館利用教育委員会によるものなどがある。

(1)全国 SLA によるもの[35]は、指導内容を体系表にまとめたもので、名称はガイドラインではないが、目的や機能には共通するものが多いと思われる。82年の「自学能力を高める学校図書館の利用」の全面改訂版であり、名称の変化にも表れているように、"図書館を中心に行われる教育活動は、まさにこうした情報処理能力[筆者注：情報リテラシーとほぼ同義で用いられている]の育成をめざすもの"という考えが反映されている[36]。

(2)NDL の研究班によるもの[37]は、主に公共図書館における「利用ガイダンス」のガイドラインをまとめたものである。公共図書館における不特定多数の利用者に対して、どのような支援が行えるか、という調査研究[38]に基づくものであり、「ガイダンス」という用語・概念にその考え方が表れている。綿密な調査の裏づけがあり、図書館の現状に則した具体的な内容が評価できる。

(3)JLA 図書館利用教育委員会によるものは、まだ正式版が公表されていないが、これまでの経過から見るかぎり[39][40]、館種を超えたトータルな視点を重視しているところに特色がある。すなわち、館種ごとではなく、図書館界全体の問題として協力連携体制も築きながら利用教育にあたろうという従来なかった意図が伺える。しかし、大きな目的だけに現状とのギャップも大きく、実行段階では制度的な側面などの課題も多いと予想される。

諸外国でも、様々な協会団体等が積極的な取り組みをなしている。ここでは、米国における(1)大学研究図書館協会(ACRL)によるもの、(2)米国学校図書館員協会(AASL)によるものについて触れる。

(1)ACRL の Instruction Section は、96年にガイドラインを公表した[41]。これは、77年のガイドライン[42]を改訂したものであるが、タイトルを比較すれば判るように、「bibliographic　instruction」が単なる

「instruction」になっていることに注目したい。すなわち、上述のように、図書館利用能力を指導する図書館利用教育(bibliographic instruction)から、より広い(情報)教育の一環として図書館における利用教育を位置づけようという意図の表れであると理解できる。

(2)AASLによるものは、『インフォメーション・パワー』[43]という学校図書館メディアプログラム全体のガイドラインの中で、利用指導に関する方針・指針が詳しく示されている。今となってはやや古いため、今日のネットワーク環境には必ずしも対応しないかもしれないが、80年代の研究成果を反映し「情報」教育が重視されている点は、今日でも充分評価できる。

以上のような各種団体等によるガイドライン等をめぐる動向を総括するならば、図書館として情報リテラシー教育への貢献を意図してはいるものの、取り組みは始まったばかりで、目標や内容、方法、図書館相互あるいは他機関との機能分担、指導者の素養等について、充分な検討、合意はできておらず、試行錯誤の状態であるということができる。すなわち、図書館界が現在のような情報利用環境下にあって、図書館利用者に対する「指導サービス」をどうすべきか、あるいは、そもそも情報利用教育において図書館ないしは図書館員が果たす機能はどうあるべきか、といった意思統一がはっきりなされているとはいえない。その理由としては、結論を出すのに充分な実践および研究がなされていないことを指摘できよう。

付記するならば、この傾向は図書館(図書館員)以外でも同様であり、3.2.2で見たように、学校教育において、現在の指導要領で取り入れられた(さらに、次期指導要領でいっそう重視されるであろう)「情報教育」に関して、教員等学校関係者の間で誰がどのような責任を果たしつつ教育体制を構築していくか、という実践レベルでの共通理解、およびそれを裏づける研究成果は必ずしも得られているとはいえない。ただし、各学校で様々な実験的取り組みや研究活動は盛んになされていることも確かである。

### 3.3　情報利用教育研究の問題点と課題

このように、情報リテラシー習得の支援体制(指導・育成システム)の整備は遅れており、その背景には、研究上の問題点や課題が存在していると考えられる。本節では、それらについてまとめてみたい。教育学などの関連領域にも配慮はするが、図書館情報学(利用者教育)を中心に扱うことにする。

利用者教育における情報リテラシー教育への貢献という問題については、これまでいろいろな指摘がなされてきた。多くは館種や利用者層を限定した議論であるが、以下に挙げるようにそれらの中には館種を超えた普遍的な指摘もある。なお、本稿は情報利用教育研究の問題点を記述し、今後の「情報利用学」における課題材料を明らかにすることに主目的があるので、必ずしも同じレベルでないものも単純に並べてある。

(1)情報リテラシーの定義と構造化

まず、「情報リテラシー」とはそもそも何であるのか、という点の理解がまちまちである。3.2.1および3.2.2で見たように、情報リテラシー自体は抽象的な概念であり、実際に指導・育成する際には、教育可能なレベルまで具体化しなければならない。すなわち、教育内容あるいは目標としての情報リテラシーとは何であるか、という根本的な問いに対して、理論的な考察、あるいは実践の積み上げにより、解答を見い出していく努力は依然として必要である。

これは、情報リテラシー自体の定義の曖昧さの問題といってもよい。3.2.1で挙げた定義からも判るように、情報リテラシーは非常に抽象的かつ包括的な概念であり、しかも社会的文脈、時代的背景等によって変容するものである。よって、情報リテラシーを総体で捉えるには、何らかのレベルでの枠組み設定が不可欠であることを定義自身が既に内包している。つまり、情報リテラシーの実体を捉えるために枠組みを決めた際に、その枠組みから外れたところは排除されてしまうのである。今後の課題は、その排除された部分をどう活かしていくか、つまり、異なった枠組みで捉えた情報リテラシーをどう統合化して、その総体を構造的につかんでいくか、ということになろう。こうした動きはないわけではないが、今のところ、異なった枠組みのものをまとめるには、極論すれば、抽象的に過ぎてしまうか、非体系的な集合として記述するか、ということになっている。

(2)関連概念の整理

図書館情報学以外にも、教育学、社会学などにおいて、同類の概念が提起され、議論されている。例えば、社会学などでしばしば用いられるメディアリテラシーという概念がある。しかし、情報リテラシーとメディアリテラシーはどのように違う(あるいは同じ)概念であるか、という点に対する考え方は、図書館情報学と社会学分野では食い違いがある。しかし、概念的定義でみれば、両者には共通部分も多く、お互いの研究成

果を援用できると思われる側面も少なくない。それを可能にするには、こうした概念の整理を試みることが必要であろう。図書館利用能力(図書館リテラシー)、情報リテラシー、コンピュータリテラシー、メディアリテラシー、ディジタルリテラシーなど、関連概念(用語)はたくさん提案されている。また、これらを整理しようという試みもなされているが[44]、まだ同意が得られるような整理はなされていない。しかし、今後こうした整理が進めば、情報リテラシーをいくつかの概念の複合体として捉えるなど、様々な可能性が広がることは予想される。(2)は(1)とも関連する問題であるといえよう。

(3)育成にあたっての機能分担

これはある意味では、(1)の裏返しの問題といえるかもしれない。総体としての情報リテラシーは非常に大きなものであり、一つの図書館(あるいは、学校、授業など)だけで完結して指導できるものではない。当然ながら、他の図書館等と連携をとって指導体制を整備する必要がある。すなわち、一つの図書館でみれば、例えば学校図書館と教員(授業)との連携が、図書館どうしでみれば、例えば学校図書館と公共図書館との連携が必要であり、これらはいわば「ヨコ」のつながりと呼べるだろう。また、図書館界全体でみれば、例えば小学校図書館、中学校図書館、高校図書館、大学図書館という、一人の利用者が時間を追って利用する図書館どうしの、「タテ」のつながりが必要である。しかし現在、これらのつながりは充分とはいえず、特に「タテ」のつながりの不足は、次の(4)の問題点にも大きく関連する。

こうした状況に対して、福永[45]は、学校図書館における利用者教育の今後の研究課題として、“学校図書館における利用者教育の責任と限界が不明確である”ことを挙げている。このことは、もちろん学校図書館に限ったことではない。ほぼすべての図書館において、その利用者教育に果たす責任と限界(すなわち機能分担)は不明確なままである。

確かに、各館種相互の理想的な結びつきのあり方は摘されている。すなわち、“理論的には教育内容を体系化し、(1)基礎の部分は義務教育段階の学校教育の中で全国民に指導する、(2)その上で各館種に特有な機能や資料に関する指導内容は学校図書館、大学図書館、専門図書館と積みあげる形で図書館サービスとして指導をする、(3)公共図書館はそれぞれの館種に属する利用者と、それ等のどこにも属さない利用者の両方が対象であるから、他の館種での指導を補完する形で全指導内容を必要に応じて取り入れながら指導をする、ということ”[46]が理想的には提言されている。しかし、特にわが国では、学校教育に占める図書館の位置づけの軽さや、そもそも図書館員の配備の遅れなどの理由から、現実的には“それぞれの館種で前段階の指導が行われていることは期待せず、利用者の情報利用技術がどのレベルであるかを判断し、それに合わせて基礎のレベルの指導も取り入れながらプログラムを組むこと”[46]が必要である、という状況にある。

ただし、仮に理想的な状況にあったとしても、必ずしも連携がうまくいくとは限らない。理論的なレベルで、図書館相互、図書館と他組織間の有機的な結びつきがどのようにあるべきか、という議論は必ずしも充分ではないからである。これは、具体的には、(4)の問題にも関わってこよう。

(4)体系化と発達段階への対応

(3)の問題は、教育内容・目標としての情報リテラシーの体系化の遅れを反映している(この場合は、(1)で述べた情報リテラシーの総体の構造化とは違い、ある枠組み設定の中での情報リテラシーについての体系化を意味している)。構造的に教育内容・目標を捉え、どの部分をどの機関が担当し、どのように相互の連携をとっていくか、という問題が、図書館相互でとれていないのは、この指導にあたっての体系化の遅れも大きな要因の一つである。

さらに、特に学校教育にあっては、体系的指導にあたって、発達段階に応じた教育プログラムの開発が重要である。しかし、情報については、他の教科目と比べ、そうした側面での研究には遅れがある。様々な考察が進められているが、まだそれらを総体的に捉えた理論はなく、必ずしも実践に充分に活かせる段階ではない。

(5)指導者の育成

(1)～(4)の問題が、部分的にではあるが、解決に向けた取り組みが進められている中で、実践に移す段階で問題化しているのが、充分な素養、知識・技術を持った指導者の不足である。これは、図書館員(教員)養成のプログラムの問題にも関連する重要な課題であるということができる。

## 4　「情報利用学」の構築に向けて

### 4.1　「情報利用学」の位置づけと意義：学術情報センターを例に

本稿の冒頭で述べたように、1997年度から学術情報センターに、わが国初の研究部門である「情報利用学研究部門」が設置された。ここでは、学術情報センターにおける情報利用学研究部門の位置づけと、期待される研究成果・効果について考察し、「情報利用学」の意義を検討する。

学術情報センターは、総合目録システムであるNACSIS-CAT/ILLの提供を通じて、わが国の大学図書館をはじめ、海外の大学図書館等、国内の県立図書館等を対象に、学術情報を扱う書誌ユーティリティとして機能している。また一方、NACSIS-IR、NACSIS-ELSといった各種データベースシステムの提供を通して、主に内外の研究者らを対象に、各種情報検索サービス等を提供する一種のデータベースプロバイダ・ディストリビュータとしても機能している。

このようなサービスを提供する一方で、サービスに関わる研究開発も行なっている。研究開発を行う研究開発部は現在、学術情報研究系、システム研究系、研究動向調査研究系という三つの系からなっている。このうち情報利用学研究部門が設置された学術情報研究系は次のような研究開発等を行なっている[47]。

> 広く学術情報システムの構築・形成の在り方から始まって、目録所在情報の標準化、大学図書館のハウスキーピングの機械化、全文データベースの構築と利用を前提とした電子図書館、電子出版などの研究開発、数値・画像等のファクト情報、抄録・索引等の二次情報などの各種学術情報データベースの構築・管理手法の研究開発、キーワードの自動抽出法等情報検索の自動化技術、データベース形成の効率化、評価、品質管理等の研究開発、学術情報システムの最適利用に関する研究開発、更に、国際サービス、特に東アジア地域におけるサービスに必須な各文字種表記に対応するシステムの研究開発等を行なっている。

学術情報センターの使命の一つとして、「学術情報流通・コミュニケーションにおいて、大学を中心とする研究者らの情報要求に的確に応えること」を挙げるとすれば、それは、直接的にはNACSIS-CAT/ILLやNACSIS-IR、NACSIS-ELSなどのシステムを通して達成されるものである。しかし、これらのシステムを利用者に最適化したかたちで開発、展開し、また、各種マニュアル、講習会などの研修・教育活動などを提供していくにあたっては、利用者のおかれている状況、情報ニーズ、およびそれらを満たすために取られている行動、結果に対する満足などについて、その研究方法論も含め総合的に研究していくことが必要である。情報に関わる研究機関の一つである学術情報センターには、そうした研究をとおして、情報利用をめぐる研究の基盤を構築する役割が求められていくだろう。

実際に学術情報センターでは利用者および利用に関する調査を行なってきたが、基本的にそれらは、センターで提供しているサービスの「現利用者」を調査対象の中心とするものである。もちろん、そうした調査も重要であり、明らかになる事柄も多く、サービスの改善、開発に活かされてきた。しかし、学術情報センターは、学術コミュニティ全体に対して開かれているべきであるとするならば、その中には、センターが提供している諸サービスの利用頻度の少ない利用者、さらに未利用者、非利用者までが含まれる。ここで情報利用研究として問題とすべき点は、学術コミュニティの成員がどのように各自の情報世界を構築しているのか、独自の情報世界を持つ個々の成員がどのように共存しているのかということである。つまり、非利用者に関していえば学術情報センターのサービスを利用しない人々がどのような情報利用行動をとっているのか、なぜセンターのサービスを利用しないのかなどの点が問題となる。そして、この点に関する研究成果をサービスの改善に還元するならば、これら具体的な結果に基づいて新しいサービスの可能性を模索するということになろう。したがって、より調査対象を広げ、かつ適切な調査方法を用い、適切な分析、評価枠組みを設定する必要がある。すなわち、「情報利用」を体系的、構造的、総合的に捉えた上で行う調査でなくてはならない。この点において、情報利用行動研究、情報利用教育研究を中心とした情報利用学の視点が有効である。

### 4.2　「情報利用行動研究」「情報利用教育研究」から「情報利用学」へ

情報利用環境の変化による情報利用行動の多様化、求められる情報技能・知識(情報リテラシー)の高度化という現象に対し、現在、実践的取り組みは遅れている面があり、その背景には研究の不充分さがある。こ

れに対応するためには、情報利用行動研究においては、旧来からある批判に耐えうる方法論の確立と、現在および今後の情報利用環境に対応した理論的枠組みの確立、情報利用教育研究においては、関連学問領域まで視野に入れた理論的な概念整理と、プログラム開発など応用的な観点からの指導機能分担とそのための体系化が大きな課題である。これらの課題の解決には、研究を進める上で共有できる部分の多い両研究を有機的に結びつけ、相互に補完できるようにすることが有効である。これを「情報利用学」と位置づけることができる。

従来の情報利用行動研究は、2.1で見たように一定の成果を挙げてきたが、2.2で述べたように現在のような電子化・ネットワーク化された情報環境を必ずしも想定していない。そのため、今後において汎用的に適用可能な理論的枠組みを提供できるとは限らない。今後の情報利用学研究においては、旧来のモデルに全面的に基づくのではなく、現在の情報環境下における情報利用行動について、これまでの批判を踏まえつつ、改めて質的、量的な実態調査を行い、実証的な見地から検討を加え、新たなモデルの構築をめざすとともに、情報利用に必要な技能を情報リテラシーという概念から捉え、利用者が情報リテラシーを習得する過程、およびその支援体制までを含めた総合的な観点から考察を行うことが求められている。そのためには、旧来からある研究成果を統合的、俯瞰的な視点から再構築することが必要である。

情報利用学研究が進展すれば、次のような種々の効果も期待できる。すなわち、情報利用学の研究から得られる成果により、電子情報環境下における情報利用行動の新たなモデルの構築が期待でき、関連学問領域に対して学術的な貢献ができる。加えて、新しいモデルをもとに、様々な活動にとって効果的・効率的な電子情報資源の組織、提供の在り方、および利用支援体制の在り方について提言でき、電子情報関連分野における開発活動への示唆も少なくない。将来のわが国の学術研究活動にとっての波及的な意義も高いといえよう。さらに、電子環境および情報リテラシーという、具体的なかたちを持たない対象についての研究手法はまだ充分に検討されているとはいえず、情報利用学研究の遂行により、従来の方法論上の問題点を洗い出し、その具体的な解決法を示唆することも期待される。

今後において、情報利用行動研究と情報利用教育研究の有機的、相互作用的な結びつきを考慮した、新しい視点に立った(あるいはこれまでの視点を統合した)情報利用学の必要性は非常に高いということができる。なお、情報利用学は、情報利用行動、情報利用教育の研究を中心的な領域としながらも、他にも重要な領域を持ちうるが、本稿では言及できなかった。これについては、別の機会において論じたい。

## 5　おわりに

本稿は、情報関連の研究の中でも、現在および将来において課題性の高い領域の一つである「情報利用学」に関して、適切な議論を構築するための予備的な考察を行なったものである。「情報利用学」として今後構築していくべき新しい学問分野、あるいはその確立のために解決していくべき課題とその方向性はある程度指摘することができた。しかし反面、充分に掘り下げた検討ができなかった面も否めない。今後の課題としたい。

また、本稿では、考察の足掛かりとして主に図書館情報学に検討材料を求めたため、他領域に関する検討が不充分となった。これも今後の課題としたい。

今後は、これらの課題も視野に入れつつ、本稿を足掛かりにして、研究を進展させていくことになる。最後に、当面の計画として想定している事柄を挙げておく。一つは、実証的なデータを取り入れるため、比較的規模の大きい情報利用行動の調査である。もう一つは、情報利用教育プログラムに関する調査、およびそれをもとにしたモデルプランの開発である。

## 謝辞

本稿の執筆方針については、内藤衛亮先生(学術情報センター研究開発部学術情報研究系研究主幹・教授)に貴重なアドバイスをいただきました。この場を借りて、深く感謝の意を表します。

### 注・参考文献

[1] 本稿で述べられていることは、個人の見解であり、所属する機関、および言及した機関・組織の見解を述べたものではない。

[2] 本稿では「利用者」といった場合には原則として潜在的利用者も含む。潜在的利用者をどう定義するかについては論義があるが、ここでは未利用者、非利用者一般を指す言葉として用いる。

[3] 日本図書館学会用語編集委員会編，「図書館情報学用語辞典」，丸善，1997．ただし見出しは「利用者調査」(英語は「user study; user sur-

vey」)となっている。

[4] Menzel,H., "Information Needs and Uses in Science and Technology", *Annual Review of Information Science and Technology*, Vol.1, pp.41-69, 1966.

[5] Hewins,E.T., "Information Need and Use Studies", *Annual Review of Information Science and Technology*, Vol.25, pp.145-172, 1990.

[6] Line,M.B., "The Information Uses and Needs of Social Scientists: An Overview of INFROSS", *Aslib Proceedings*, Vol.23, pp.412-434, 1971.

[7] American Psychological Association, *Report of the Project on Scientific Information Exchange in Psychology*, Washington, D. C., 3 Vols., 1963-1969.

[8] Dervin,B.; Nilan,M., "Information Needs and Uses", *Annual Review of Information Science and Technology*, Vol.21, pp.4-33, 1986.

[9] Wilson,T.D., "On User Studies and Information Needs", *Journal of Documentation*, Vol.37, pp.3-15, 1981.

[10] Dervin,B., "Communication Gaps and Inequities: Moving toward a Reconceptualization", *Progress in Communication Sciences*, Vol.2, pp.73-112, 1980.

[11] Belkin,N.J. et al., "ASK for Information Retrieval", *Journal of Documentation*, Vol.38, pp.61-71, 145-164, 1982.

[12] 田村俊作,「訳者解説:利用者研究とその領域」, Varlejs,Jana ed.,(池谷のぞみ他訳)情報の要求と探索, 勁草書房, 1993, pp.122-157.

[13] Wilson,T.D., "Information Behaviour: An Interdisciplinary Perspective", *Information Processing and Management*, Vol.33, No.4, pp.551-572, 1997.

[14] Mey,Marc de,(村上陽一郎他訳)「認知科学とパラダイム論」, 産業図書, 1990.

[15] Fidel,Raya, "Qualitative Methods in Information Retrieval Research", *Library and Information Science Research*, Vol.15, No.3, pp.219-247, 1993.

[16] Wilson,T.D., "On User Studies and Information Needs," *Journal of Documentation*, Vol.37, pp.3-15, 1981.

[17] Dervin,Brenda, "Useful Theory for Librarianship: Communication, Not Information", *Drexel Library Quarterly*, Vol.13, No.3, pp.16-32, 1977.

[18] Ellis,David, "A Behavioural Approach to Information Retrieval System Design", *Journal of Documentation*, Vol.45, pp.171-212, 1989.

[19] Dowler,Lawrence ed., *Gateways to Knowledge: The Role of Academic Libraries in Teaching, Learning, and Research*, MIT Press, 1997.

[20] 大城善盛, 鍛治宏介,「わが国のインターネットOAPCの現状」, 図書館学会年報, Vol.43, No. 3, pp.103-116, 1997.

[21] 林賢紀,「日本国内OPACリスト」, <http://ss.cc.affrc.go.jp/ric/OPAC/>.

[22] 矢崎省三, 豊田裕昭,「東京農工大学附属図書館におけるインターネット:WWWホームページの制作と利用教育, 大学図書館研究, No.48, pp.36-41, 1996.

[23] 野末俊比古,「ネットワーク環境における利用教育の可能性」, 第45回日本図書館学会研究大会発表要綱, 日本図書館学会, 京都, 1997, pp.37-40.

[24] 越塚美加,「情報利用行動調査の一技法としての具体例叙述法」, 図書館学会年報, Vol.39, No.1, pp.1-12, 1993.

[25] 越塚美加,「文献のブラウジングが研究過程に与える影響」, 学術情報センター紀要, No.8, pp.131-142, 1996.

[26] 館種によって、あるいは文脈によって、利用教育、利用指導、利用ガイダンス、利用案内など異なる用語が使われ、また、概念も異なることがある。ここでは、館種にかかわらず図書館が(他の組織等と協力しながら)行う利用者の図書館・情報利用を支援する指導・案内的な活動(サービス)、およびそれに関わる研究領域を包括的に表す言葉として「利用者教育」という用語を用いる。

[27] 日本図書館学会研究委員会編,「図書館におけ

る利用者教育：理論と実際」(論集・図書館学の歩み　第14集)，日外アソシエーツ，1994.

[28] 福永智子，「米国の学校図書館における利用者教育の理論化：Carol C. Kuhlthauを中心に」，図書館における利用者教育：理論と実際(論集・図書館学の歩み　第14集)，日本図書館学会研究委員会編，日外アソシエーツ，1994，pp.137-152.

[29] American Library Association Presidential Committee on Information Literacy, *Final Report*, Chicago, ALA, 1989.

[30] Doyle,C., *Summary of Findings: Outcome Measures for Information Literacy within the National Education Goals of 1990* (Final Report to National Forum on Information Literacy), 1992.

[31] 文部省，「情報教育に関する手引」，第3版，ぎょうせい，1993.

[32] 情報化の進展に対応した初等中等教育における情報教育の推進等に関する調査研究協力者会議，「体系的な情報教育の実施に向けて：第一次報告」，文部省，1997.

[33] 教育課程審議会，「教育課程の基準の改善の基本方向について(中間まとめ)」，文部省，1997.

[34] Rader,Hannelore B., "Library Instruction and Information Literacy--1995", *Reference Services Review*, Vol.24, No.4, pp.77-96, 1996.

[35] 全国学校図書館協議会，「『資料・情報を活用する学び方の指導』体系表：『学校図書館利用指導』の内容と展開」，学校図書館速報版，No.1359，p.5，1992.

[36] 笠原良郎，「『資料・情報を活用する学び方の指導』体系表をまとめて」，学校図書館，No.501，pp.9-13，1992.

[37] 国立国会図書館図書館研究所「不特定多数を対象とする図書館における利用者ガイダンスのあり方」研究班，「公共図書館利用者ガイダンス・ガイドラインとその考え方」，現代の図書館，Vol.34，No.4，pp.212-216，1996.

[38] 不特定多数を対象とする図書館における利用者ガイダンスのあり方研究班，「利用者ガイダンスの視点：公共図書館等を中心に」，図書館研究シリーズ，No.31，1994.

[39] 日本図書館協会図書館利用教育委員会，「図書館利用教育を全学生の必修に！ カリキュラムに組み込んでいくための実績づくり：図書館利用教育ガイドライン(大学版)第2次案」，図書館雑誌，Vol.90，No.6，pp.408-411，1996.

[40] 日本図書館協会図書館利用教育委員会，「学校図書館を情報教育の拠点に：図書館利用教育ガイドライン(学校図書館:高等学校版)案」，図書館雑誌，Vol.90，No.10，pp.796-799，1996.

[41] American Library Association/Association of College and Research Libraries/Instruction Section, "Guidelines for Instruction Programs in Academic Libraries", *College and Research Libraries*, Vol.58, pp.264-266, 1997.

[42] American Library Association/Association of College and Research Libraries/Bibliographic Instruction Task Force, "Guidelines for Bibliographic Instruction in Academic Libraries", *College and Research Libraries News*, Vol.38, pp.92, 1977.

[43] American Association of School Librarians; Association for Educational Communication and Technology eds.,(全国学校図書館協議会海外資料委員会訳)「インフォメーション・パワー：学校図書館メディア・プログラムのガイドライン」，全国学校図書館協議会，1989.

[44] McClure,Charles R., "Network Literacy: A Role for Libraries?", *Information Technology and Libraries*, Vol.13, No.2, pp.115- 125, 1994.

[45] 福永智子，「学校図書館における新しい利用者教育の方法：米国での制度的・理論的展開」，図書館学会年報，Vol.39，No.2，pp.55-69，1993.

[46] 丸本郁子，「図書館サービスとしての利用者教育の意義」，図書館における利用者教育：理論と実際(論集・図書館学の歩み 第14集)，日本図書館学会研究委員会編，東京，日外アソシエーツ，1994，pp.7-30.

[47] 学術情報センター，「学術情報センター要覧 平成9年度」，学術情報センター，1997.

学術情報センター紀要　第10号(1998)

研究論文

# 言語における共時性と通時性

# Sinchrony and Diachrony in Language

学術情報センター　影浦　峡
Kyo KAGEURA
National Center for Science Information Systems

**要旨**

言語における「共時性」と「通時性」の概念の基本的な枠組みは、今世紀前半にソシュールによって示唆された。けれども、それ以来、これらの概念を巡っては、様々な解釈が与えられてきた。本稿では、それらの解釈の代表的なものを参照しつつ、具体的・技術的な言語研究を健全なかたちで進めるために必要な、「共時性」と「通時性」という概念の枠組みを、特に専門用語研究のあり方を想定しながら、整理する。

**ABSTRACT**

The conceptual framework of 'synchrony' and 'diachrony' in language was indicated by Saussure, at the first half of the 20th century. Since then, various interpretations have been made of these concepts. In the article the author tries to clarify the diversity of conceptual spectrum implied by these two concepts, in order to put the research in terminology in a proper theoretical setting.

[キーワード]　共時性、通時性、言語体系、言語運用、均質空間
[Keywords]　synchrony, diachrony, language system, language performance, uniform space

## 1　はじめに

専門用語研究の言語研究における位置づけを巡る議論の中で、影浦は以下のように述べている。

> …言語体系が共時的なものとして解釈されるのに対して、専門用語の体系は同じ意味では共時的ではなく、通時的な流れのなかでの一時的な体系として、通時的な観点において捉えられなくてはならない…[1]

いささか性急に述べられたこの言葉は、専門用語研究の位置づけについて誤解を招きやすいばかりでなく、極端な背景の省略ゆえに、共時性と通時性という言葉がそれぞれ一つの概念を表していることをあたかも皆が受け入れているかのような、誤った印象を暗黙に与えかねないという意味で、不正確である。

そこで、本稿では、共時性と通時性という概念設定を検討し、言語に対する認識を巡る場の構成に両概念がどのように関わっているか(あるいは逆に両概念の捉え方に見られる相違に、言語に対する認識論的付置の相違がどのように現れるか)を整理した上で、言語研究の構図におけるその役割を検討する。

## 2　共時性と通時性を巡って

### 2.1　一般的な解釈

例えば、『現代言語学辞典』では、「共時的」という語は、ソシュールの用語と断った上で、「歴史的な考察を加えず、一時期の言語(およびその部分)の状態を体系的に考察・研究する場合に用いる」用語とし、「歴史的」という意味の「通時的」という用語と対比している[2]。言語の認識おける時間というものを、我々が常識的に理解している時間と同じものと考えるならば、これは、非常に常識的で、一見して理解しやすい定義である。いずれにせよ、本稿の文脈では、これらの用語がソシュールを起源とすることは明らかであるから、以下ではソシュールの概念規定に対する解釈を一応の骨子として議論を展開していこう。

ソシュールに関する専門の辞典でも、記述がより丁寧である以外は、基本的に同じ類の解説が与えられている。多少長いが、以下に引用しよう。

ある科学の対象が価値体系(système de valeurs)として捉えられるとき、時間の軸上の一定の面における状態(ètat)を共時態と呼び、その静態的事実を、時間(temps)の作用を一応無視して記述する研究を共時的言語学という。これはあくまでも方法論上の視点であって、現実的には、体系は刻々と移り変わるばかりか、複数の体系が重なり合って共存していることを忘れてはならない。このタームはさらに厳密に特定共時態(idiosynchronie)もしくは特定共時的(idiosynchronique)と定義し直された。「共時的(＝言語(langue)の一定時期に属する)という語は少々不確かな点もある。同時的なものはすべて同じ秩序を構成するように思わせかねない。特定共時的(＝特定の一言語に対応する独自の秩序における)とつけ加えるべきである…。これに対して、時代の移り変わるさまざまな段階で記述された共時的断面と断面を比較し、体系総体の変化を辿ろうとする研究が、通時言語学(linguistique diachronique)であり、そこで対象とされる価値の変動(déplacement)が通時態である。「定義：通時的次元とは諸価値の変動のことであり、それは有意単位の変動ということにほかならない」・・・。「定義：特定共時的次元とは、刻々と樹立される形で現わされた、諸価値の特定の均衡のことである。これは通時的次元と性格を異にする。通時的次元と特定共時的次元との対立は、動態と静態の対立である」…[3]。

さらに、「ソシュール研究」における世界的な権威といわれる丸山圭三郎の『ソシュールの思想』における、共時性と通時性の解釈も、より綿密な関連概念の議論を含むとは言え、その骨子は、概ね上記の引用に現わされている[4]。また、カラー[5]やムーナン[6]も共時性・通時性という概念に関して、本質的には丸山と同様の立場を取っているように思われる。

それでは、これらに対して、ソシュール自体はどうであったであろうか？彼の『一般言語学講義』では、共時性と通時性という言葉について、以下のようないささか奇妙な述べ方がされている[7]。

**共時言語学**　は、共存し・かつ体系を形づくる諸辞項をむすぶところの論理的および心理的関係を、同一の集団意識によって知覚されるままに取り扱う

**通時言語学**　は、これに反して、同一の集団意識によって知覚されず・かつたがいのあいだに体系を形づくることなくつぎつぎと**置きかわる**継起的辞項をむすぶところの関係を、研究する

これらの規定を行う前に、ソシュールは、当時の歴史的言語研究の状況に対する考察を加え、言語研究における時間の位置づけを検討している。けれども、ソシュール自身が自ら共時性と通時性という概念を規定しようと試みていると取れる部分では、実は「時間」という言葉はあまり使われないのである。

### 2.2　「時間」と「方法論上の視点」を巡って

上で引用した『現代言語学辞典』と『ソシュール小事典』において、共通しているのは、ともに、共時性と通時性を説明するために、「時間」ないしは「時代」という概念に訴えている点である。しかしながら、我々が常識的に受け入れている時空の認識論的配置の中では、当然、「共時性」を抽象化できる現実的な契機はどこにもない。それゆえ、『ソシュール小事典』では、共時性の概念が、「方法論上の視点」として、すなわち、いわば虚構として導入されたものであることを強調しているのであろう。

もちろん、「共時性」という概念について言うならば、それが方法論上の虚構であることは、ほとんどその定義によって十分明らかである。けれども、問題となっているのは、そうでなはい。問題は、むしろ、「共時性」と「通時性」の概念的対立が方法論上の虚構であるということが、一体、言語を巡る認識論的配置の中で何を意味するか、ということであるように思われる。ここにおいて、上記の定義における「時間」(ないしは時代)という言葉の意味するところが問題となってくる。

上述の定義における「時間」の意味そのものを理解することは容易である。すなわち、デカルト以降の均質なユークリッド的時空における時間軸を想像すればよい。その中での、特定共時態の方法論的抽象と、「歴史的」に認識される言語の通時的変遷という概念規定が、「共時性」と「通時性」の本質なのであろうか。もしそうだとするならば、いくつかの疑問がわいてくる。

第一に、ソシュールにおいては、徹頭徹尾、共時態が通時態に先行することが強調されていること、第二に、ソシュール自身は、共時性と通時性の定義において、時間という言葉をほとんど使っていないこと、である。ここでは、別にソシュールの文献考証を展開するわけではないが、それにしても、これら二点は、以下のような論理的な問題を示唆しているように思われ

る。

すなわち、共時性と通時性との対立は、デカルト的時間の配置の中で言語研究を進めるために導入された方法論上の虚構に過ぎないのではなく、むしろ、デカルト的時間とは全く関係ないかたちで、言語を巡る全体的な認識論的配置を構成する一次的要因として導入されたものではないか、ということである。誤解を覚悟で簡潔に言うならば、共時態とは、それなしでは言語に関する認識がそもそも成り立たないから、理論的に存在を仮定せざるを得ないものであり、デカルト的時間配置とは全く関係がない。第二に、通時態とは、共時態の前後関係として、すなわち順序関係として形式的に認定されるものであり、従って、これも、デカルト的時間配置とは全く関係がない。そして、このように設定される共時性と通時性によって規定される認識論的配置こそが、言語を言語たらしめているものなのだ、ということである。

言葉を変えよう。言語において、そもそも「時間」というものが存在するならば、それは共時態の系列としての順序として存在するのであり、それは我々が現在常識的に理解しているような、デカルト的時間とは全く異なると同時に、そこから二次的に派生するものでもない。順序関係を規定する認識論的配置を、(せいぜいよくて近似的に)デカルト的時間軸に再配置したときに、『ソシュール小事典』において使われているような「時間」が誕生するのであり[8]、時間を方法論的必要性から抽象化したところに共時態が生まれるわけではないのではないか。それゆえ、両者は存立構造に関することであり、「歴史」に関することではないのではないか。

ソシュールにおいては、このように考えて、初めて、ラングとパロールの二分法が(歴史的事実がラングの分析に関与しないといった、いわば手続き的関連とは別に)、共時態と通時態の二分法と論理的に独立したものとして設定されていることの意味も理解できることになる。デカルト的時間軸を前提して考えるならば、例えば「方法論上の視点」であるところの通時態が、スケールの違いはあるにせよ、なぜパロールと論理的に区別されなくてはならないのかという点、さらにそれを介してラングと共時態が論理的に区別されなくてはならないかという点が理解できないことになる。実際のところは、両者の位相が違うのだ。ラングとパロールの対立と共時性と通時性の対立という二つの対立を別々に見ている限りでは、このことは当たり前のように理解されていると感じられるのであるが、両者の連関を改めて考慮すると、少なくとも共時性と通時性を歴史的に理解する限り、問題はそれほど簡単ではないのである。共時態と通時態の対立が上述の通りであると解釈し、そのように規定された特定共時態という概念の中に、デカルト的時間軸を導入することによって、パロールが言語事実と我々が考えるものと結びつくのである。

### 2.3　整理

以上から、共時性と通時性について、通常の解釈に加え、もう一つの解釈が可能であり、さらにその方が、ソシュールが導入した概念的諸装置の配置を考えるにあたっては都合がよいことが示唆される。整理してみると、以下のようになる。

**共時性**　言語に関する認識を成立させる認識論的契機
**通時性**　共時性の観点から規定された共時態という要素の順序関係

注意しなくてはならないのは、この両概念の分割に先だって、分割対象となる全体が存在するわけではないということである。逆に、分割することによって初めて全体が成立するような性質のものなのである。分割以前の全体を実体的に仮定することは、結局、通常の解釈のように、改めて外部的な要素に訴えて意味付けするのと同じことになってしまう[9]。

## 3　言語研究における意義

それでは、前節のように共時性と通時性とを整理することは、言語の具体的な研究において、どのような意味があるであろうか。第一に、これは厳密には意義とは言えないが、論理的にそのようでしかありえないのだからしょうがない、というものである。この意味では、言語研究の前提として明確化されなくてはならない背景を明確化したというだけであり、その意義をあえて考えるならば、言語の個別研究に外部から明確化されない神秘的要因を無意識に移入することを排除することになろう。また、同時に、無意味な議論を排除することにもなろう[10]。

第二に、このように考えて、はじめて、我々が通常理解しているような時間性・歴史性の問題を、改めて言語分析に取り込むことができるようになる。ソシュールの解釈における、共時性と通時性の二分法の視点と、ランガージュ・ラング・パロールの対立の視点との関係は、後者の概念の多義性やつかみにくさも

あって、正面からは扱われてこなかった。例えば丸山[11]は、前者をいわば研究方法論における認識の枠組みとして、後者を言語に対する認識論的な枠組みとして(しかし両者はどう違うのだろう)解釈し、それらの関係については積極的な言明を控えているように思われる。一方、カラーは、最も明白な事実的関連「通時的ないし歴史的事実がラングの分析に関与しない」[12]を述べるにとどまっている。

別に、具体的な言語の研究を進めるにあたって、こうしたソシュールの概念規定の相互関係を改めて明確にする必要はあるまい。必要なのは、こうした錯綜した概念規定の背景にあるアポリアを認識し、具体的な言語の研究を進めるにあたって、何を避けなくてはならないか、言葉を変えて言うと、恣意的に前提しなくてはならない部分に何を持ってくるか、を自覚することである。こうした自覚なしには、個別の言語研究の成果が「言語」の研究にどのようなかたちで貢献するのかについて立ち戻った議論ができないからである。

## 4　おわりに

本稿では、通時性と共時性の問題を、主にソシュールとその解釈を参照しつつ議論してきた。両概念に関する、一応の特徴付けは

**共時性**　言語に関する認識を成立させる認識論的契機

**通時性**　共時性の観点から規定された共時態という要素の順序関係

とすることができる。

では議論のポイントはどこにあるのか？議論のポイントは、むしろ非常に消極的な言明にある。すなわち、何に言及すると具体的な言語の研究が成り立たなくなるか、従って、意識的に整理しつつも、具体的な言語の研究として恣意的に前提する以上に立ち戻ることを避けなくてはいかないことは何か。

共時性と通時性を巡って本稿で論じたのは、デカルト的時間軸に沿った解釈が、言語の研究において、ほとんど無意味であり[13]、むしろ本稿が結論づけたように、共時性と通時性という概念を捉えなくてはいけないこと、それと同時に、具体的な言語研究を進めるにあたって前提とされなくてはならないレベルでは、この区別が純粋に恣意的なものであり、その恣意性を越えて立ち戻ることは言語の具体的研究を不可能にすること、さらに、にもかかわらず、この恣意性に自覚的であることのみが、言語の具体的な研究を展開可能にする要因であること、である。大まかな議論ではあるが、具体的な言語研究の主発点に置くという観点からは、少なくとも深さを装った神秘化には勝るであろう。

## 注・参考文献


[1] 影浦峡,「「語」と「専門用語」—専門用語に関する理論的研究へ向けての試論—」『学術情報センター紀要』第7号, pp. 225, 1995.
また、
Kageura, Kyo., "Toward the Theoretical Study of Terms," *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*. Vol. 2, No. 2, pp. 239-257, 1995.
も参照。

[2] 田中春美他『現代言語学辞典』成美堂, 1988, pp. 658.「synchronic《共時的》」の項、およびpp. 163.「diachronic《通時的》」の項。

[3] 丸山圭三郎編,『ソシュール小事典』大修館, 1985, pp.309-310.

[4] 丸山圭三郎,『ソシュールの思想』岩波, 1981. 352p.特にpp. 92-116.

[5] J・カラー,『ソシュール』川本茂雄訳, 岩波書店, 1992, (Culler, J. *Saussure*. London, Fontana, 1976) pp. 47-62.

[6] G・ムーナン,『ソシュール』福井芳男, 伊藤晃, 丸山圭三郎訳, 大修館, 1970, (Mounin, G. *Saussure ou le Structuraliste sans le savoir*. Paris, Editions Seghers, 1969) pp. 50-60.

[7] F・ソシュール,『一般言語学講義』小林英夫訳, 岩波書店, 1940, (Saussure, F. de. *Cours de Linguistique Generale*. Bayee & Sechehaye, 1949.)主に第3章
これに対して、いわゆる「原典」資料が色々と出ているが、ここでは文献考証を目的としているわけではないし、また、ソシュールの通時性と共時性とを巡る概念の不透明さを指摘し、本稿での問題提起の前提とするには上記の文献で十分である。

[8] こうした再配置を暗黙の前提とした言語認識は、しばしば、政治的問題として現れる。この点に関する日本語を巡る論考としては
『現代思想』Vol. 22, No. 9, 1994. 特集＜日

本語＞の現在
李妍淑『「国語」という思想：近代日本の言語認識』岩波書店，1996.
等がある。

なお、実は『ソシュール小辞典』からの引用自体は、「時代」というデカルト的な時間概念を前提としながら、共時的なものの前後関係から通時性を定義しようとしているようにも読めるという意味で、二義的である。

[9] 『ソシュールの思想』を出発点とする丸山のその後の諸論考には、こうした認識論的配置に関する倒錯した解釈がまま見られる。最も典型的なものの一つは、言語の分節化を、「身分け」と「言分け」における循環のダイナミズムとして説明している点に見られる。意味を生成し言語の存立構造を支える基盤として理論的と同時に実体的に導入された「身分け」だが、言語を巡る本来的な問題は、常に・既に言語が存在している状況で、論理的に想定し得る「身分け」の構造が、常に・既に「言分け」からの差分として、「言分け」に規定されているものでしかありえない、という点にあったのではなかったか。言語学者が意識的にあるいは半ば無意識に、言語の部分現象の記述に研究対象を限定し、存立構造論への言及を避けてきたのは、丸山が導入した「身分け」のような、一見言語の存立を基礎づける外部的根拠が、常に・既に言語によって派生したものでしかありえないという論理的事実に、具体的な言語記述の場で直面してきたからではなかったか。「身分け」が、言語の存立構造論上の説明概念として何か意味を持ち得ると考えるのは、その意味では、宗教的な満足を得られる人がいるにせよ、論理的な解決ではない。

[10] ただし、この点については、共時性・通時性よりもはるかに多く誤解され、無意味な議論が交わされている概念装置は多い。やはりソシュールに帰されることの多い、関係・実体・構造は、その典型的なものであろう。当然、この三者は論理的に言えば、どれが先行するとは言えない類のものである。それゆえ、関係の優先性を唱えることが意義を持つのは、例えば実体論の弊害が、方法論上現れてきた限りにおいてであり、それ以上のものではない。それを越えて、関係の優先性をとりあえずの出発点と置くことの具体的な方法上の利点を示すことなしに、実体から関係あるいは構造へという議論を何か重要なものであるかのように見なすのは、健全な言語研究を阻害するものであると同時に、言語哲学的にも無意味である。

[11] 丸山，『ソシュールの思想』，*op. cit.*

[12] カラー，*op. cit.*

[13] むろんこれは、歴史的な観点からの言語研究の意義を否定しているのではない。

学術情報センター紀要　第10号(1998)

研究論文

# 情報検索システムのオンライン更新における一貫性維持方式

# A Mechanism to keep Consistency on On-line Update of Information Retrieval Systems

学術情報センター　大山　敬三
Keizo OYAMA
National Center for Science Information Systems

**要旨**

情報流通のネットワーク化、グローバル化に対応するために、情報検索システムにも、サービスを中断することなくデータベースを更新して新しい情報を提供できる、オンライン更新機能が求められている。WWWのインデクスサーバ技術を適用してこれを実現しようとすると、検索結果を分割して表示する際に不整合が生ずるなど、検索集合の一貫性に問題が起こる可能性がある。また、セッションレス型のサービス形態で集合演算機能などを提供するためにも検索集合の管理方式を効率化する必要がある。本稿ではこれらの問題点を分析し、検索集合とデータベースレコードのバージョンを制御することにより解決する方法を提案している。

**ABSTRACT**

In order to cope with expanding information flow over the Internet and its globalization, quasi-immediate online update functionality which enables providing new information without stopping services is required for information retrieval systems. When adopting the index server technology used in the World Wide Web to realize it, some problems regarding consistency of result sets may arise in such a case that the result records are fetched as divided segments over some time period. At the same time, an efficient method of result set management is required in order to provide conventional information retrieval functions such as set operations in a session-less service mode. This paper analyzes those problems and presents a solution by managing versions of result sets and database records.

[キーワード]　情報検索、検索集合、集合演算、オンライン更新、イントラネット、バージョン管理
[Keywords]　information retrieval, result set, set operation, on-line update, intranet, version management

## 1　はじめに

インターネットに代表される情報通信基盤の整備によって情報システムにおけるデータ収集の効率化がはかられ、データベースの構築過程もオンライン化、リアルタイム化が進められてきている。このような動向に対し、情報検索においても即時更新の要求が強まっている[1]。

また、世界的規模でのネットワークの相互接続により、情報検索サービスの利用のグローバル化も急速に進んでおり、システムには24時間体制のサービス性が求められてきている。

情報利用者の環境も急速に変化しており、WWW(World Wide Web)技術がその中核をなすようになってきている。ここではセッションレス型、ステートレス型の通信が一般的であり、情報検索システムにおけるセッションを前提とした資源管理方式はこのような環境に適応しきれなくなってきている。

さらに、イントラネット(intranet：インターネット、特にWWWの情報技術を用いて構築される組織内ネットワーク)の普及により、経済的、社会的に重要な情報がWWWの仕組みを通して提供されるようになると、従来のインターネットとは異なって情報源に

高度の信頼性(ここでは可用性、一貫性、安定性などを指す)が求められるようになる。

ところが、既存の情報検索システムの多くではデータベース更新はオフライン、切換時はサービス停止という運用形態をとっており、新しい利用環境からの要求に十分に応えられない。

一方、WWW検索の新しい技術として登場してきたロボット型のインデクスサーバでは、インターネット上の情報サーバからロボットと呼ばれるプログラムで自動的に収集した情報に機械的にインデクスを付与して検索機能を提供しており、即時性はあまり高くはないがオンライン更新を可能としている例もある[2][3]。しかし、これらは対象であるインターネット上の情報自体が不安定であることから、検索結果の厳密性についてのシステム上の配慮はあまりなされていない。

インデクスサーバ技術を情報検索システムのオンライン更新にそのまま適用すると、後述のように時間経過にともなってデータの再現性(通常の情報検索の評価尺度としてのrecallではない)に問題が発生し、検索結果の一貫性が崩れる。イントラネットや情報検索サービスでは特許や症例、判例などといった経済的・社会的に重要な情報も提供しており、システムの不備によりユーザに提供する情報に漏れが発生することは、たとえ頻度が低くとも看過できない問題である。

また、WWW技術をベースにしてサービスシステムを構築しようとすると、セッションレス型であるために接続時間に基づいた資源管理の有効性が薄れる。このため、インデクスサーバでは、コストが接続時間に依存するような資源の消費を削減する目的から、検索結果を(少なくとも検索エンジン内部では)レコードの集合としてではなく検索条件として保持するようになっている。

情報検索システムにおいては、これらの条件を満足し、従来からの検索機能を維持しつつオンライン更新を実現するための技術開発が求められている。本論文では、その基礎的検討として、オンライン更新時の検索集合の一貫性維持方法について述べる。

## 2　一貫性に関する問題

本章ではまず、従来型の情報検索システムとWWWのインデクスサーバシステムにおける検索集合の保持の方法について述べ、次に、これらのシステムにおいて検索サービスを提供しながらデータベースの更新を行う場合(オンライン更新)に、検索集合にどのような影響を及ぼすかを明らかにする。

### 2.1　検索集合の保持方法

従来の端末接続型の情報検索システムにおいては、セッションが継続する間は、複数の検索集合を保持して、任意の時点でその集合要素を表示したり、集合演算を実行して新しい検索集合を作成したりできる。これは、情報検索用標準プロトコルであるZ39.50接続などによるセッション型のシステムでも同様である。

システムによっては検索集合を外部記憶装置に保存し、セッションを越えて保持できるものもあるが、基本的な考え方は同様であるので、ここでは簡単のために検索集合の保持期間はセッション中に限ることにする。

検索システムの内部(検索エンジン)では通常、検索集合要素の全レコードを識別するためのハンドラをリストとして記憶することにより検索集合を保持する(レコードリスト保持方式)。ハンドラは通常、レコード番号などとするが、全文検索システムでは文書要素の開始と終了の位置情報を用いるものもある。

一方、WWWは基本的にセッションレス型でありサーバは状態を持たないため、インデクスサーバ上には検索集合は存在せず、検索と表示は一体の操作として実行される。しかし、検索結果とともに検索条件をHTMLデータに埋め込むことにより、クライアント側に状態を保持させることが可能であり、ユーザからみると疑似的に検索集合を保持しているように見える。表示ウィンドウを複数使用することにより複数の状態を保持することもできる。クッキーなどの仕組みを用いればサーバ側に状態を保持してクライアントと連携することも可能となる。これらの場合はいずれも検索集合を検索条件として保持する(検索条件保持方式)。

### 2.2　オンライン更新における検索集合

データベースは通常、レコード自体を集めたレコードファイルと検索のために作成したインデクスファイルからなる。更新には一般に、レコード単位での追加、置換、削除がある。ここで、レコードとは検索や表示において対象とするデータの単位である。従って、データベースの更新には、レコードファイルとインデクスファイルのそれぞれについて追加、置換、削除が必要となる。

検索サービスを提供しながらデータベースの更新を行うオンライン更新においては、検索時と表示時の間に検索集合に含まれるレコードに更新があると、検索集合が影響を受け一貫性に障害が発生する可能性がある。

レコードリスト保持方式の場合、集合要素は検索時に確定するので、更新前後で集合に含まれる要素レコードが変化してしまうことはないが、レコードファイルの変化の影響を受けるため、以下のような問題が発生することがある。

・検索レコードが削除されて存在しなくなる
・検索レコードが置換されて検索条件に合わなくなる

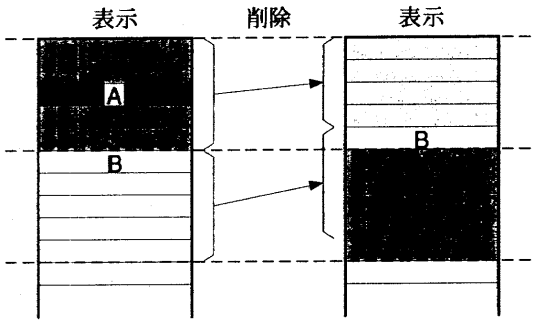
検索条件保持方式の場合は、表示時にインデクスファイルにアクセスして要素レコードを再検索するので、インデクスとレコードの間に不整合が生ずることはないが、データベースの変化に応じて集合自体が動的に変化するため、以下のような問題が発生する。

・検索結果の件数が変わる
・レコードの表示順が変わる

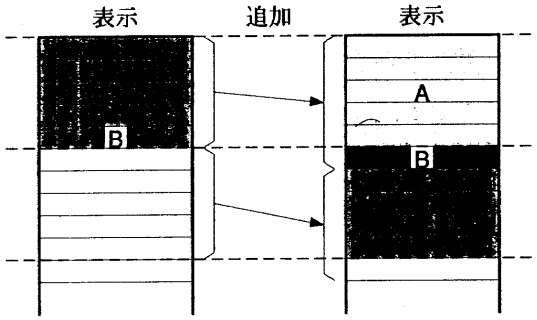
表示順の変動は、検索集合のレコード中で直接の検索対象となったデータ部分が更新された場合だけでなく、表示順をソートするためのキーとなるデータが変更された場合や、さらには tf-idf などの手法によるランキングにおいて、更新によってデータベース全体の統計情報に変化があった場合にも起こり得る。

これらの結果、検索集合を何レコードずつかに分割して表示する場合に、図1に示すように一部のレコードが表示されなかったり重複して表示されたりする可能性がある。

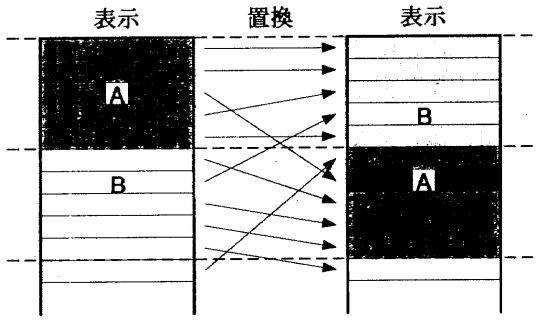
まだ表示していない部分で順序が変動してもユーザには全く影響なく、このような問題が起こる可能性が通常では無視し得るほど低い場合も多いが、厳密性を要求されるデータベースでは頻度にかかわらず重大な問題である。




(a)1-5件目を表示後、レコードAを削除し、6-10件目を表示するとレコードBは表示されない。




(b)1-5件目を表示後、レコードAを追加し、6-10件目を表示するとレコードBを重複して表示する。




(c)1-5件目を表示後、レコードAとレコードBの置換をし、ソートやランキングのために順番が入れ替わったため、6-10件目を表示するとレコードAは重複して表示され、レコードBは表示されない。

図 1　レコード順の変動と表示の問題

### 2.3　一貫性に関する対処方法

情報検索サービスでは一貫性が崩れた状態のまま情報の提供を行うことは許されない。上記の一貫性の問題に対する対応策としては、以下のような対処方法がある。

(a) 一貫性が崩れている検索集合にアクセスしようとしたときに、ユーザに通知するとともに集合を破棄または再作成する

(b) レコードファイルやインデクスファイルの構成を工夫することにより一貫性が崩れないようにする

データベースの更新の頻度が少なければ、システムのオーバヘッドが少ない(a)の方法が適当な場合もあ

ろう。データベースの更新があったときに、単純にすべての検索集合を破棄することもできるが、一貫性の破れを検索集合ごとに厳密に評価してユーザへの影響を最小限に留めることが望ましい。

しかし、更新頻度が高くなると、ユーザの負担が大きくなるため(a)の方法は採用しづらい。料金を取ってサービスしている場合は、ユーザに再検索や再表示を強いるため、一層管理が困難になる。このため(b)を実現する必要が生ずる。

ただし、検索集合を維持している期間は何らかのシステム資源を消費するため、(b)を実現する場合でも、検索集合の持続時間に制約を加える必要がある。そこで、セッションの継続時間や検索集合の有効期限というような形で(a)の方法を組み合わせて用いることになる。

(a)は(b)の補助として考える限り実現は比較的容易であるので、以下の章では(b)の実現方式について検討することにする。

## 3　一貫性の維持方法

### 3.1　一貫性の条件

前章の検討より、個々の検索集合の一貫性を維持するためには、検索結果を表示するときに検索時のレコードデータと検索集合要素の状態(集合内の順序を含む)を再現できればよいことがわかる。

#### (1) レコードデータの再現方法

データベースの更新後においても更新前のレコードデータを再現できるようにするためには、置換、削除の更新について、アクセスの可能性のある過去のバージョンのレコードデータを保持しておく必要がある。

各バージョンのデータを同一レコード中に保持するか、バージョンごとに独立のレコードとするかは、記憶資源と計算資源のトレードオフとして検討する必要があるが、いずれも実現は容易である。

#### (2) 検索集合要素の再現方法

レコードリスト保持方式の場合は記憶している集合要素が変化しないので特別な仕組みは必要ない。一方、検索条件保持方式の場合は、追加、置換、削除のすべての種類の更新について、任意のバージョンを指定して検索できるインデクスの仕組みが必要になる。

計算資源に関しては、(1) は各レコードを実際に表示するときに処理するため全体の効率にはあまり大きく影響しない。しかし、(2) は検索時に常にオーバヘッドとなるため、その実現方法が検索システムの記憶および計算資源の利用効率を大きく左右する。そこで、以下ではインデクス上でバージョン管理を行う方法を中心に検討する。

### 3.2 バージョン切換単位

データベースの更新が頻繁であると、ユーザがシステムを利用中にもバージョンは次々に新しくなる。検索対象のバージョンを最新のものに切り換えるタイミングには以下のような選択があるが、これによって、ユーザの使い勝手が幾分変わってくる。

#### (a) セッション単位

セッション接続時点の最新バージョンをセッション終了まで使用する。再接続するまでは新しい情報にアクセスできない。セッション中は同じ条件で検索すれば同じ結果が得られる。集合演算も同一バージョン上で行われる。

#### (b) 集合単位

検索実行時の最新バージョンを使用する。検索集合ごとにバージョンが変化する可能性がある。常に最新の情報を検索できるが、表示は検索集合作成時のバージョンで行われる。同一セッション中で同じ条件で検索しても結果が異なる可能性がある。集合演算は既存集合のバージョンが異なると矛盾が生ずる可能性があるので、原則的に再検索となる。

これらの選択はセッション型かセッションレス型かによって制約を受ける。セッション型の場合、(a)、(b)の両方とも選択可能である。セッションレス型の場合、自動的に(b)となる。ただし、WWW でもクッキーなどを用いて疑似的なセッションを作ることは可能であり、この場合は(a)を選択することも可能である。

## 4　インデクスの実現方式

### 4.1　インデクスファイル構成

複数のバージョンのインデクスを保持するインデクスファイル構成には、以下の2通りの方式が考えられる。

**(1) 個別インデクス方式**

各バージョンのデータベースごとにインデクスを作成する。検索集合作成時には対象バージョンによってインデクスファイルを切り換えて使用する。

更新頻度が増えると、ファイル容量だけでなく主記憶容量もそれに比例して必要になるばかりでなく、入出力バッファの効率が悪くなり性能低下をもたらす。

**(2) 統合インデクス方式**

複数のバージョンのレコードを含むデータベースに対して一括してインデクスを作成する。バージョンに関する情報もインデクスに保持する必要がある。検索集合作成時にはバージョンを検索条件に組み合わせて検索対象を限定する。

インデクスの実体は1個であるので、更新を繰り返してもファイル容量、主記憶容量ともに増加分はわずかですむが、バージョンを条件とした検索結果の絞り込みに計算資源を消費する。

### 4.2 検索処理方法

インデクス検索の実現方法には、前述の2つの検索集合保持方式と2つのバージョン切換方式のそれぞれについて上記の2つのインデクスファイル構成を用いる組合せが考えられるので、全部で8通りの場合がある。しかし、既存の検索エンジンをそのまま使うか、最小限の変更に留めることを考えると、実現可能性はおおよそ表1に示すようになる。

従来型の情報検索用エンジンにおいて、セッション開始時の最新のインデクスをセッション終了時まで継続して使う場合は(レコードリスト保持＋セッション単位切換＋個別インデクスファイル)にあたる。処理手順は明白であるが、適用可能なサービス環境は限られる。

表 1　インデクス検索方式の実現可能性

◎：実現性高い　○：実現可能
△：実現困難　×：実現不可能

| 集合保持方式 | インデックスファイル | |
|---|---|---|
| バージョン切換単位 | 個別 | 統合 |
| レコードリスト | | |
| セッション単位 | ◎ | ○ |
| 集合単位 | × | × |
| 検索条件 | | |
| セッション単位 | ○ | ○ |
| 集合単位 | △ | ◎ |

WWW用の検索エンジンにおいて、バージョンを維持しつつセッションレスの利用形態で使う場合は(検索条件保持＋集合単位切替＋統合インデクスファイル)にあたる。WWW環境で更新頻度が高い場合に有効であり、処理手順は以下のようになる。

**(1) 検索集合作成時**

インデクスファイルをオープンし、与えられた検索条件に最新のバージョン番号を組み合わせて検索実行し、一致件数を取得する。検索条件とバージョン番号を記憶する。

**(2) 表示時**

インデクスファイルをオープンし、記憶しておいた検索条件とバージョン番号で再検索し、与えられたレコード番号のレコードと記憶しておいたバージョンに該当するレコードデータを取得する。

この手順では表示時に再検索を要するので効率が悪いが、キャッシュを適切に使うことにより再検索の頻度を抑えることができる。処理がやや複雑であり、検索、表示ともオーバヘッドを小さくする工夫が必要となるが、更新頻度が高くなってもそれによる効率の低下は少ない。

この処理を効率的に実現できれば、バージョン切換単位やセッション型・セッションレス型にかかわらず即時に近いオンライン更新が可能となる。

## 5 バージョン管理方式の一例

上記の組合せを検索エンジンを用いて効率的に実現するためのレコードデータとインデクスのバージョン管理、および、その更新処理の一方式を示す。

**(1) レコードのバージョン管理**

論理的に同一のレコードでもバージョンが異なる場合は物理的に別のレコードとして扱う。こうすることにより、インデクスの各ポスティングにおいてレコードのバージョン情報を保持する必要がなくなる。

レコード中には、レコードデータの他にそのレコードの有効期間の開始時刻と終了時刻を収納しておく。ただし、時刻はデータベースのバージョン番号で表される論理的な数値であり、開始時刻はレコード作成時、終了時刻はレコード削除時の値である。また、最新バー

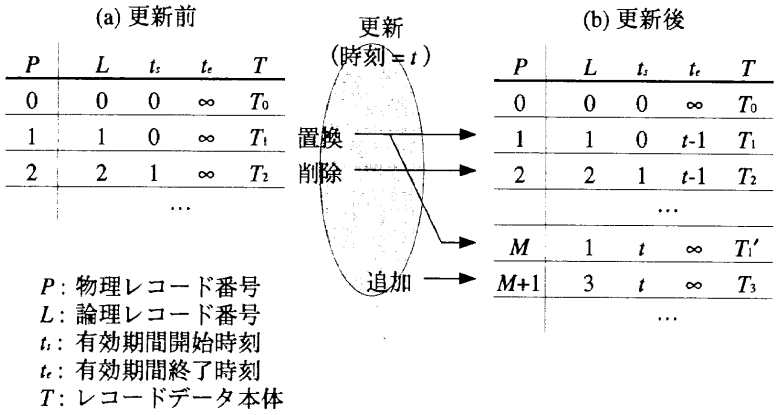



図 2　データレコードの更新方式

ジョンにおいて有効なレコードの終了時刻は無限大とする。

インデクスには、レコード中のデータのインデクスの他に、レコード有効期間の開始時刻と終了時刻のインデクスを作成する。

図2にレコードの更新時の処理を示す。削除レコードについては、物理レコードは実際には削除せず、有効期間の終了時刻のみを現在の時刻から1を引いた値に置き換える。追加レコードについては、新しい物理レコードを、有効期間の開始時刻を現在の時刻の値、終了時刻を無限大として作成する。置換レコードについては、現在の物理レコードは有効期間の終了時刻を現在の時刻から1を引いた値に置き換え、同時に、更新されたデータを含む新しい物理レコードを、有効期間の開始時刻を現在の時刻の値、終了時刻を無限大として作成する。レコードデータの更新が終了したらインデクスの更新を行う。

### (2) データベースのバージョン管理

データベースの各バージョンごとにその値とその時刻におけるデータベースの統計情報などを専用のバージョン情報レコードに収容してインデクスを作成することにより、アプリケーションからアクセスできるようにする。これにより、現在時刻がわかるだけでなく、tf-idf(term frequency / inverse document frequency)などによるランキングもバージョンごとに計算できるようになる。

図3にバージョン情報レコードを示す。物理レコードは通常のデータレコードと同一のデータベース中にあっても、あるいは、別データベースであっても構わないが、インデクスの更新は同期している必要がある。

| P | $t_v$ | S |
|---|---|---|
| $P_0$ | 0 | $S_0$ |
| $P_1$ | 1 | $S_1$ |
| $P_2$ | 2 | $S_2$ |
| | … | |
| $P_t$ | t | $S_t$ |

P：物理レコード番号
$t_v$：更新時刻（バージョン）
S：データベースの統計情報
t：現在時刻

図 3　バージョン情報レコード

データレコードを更新するときは、新たに現在時刻と更新後のデータベースの統計データを含む物理レコードを作成し、インデクスを更新する。データレコードとは別データベースにする場合は、データレコードを含むデータベースのインデクス更新を先に行う必要がある。

### (3) 検索集合作成時の処理

最新のバージョン情報レコードをフェッチして現在時刻を取得し、検索条件に式(1)の条件を加えてレコードを検索する。

$$t_s \leqq t_c \leqq t_e \qquad (1)$$

(ただし、$t_s$, $t_e$はそれぞれ各レコードの有効期間の開始時刻と終了時刻、$t_c$は現在の時刻)

### (4) 表示時の処理

検索条件に式(2)の条件を加えてレコードを再検索し、データを表示する。

$$t_s \leqq t_c \leqq t_e \qquad (2)$$

(ただし、$t_c$は表示対象の検索集合の作成時刻)

なお、検索エンジンがオンライン更新機能を備え、

データベースの更新において物理レコードが維持され、既存の検索集合のレコードリストをそのまま使える場合は、検索集合作成時に作成した集合のレコードリストを使用しても矛盾は生じない。

## 6　おわりに

情報検索システムにおいてオンライン更新を実現する方式について検討した結果、バージョン管理を行うことにより、比較的単純な手法で検索集合の論理的な一貫性を保てることがわかった。

今後は本稿で提案したバージョン管理方式の効率を実験を通して検討する予定である。

## 参考文献

[1] 大山ほか:総合目録オンラインDBと情報検索システムの連携方式, 学術情報センター紀要 Vol.9, pp.83-90, 1997.

[2] "Livelink Search Index Administrator's Guide", Open Text Corp, 1997.

[3] "Oracle ConText Option", http://www.oracle.co.jp/oco/top.html.

学術情報センター紀要　第10号(1998)

研究論文

# テキストの機能構造を用いた検索方式の比較：検索要求文の役割分析と同義語自動獲得による検索式拡張

# Comparison of Query Construction Methods using Text-Level Structure : Role Analysis and Query Expansion using Automatic Synonym Extraction

学術情報センター　神門　典子
Noriko KANDO
National Center for Science Information Systems

**要旨**

本稿では、テキストの機能構造を表わす構成要素カテゴリを付与した小規模な実験的日本語全文データベースを用いて、各種の検索方式の比較を行なった。その過程で、日本語全文検索の索引・検索要求解析法を検討し、日本語の自然言語で記述された検索要求文からの検索式構築を自動化し、検索語の重みを個別に考慮する方式へランキング方式を改善した。その結果、本稿で提案した概念役割を用いた検索、特定のディスコース・タイプを指定した検索、これらに同義語自動獲得による検索式の拡張を組み合わせた方式は、従来の方式よりも、検索効率が平均でそれぞれ28.1%、28.9%、32.3%、31.8%向上した。データベースからの同義語獲得は、テキストの機能構造を用いることにより、処理対象を限定した効率的抽出が可能となった。テキスト構造の全文検索への意義と今後の課題についても検討した。

**ABSTRACT**

This paper suggests an approach for textual information retrieval using text-level structure, and compares the effectiveness of various strategies of query construction from Japanese search topic sentences. The results of the preliminary experiments using a small- scale structure-tagged fulltext database of Japanese research papers shows that the role analysis of query terms using text-level structure, the default categories, and the role analysis and default categories with query expansion using synonyms automatically extracted from the database produce improvements of 28.1%, 28.9%,32.3%,31.8% respectively over the baseline. It seems that the text-level structure is effective for automatic synonym extraction through specifying the passages which likely contain synonyms effective for query expansion. The paper also discusses the implication of the text-level structure for textual information retrieval.

[キーワード]　情報検索、テキスト構造、ランク付け方式、役割分析、検索式拡張、日本語全文検索
[Keywords]　information retrieval, discourse-level structure, text structure, ranking methods, role analysis, query expansion, fulltext search with Japanese language

## 1 はじめに

テキストは、語や文の寄せ集めではなく、構造がある。従来の情報検索研究では、語や文レベルの現象に着目してきたが、テキストの構造的な特性を利用することにより、より効果的な検索と柔軟な利用が可能になると期待される。テキスト構造分析にはさまざまなアプローチがあるが、ここでは、テキストのジャンルに応じた特徴的な構成要素に着目する。これは、たとえば、学術論文なら、「背景」、「目的」、「方法」、「結果」、「考察」、「結論」といったテキスト内での役割や機能を表わすものであり、さらに細かい要素を認定することもできる(図1参照)。これらの特徴的な構成要素に

よって記述される構造(機能構造)は、テキストの利用者にとって自然なテキスト内容の構成であり、したがって、検索の過程でも利用可能だと期待される[1]。

このようなテキストの特徴的な構造を用いた検索は、各概念がテキスト中で果たしている役割や機能を識別し、ディスコース・タイプ[1]やテキストのジャンル、利用者の状況やタスクとも関わることから、テキスト構造を用いない検索よりも有効であると考えられる。テキスト単位の検索[2-4]だけでなく、パッセージ単位の検索、テキスト内のブラウジング、スキミング[5]など柔軟なインタフェースを提供する基盤となり、情報抽出、自動抄録[6-8]、電子雑誌の設計[9]などでも用いられている。

著者は、これまで、日英の複数分野の学術論文[10-13]、新聞記事[13-15]、看護日誌などの医療文書[13]などのテキスト構造を分析し、分析を自動化し[11-13,17-18]、テキストの機能構造を示すタグを付与した実験的な日本語全文データベースを用いて、テキスト単位およびパッセージ単位の検索の精度向上、テキストのスキミング、他文書中のパッセージとの関連付けなどにおいて、テキスト構造が有用であることを示してきた[16]。

本稿では、テキストの機能構造の特性に適した検索方式を検討する。主な目的は以下の2点である。

(1) 検索要求文中の各概念の役割を分析して、検索に用いる。

(2) 検索式の構築を自動化する。それにともに、再現率向上デバイスを組み込む。

(1)については、前報[16]の検索実検においては、個々の検索要求における概念の役割は分析せず、一般的に検索精度向上に効果があるデフォルト・カテゴリの認定を行なった[16]。しかし、これは、テキストの機能構造を用いたアプローチの中で、検索において一般的に重みをおくべきテキスト中のディスコース・タイプを識別するという側面を活用しているのみであり、個々の利用者の状況との関わり、テキスト中で個々の概念が果たしている役割の識別といった他の特性を十分に活用しているとはいえない。そこで、本稿では、検索要求中の各概念の役割を分析して検索に用いた。

検索式中で、検索語と、それが果たしている役割や機能などを表わす構成要素カテゴリ(図1)とを組み合わせることは複雑な作業である。構成要素カテゴリを用いた検索式を作成するためのインタフェースについては、デフォルト・カテゴリの使用、インタラクティブなカテゴリ決定、あらかじめ設定したカテゴリごとのスロットに検索語を入力する構造化スロットの3方式を検討した[19]。構造化スロットは、検索要求の多様な側面を利用者に想起させる効果があり、入力される検索語数が他の方式よりも多かった。これは、検索効率向上には有利である。しかしながら、スロットの選択が困難な場合もあるため、自然言語文で記述された検索要求から、各概念の役割を自動解析する機能も必要である。本稿では、その準備段階として、概念の役割分析のみは人手で行ない、検索効率への効果を検証するとともに、概念役割の自動解析ルールの材料を得ることを目的とする。

(2)については、前報[16]では、人手で同義語も含めた形で検索式を作成して検索実験を行なった。これは、テキストおよび検索要求中では、同一概念がさまざまな語句で表現されており、同義語を網羅的に利用できないために検索の再現率が低下するという問題を回避し、テキスト構造が検索精度へもたらす効果のみを検討するためであった。しかし、人手による検索式の作成は検索者の負担が大きく、現実的ではない。実用システムでは、ユーザの負担軽減のために自然言語文でのデータベースへの問合せから、検索式の自動構築が必要である。

そこで、本稿は、自然言語で記述された検索要求文から検索語の切り出しを自動化し、検索式の自動構築を行なった。その際、同一概念を表わす用語の多様性に対処し、再現率を向上させるデバイスとして、テキスト構造を用いてテキストから自動抽出した同義語による検索式の拡張を行なった。

以下、第2章では、情報検索をめぐる問題を概観して本稿のアプローチを明確にするとともに、第3章で実験方法の詳細、第4章で結果と考察を述べ、今後の課題を示す。

[1] ディスコース・タイプとは、テキスト内での陳述および文体の違いを意味する。たとえば、「事実の記述」、「意見の陳述」、「会話文」などがディスコース・タイプの例である。通常、1つのテキストは、複数の性質の異なるディスコース・タイプからなりたっている。(Chris Paice (1996) "Automatic Categorisation of Genre." Unpublished Proposal)

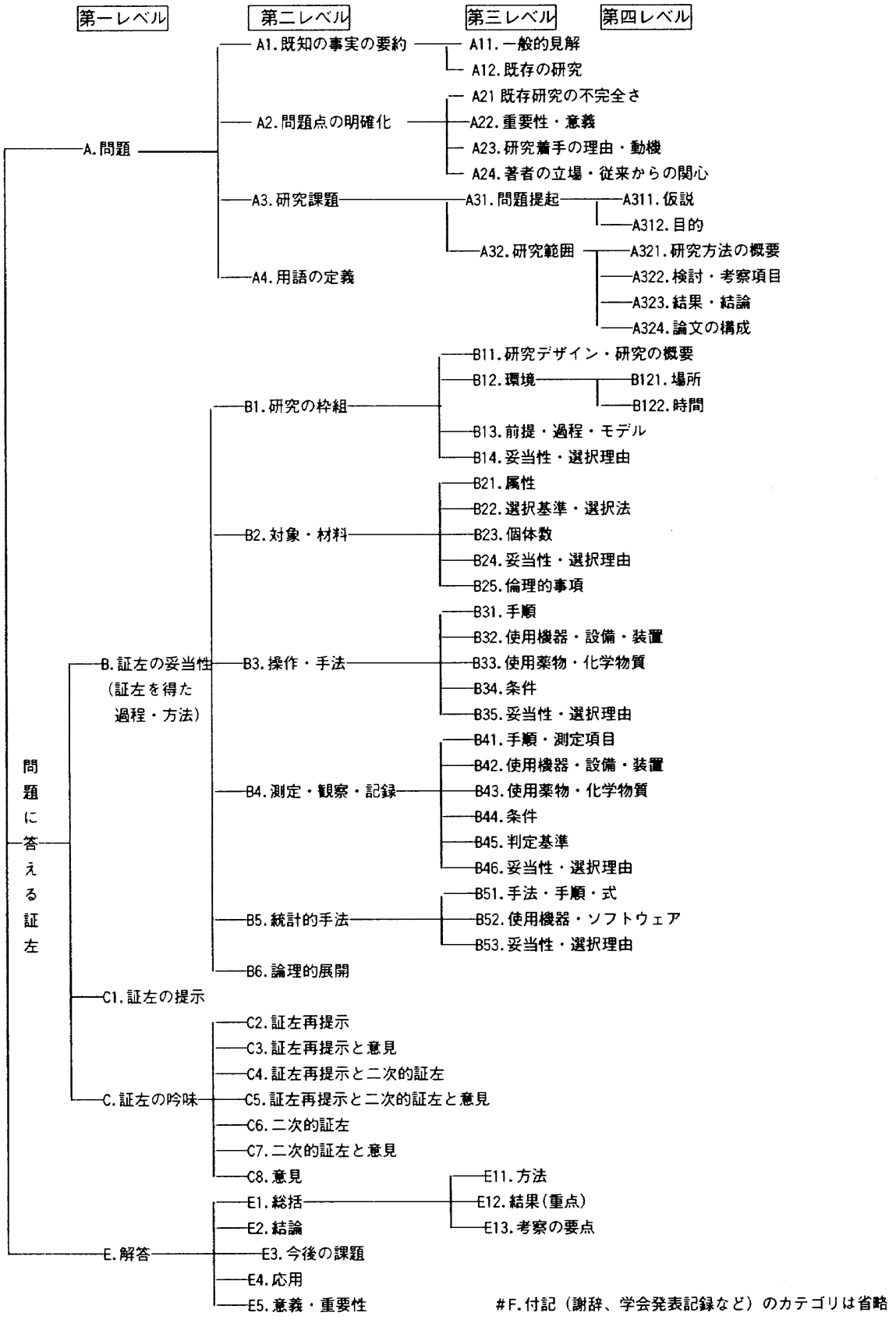



図 1　学術論文の構成要素カテゴリ

## 2 情報検索をめぐる問題とテキスト構造アプローチ

### 2.1 利用者の状況

近年、情報検索において、個々の利用者の携わっているタスクやその置かれている状況は、利用者の情報ニーズの主要な要素であり、検索結果のレレバンス判定に大きな影響を持つことが知られてきている[1,20-25]。利用者の状況は、利用者の年齢、属性といった静的な「利用者モデル」ではなく、個々の状況に応じた動的なものである。

利用者の状況そのものをシステムに組み込むことは困難であるが、利用者のタスクや状況によって、利用者がテキスト中で重視する部分や必要なテキストの種類も異なる。そのような特定の部分の認定にテキスト構造を用いることができる[1]。たとえば、論文を検索する際に、以下の2つを区別したい場合は多い。

(1) **検査法Aの有効性について論じた文献**
(2) **検査法Aの結果で被験者をグループ分けし、治療法Bの効果の差を調べた論文**

この二者は、Aという用語のデータベース中での出現回数だけでは識別できない。テキストの機能構造を用いることによって識別が可能である。また、研究過程のどの段階にいるかで論文のなかで着目する部分は異なる。新聞記事の検索でも、たとえば、新製品について、「いつ、どの会社が、なにを発売したか、その特性は」といった「事実」が知りたいのか、その製品の売り上げ予測や市場への影響などの「予測」や「意見」が必要なのかは利用者の状況によって異なる。また、同じトピックでも、電子メールか、新聞記事か、雑誌論文かといったテキストの種類によっても受け取り方は異なる。

このようなテキスト中での「役割」、意見、事実といったテキスト内での陳述の差異である「ディスコース・タイプ」、文書の社会的な機能の差異である「ジャンル」などは、利用者が、検索された文献が必要かどうか判定する際に重要である。しかし、これらの識別は、従来、情報検索が用いてきた「語の出現」や「文レベルの解析」だけでは困難である。著者が提案してきたテキストの機能的な側面に着目した機能構造分析は、これらの「役割」、「ディスコース・タイプ」、「ジャンル」など利用者の検索結果に対するレレバンス判定に影響を及ぼす要素を、テキストから自動抽出してシステムに組み込むための方策の一つである。

さらに、利用者のニーズ、関心、状況は、検索システムとのやりとりの間にも変化していく。テキスト構造は、検索されたテキスト内・間の柔軟なブラウジングやナビゲーションの基盤として、インタフェースの改善にも貢献しうる[5,16]。

### 2.2 検索式の拡張：再現率向上デバイス

情報検索において、テキストの機能構造は、検索語によって表わされる概念の役割や機能、ディスコース・タイプなどを指定することによって、主に精度向上デバイス(precision device)として機能する。検索システムでは、精度向上デバイスだけでなく、検索漏れを低減する再現率向上デバイス(recall device)も必要である。後者の主要なものは、シソーラスやレキシコンにより、あるいは、レレバンス・フィードバックによって、元の検索要求中の概念に関連の深い新しい語句を獲得して、検索式に追加する「検索式の拡張」である。

シソーラスやレキシコンを用いたアプローチは、言語横断(cross-lingual)検索でも必須である[26]が、日本語の学術文献を対象とした検索に必要な専門用語を、すべての分野について満足いくレベルで網羅したものは存在しない。これらのトゥールの構築には多くの知的作業が必要であるため、コーパスやデータベースからの自動構築の研究も進んでいる[27]。

一方、レレバンス・フィードバックは、あらかじめトゥールを用意する必要がない点で有利である。利用者によるレレバンス判定を用いるものだけでなく、上位ランクの文献から出現頻度の高い語句を抽出して検索式に追加する「自動レレバンス・フィードバック(automatic relevance feedback)」も有用であり[28-29]、さらに、テキストで検索語が出現する部分に集中して、語句を収集するLocal Context Analysisが有効であることが報告されている[28]。また、1〜数パラグラフに相当するほどの「非常に長い検索式」は、再現率、精度とも向上させることが知られているが、実用的な環境では、検索式処理とランキングの算出にかかる時間も考慮する必要があり、より関連が深いと思われる比較的少数の語を効率よく獲得することが重要である。

本稿では、検索式拡張に用いる語を効率よく抽出するために、テキスト中の「言い換え表現」用いて検索

式中の語句に意味的に関連のある語句を自動抽出し、同義語として追加することによって検索式拡張を試みた。「言い換え表現」の中には、

(1) 当該テキスト内でのみ通用する「操作的言い換え」(ex. A群：長期投与、B群：短期投与)
(2) 他のテキストでも通用する「語彙的言い換え」(ex.インターフェロン(以降、IFN))

とがある。(1)を考慮することによって、当該テキスト中でその概念を表わす語の出現頻度をより正しく計数することが可能になり、結果として、当該テキストのその語についてのランキングを引き上げる可能性はある。しかし、A群＝長期投与という「言い換え」は、当該テキストの外では成り立たない関係であり、再現率向上のための検索式拡張に用いる新しい検索語とはならない。それに対し、(2)は、テキスト外でも成り立つ「言い換え」表現であり、検索式拡張に利用できる。

しかし、(1)と(2)は、形式的に識別することは困難であり、単純なパターンマッチングによって(2)だけを抽出することはできない。そこで、10件の論文をサンプルとして抽出し、人手で分析して、言い換え表現の言語的パターンを調べ、両者の出現位置を検討した。その結果、(1)は、当該論文で用いた「方法」とその「結果」を述べた部分に多出し、(2)は、論文の冒頭からその論文で取り扱う「問題」を提起する部分と、「考察」の部分に出現することがわかった。これは、ディスコース・タイプとしては、(1)は「事実の報告」、(2)は「議論」もしくは「意見の陳述」の部分であり、図1に示した本稿で用いた分析枠組みである「構成要素カテゴリ」では、(1)は、「B(証左を得る過程)」の下位カテゴリと「C1(証左の提示)」、(2)は「A(問題)」と「C6(二次的証左)」に該当する。

そこで、本稿では、対象テキストを、データベース中の構成要素カテゴリにおける「A(問題)」の下位カテゴリとC6(二次的証左)」が付与された文に限定して、検索式拡張に利用可能な言い換え表現を認定した。これにより、単純なパターンマッチングによる関連語抽出の精度を向上させることができた。

なお、医学の原著論文では、「緒言－方法－結果－考察」という節の構造を持つものが多い。これらの節見出しをもつ節の間での文体の差異も言語学的に分析されている[30]。上述の構成要素カテゴリの「B(証左を得る過程)」と「C1(証左の提示)」は、ほぼ、「方法」と「結果」に、「A(問題)」は「緒言」、「C6(二次的証左)」は「考察」に含まれる場合が多い。しかし、このような節見出しをもたない論文もあり[2]、また、実際の論文では、他の節の方がふさわしい内容が書かれている場合もある[31]。節見出しよりも、記述内容によって付与される構成要素カテゴリの方が汎用性があるので、本稿では、構成要素カテゴリを用いた。

### 2.3 日本語テキストの索引付けと検索式構築

日本語の文では、英文と異なり、語と語の間に空白などの明示的な区切りがない。したがって、自然言語文として記述された情報要求から検索式を自動構築するには、どのような単位で文を解析するかを決定する必要がある。

日本語テキストの分割と索引付けの方式には、テキストを一文字ずつずらしながらN文字ずつに分割する「Nグラム索引」と単語単位に分割した「ワード型索引」がある。Nグラムには、次のような利点がある。

(1) テキストを語に分割するための大規模な辞書、および、そのメンテナンスが不要
(2) 索引作成が容易
(3)「分かち書き」の不一致によるモレがない

また、一般に文字の種類(アルファベットの数)が多いほどNグラムは有効であることから、近年、中国語、韓国語、日本語など東アジア諸語を対象とした情報検索研究では、Nグラム索引が多く用いられている[32-37]。しかしながら、N＝1としたuni-gramでは、ノイズが大きく、Nを大きくすると、N文字の任意の組合せということで索引の規模が飛躍的に増大するという問題があり[33]、このような問題を回避するために、uni-gramとbi-gramの組合せ[38]、Nを段階的に大きくする方式などが提案されている。

一方、ワード型索引では、形態素解析[32]や大規模辞書[36,39]を用いた分割のほか、文字ごとに算出した単語の語頭確率と語末確率に基づいて分割する統計的単語分割を用いた方式もある[40-41]。

しかしながら、検索効率については、Nグラムが有利とする報告と、ワード型がよいとするものとが混在

[2] British National Corpusに含まれる医学論文408件中、「緒言—方法—結果—考察」およびその類型に属する節構成の論文は302件(74.0%)であった。

しており、現状では、どちらがよいか判定しがたい状況である。その中で、テキスト分割をuni-gram(1文字)索引、検索文の解析をワード型とする方式が、「Ngram索引-Ngram検索」や「ワード索引-ワード検索」よりも、検索効率の面で有利であるという報告がある[32,36]。これは、索引作業を単純化し、索引の規模を最小にとどめるとともに、検索式の組み立てにおいて多様な検索方式に対応できるという利点もある。そこで、本稿では、このuni-gram索引、ワード型検索要求文分割を採用した。

さらに、東アジア諸言語の検索では、複合語の重み付けを重くしたほうが検索効率が向上するという報告[42-43]もある。そこで、本稿では、名詞の連続をフレーズとしてとらえ、検索要求文に含まれるフレーズに対して重みを加えた検索方式も検討した。

## 3　実験の概要

### 3.1　索引方式とデータベース

実験用のデータベースは、日本語で書かれたC型肝炎論文50件からなる。大規模な医学分野データベースであるJMEDICINEから、一定条件で検索した文献について、現物が入手可能なものを、実験目的での使用を条件に出版者の許可を得て、ＯＣＲで電子形態に変換したものである。そのうち、本実験では、タイトル、本文のみを使用し、著者名、著者所属機関名、抄録、著者キーワード、図表のタイトルとキャプション、引用文献、脚注などは除外した。ヨミはない。

索引は、2.3で述べたように、日本語は1文字とした。欧文(日本語文中に出現するalphabetで表わされた文字列)は、空白をデリミタとしてワード単位で作成した。検索エンジンでは、索引の各エントリが、テキスト中で次に出現する文字へのポインタを持っているため、1文字索引であっても、1文字以上の文字列からなる検索式との照合が可能でる。ひらがなも索引対象とした。

データベース中のレコードには、論文、タイトル、本文、節、小節、パラグラフなどの文書書式を示すSGMLタグと、図1に示した機能構造を示す構成要素カテゴリを付与した。この構成要素カテゴリは、複数領域の原著論文の分析を通して設定したものであり、もっとも詳細なレベルのカテゴリ(リーフカテゴリ)を各文に付与した[10-11]。テキスト中の各文に対し、約8割の精度でカテゴリ・ラベルの自動付与が可能であった[17]が、構成要素カテゴリの検索への効果をより明らかにするため、ここでは人手でカテゴリ・ラベルを付与したものを用いた。品詞、構文情報などを示すタグは付与していない。

### 3.2　検索エンジン

検索エンジンは、OpenText6 (OpenText社、カナダ)を用いた。OpenTextでは、データベースは、一つの長い文字列としてみなされ、開始タグ('<'と'>'で囲まれたもの。ex.<title>)と終了タグ('</'と'>'で囲まれたもの。ex.</title>)とで囲まれた範囲を「リージョン(region)」といい、通常のデータベースのフィールドおよびレコードに相当するものとなる。リージョンは、入れ子構造を持つことも、複数の系列のリージョンが交錯することも可能であり、多様な方式で分析したテキスト構造を柔軟に扱うことができる。リージョンは、テキストに埋め込んだタグによって設定する「物理リージョン」のほか、索引作成時に複数リージョンの組み合わせや検索時に論理的に設定することも可能である。検索時には、入れ子になったリージョン間の包含関係演算'including'と'within'が可能である[44]。検索結果は、データベース中あるいは特定のリージョン中での文字列の出現回数(ポインタセット)として、または、当該文字列を含むリージョン数として、示される。

### 3.3　検索要求の解析と検索式構築

前報とは別に、検索要求36個を、医学研究者から実験目的で収集した。自然言語の文として記述されている。検索式は、2.3で述べたようにワード型とした。検索要求文のワードへの分割は、日本語形態素解析システムChasen Ver.1.5[45]を用いた。あらかじめ定義したストップ・フレーズ(ex.文末の「文献」、「研究」、「が欲しい」、「はないか」など)を削除し、「名詞」、「未定義語」を検索語とした。名詞の中でも「サ変名詞」は動詞が後接しない場合のみ検索語とした。「形容詞語幹」は「名詞」または「未定義語」が後接する場合のみ検索語とし、「名詞性名詞接尾辞」と「名詞性接頭辞」は単独では検索語とせず、「名詞」または「未定義語」と接続する場合のみ検索語とした。

検索結果のランク付けは、OpenText6のRank-Mode "Relevance1"を用いた。これは、OpenTextのCommand Languageでは、

$$z1 = \text{region } R \text{ rankedby}.k.m \ S$$

の形式をとる。$m$ はランク付けの方式であり、ここでは"Relevance1"を指定した。$k$ はユーザ定義の重みであり、本稿では特に明記しないかぎり $k=1$ とした。Revevance1では、検索結果集合 z1は、文字列 $S$ がリージョン $R$ 内での出現する回数と、当該リージョンの長さ、データベース全体での文字列 $S$ を含むリージョン $R$ の数によって、各文献に重み付けされ、それによって検索結果がランク付きで出力される[44]。これは、多くの情報検索システムで用いられる標準的な重み付け方式である tf•idf の一形式である。

検索については、以下の7種類の方式を比較した。(ただし、$Z_1$, $Z_2,\ldots,Z_n$ は検索結果集合、$S_1,S_2,\ldots,S_n$ は検索語とする)

### (1) WORD(基本型)

$$Z_1 = \text{region}\quad \textit{ARTICLE}\quad \text{rankedby}.k.m\ (S_1 + S_2 + ,\ldots, + S_n)$$

前報[16]で用いた方式である。検索語 $S_1,S_2,\ldots,S_n$ が出現する回数によって各文献 (*ARTICLE* というregion)の重み付けをする。検索語 $S_1,S_2,\ldots,S_n$ の個別の出現回数は考慮していない。これを検索方式比較のベースラインとする。

### (2) WORD+Dtype(基本+構造型)

$$Z_1 = \text{region}\quad \textit{ARTICLE}\quad \text{rankedby}.k.m\ (S_1 + S_2 + ,\ldots, + S_n) + \text{region}\quad \textit{ARTICLE}\quad \text{rankedby}.k.m\ ((S_1 + S_2 + ,\ldots, + S_n)\ \text{within}\ C_x)$$

前報[16]で「テキスト構造を用いたランキング方式」として採用したものである。基本型で算出される各文献の重みに、検索語が特定のカテゴリに出現する場合の重みを加えている。カテゴリ $C_x$ は、固定したデフォルト・カテゴリであり、前報で、検索効率向上に効果があることが示された「B(証左を得る過程)」の下位カテゴリと「C1(証左の提示)」を用いた。(1)WORDと同様に、検索語群全体の出現回数のみを考慮し、各検索語の個別の出現回数は考慮していない。

### (3) WEIGHT(拡張型)

$$Z_1 = (\text{region}\ \textit{ARTICLE}\ \text{rankedby}.k.m\ S_1) + (\text{region}\ \textit{ARTICLE}\ \text{rankedby}.k.m\ S_2) + ,\ldots, + (\text{region}\ \textit{ARTICLE}\ \text{rankedby}.k.m\ S_n)$$

上記(2)WORD+Dtype型では、検索要求中のすべての検索語について、データベース中で同一の特定カテゴリが付与された文中の出現する場合を重視する方式である。それに対して、本稿では、検索要求中の語の役割を分析して、それぞれの役割に応じて、検索時に重みを置くカテゴリを決定する方式を試みる。そのためには、各検索語毎に独立して、検索対象リージョンを決定し、個別に出現頻度を計数したものに基づいて、各文献の重みを計算できる方式が必要である。この(3)WEIGHTは、そのための基本的なランキング方式となる。

また、上述の(1)WORD型、(2) WORD+Dtype型では、検索語 $S_1,S_2,\ldots,S_n$ は、データベース中で出現頻度が少ない語も、出現頻度が高い語も同じ重み付けで、それらの語を含む文献の重みが計算されているのに対し、この(3)のWEITGHT型では、個々の検索語のデータベース中での出現の程度が考慮された重み付けになる。これによって、精度の向上が期待される[3]。

### (4) PHRASE(フレーズ型)

名詞、未定義語、形容詞語幹、名詞性名詞接辞のいずれかが、他の品詞の語句やストップフレーズを挟まずに、連続して出現する場合は、それを一つながりの句、すなわちフレーズとして抽出し、フレーズに対する重み付けを、(3)のWEIGHT型に追加した。

用語の表記の多様性に対応するため、フレーズは、フレーズを構成するワード間の距離を3文字以内とする順序指定の近接演算とした。これによって、たとえば、検索要求文中に「情報検索」というフレーズがある場合、データベース中の「情報検索」、「情報の検索」、「情報の高度検索」、などの文字列が検索されるが、「情

[3] WEIGHTの方式は、新NACSIS-IR開発過程での大山敬三助教授の指摘による。ここでは、常に論文単位で検索を行なうため、単純化した式を示したが、本来の式は、4.2で後述するように特定regionを指定した検索でも矛盾なく処理できる形式である。

報を用いた検索」、「検索情報」などは検索されない。

$Z_2 = Z_1 +$ (region *ARTICLE* rankedby.$k$.$m$ ($S_i$ fby.d $S_{i+1}$))

(ただし、$Z_1$はWEIGHTによるランク付き検索結果集合。fbyは順序指定の近接演算子。$S_i$の先頭から$S_{i+1}$の先頭までの距離がd以内のものがマッチする。ここでは、dは「$S_i$の長さ+3」)

**(5) SYNONYM(同義語型)**

検索式中の任意の検索語$S_i$について、同義語$S_i'$,...,$S_i''$がある場合、検索式中の

(*region ARTICLE* rankedby.$k$.$m$ $S_i$) を

(*region ARTICLE* rankedby.$k$.$m$ ($S_i$ + $S_i'$ +,...,+ $S_i''$)) によっておきかえる。

同義語$S_i'$,...,$S_i''$は、検索要求中の検索語$S_i$の同義語として、下記3.4の方式で、データベースから自動的に獲得したものある。$S_i$から$S_i''$は同義語として扱い、この中のどの語が出現しても同等のものとして考える。個々の語の重みではなく、全体として1つの語のグループとして考えるため、この方式となった。

**(6) ROLE(ロール分析型)**

$Z_1$ = (region *ARTICLE* rankedby.$k$.$m$ $S_1$)
+ (*region ARTICLE* rankedby.$k$.$m$ ($S_1$ within region $C_1$))
+ (*region ARTICLE* rankedby.$k$.$m$ $S_2$)
+ (region *ARTICLE* rankedby.$k$.$m$ ($S_2$ within region $C_2$))
+ ,...,
+ (region *ARTICLE* rankedby.$k$.$m$ $S_n$)
+ (*region ARTICLE* rankedby.$k$.$m$ ($S_n$ within region $C_n$))

ここで、カテゴリ$C_1$, $C_2$,...、$C_n$は、任意の構成要素カテゴリのタグが付与されたリージョン名とする。region *ARTICLE* rankedby.$k$.$m$ ($S_i$ within region $C_i$)は、各検索語$S_i$が、カテゴリ$C_i$で表わされる役割を果たしているとき、$S_i$のデータベース中で$C_i$が付与された文中での出現についての重み付けを求めている。この検索方式は、このように、(3)のWEIGHT型に、各検索語について、データベース中の当該役割に応じた部分に出現する語の重みを加えたものである。これは、言い替えれば、役割に適した部分に出現した検索語にマッチする語について重み付けを2倍にしている。

なお、検索式は、自然言語文で記述された検索要求から自動構築しているが、役割分析の部分のみは、人手で行なっている。

**(7) D-TYPE(DiscourseType：ディスコース・タイプ型)**

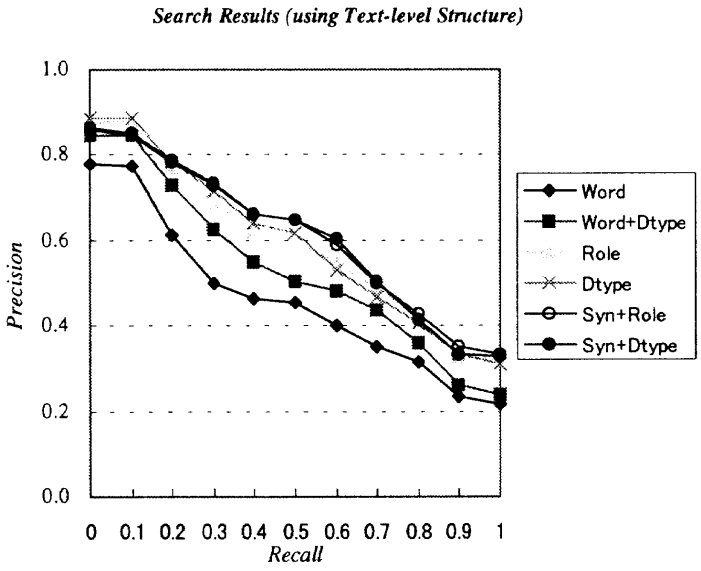
前報であきらかになった、検索精度改善に効果があるデフォルトカテゴリ「B(証左を得る過程)」の下位カテゴリと「C1(証左の提示)」を用いた。これは、(3)のWEIGHT型に、各検索語が「B(証左を得る過程)」の下位カテゴリと「C1(証左の提示)」が付与された文に出現する場合の重み付けを加えたものであり、(6)ROLE型の、カテゴリを指定する'region C *'の部分を、'(region B+ region C1)'に置き換えたものである。

### 3.4 同義語の収集

データベースから、検索語を含むパタンのパタンマッチングにより、同義語を収集した。

パターン1: X([フィラー1] Y [フィラー2])
または、
X([フィラー1] Y [フィラー2]：
パターン2: Y([フィラー1] X [フィラー2])
または、
Y([フィラー1] X [フィラー2]：
パターン3: (X：[フィラー1] Y [フィラー2])
パターン4: (Y：[フィラー1] X [フィラー2])
(ただし、Xは、検索要求中の語句。YがXの同義語として抽出される語。フィラー1は、「以降」、「以下」、「以後」など。フィラー2は、「とする」など。いずれも省略可能)

パタンマッチングは、データベース中のタイトル(<title>)、構成要素カテゴリ(図1)の「A(問題)」のすべての下位カテゴリ(<A12>、<A21>...など)、「C6(二次的証左)」のカテゴリ・ラベルが付与された文に対してのみ行なった。これは、2.2で述べたようにテキスト外でも有効な言い換え関係を効率よく抽出するためである。データベース中のテキストには、形態素解析、構文解析などの言語学的処理はほどこされてなく、プレーンテキストとのパタンマッチングによって抽出した。パターン2では、文頭、ひらがな、「、」の




| | precision | | | | | |
|---|---|---|---|---|---|---|
| recall | Word (baseline) | WordDtype (%change) | Role (%change) | Dtype (%change) | SynRole (%change) | SynDtype (%change) |
| 0 | 0.7804 | 0.8483( 8.7%) | 0.8790(12.6%) | 0.8854(13.5%) | 0.8629(10.6%) | 0.8616(10.4%) |
| 0.1 | 0.7732 | 0.8483( 9.7%) | 0.8790(13.7%) | 0.8854(14.5%) | 0.8496( 9.9%) | 0.8482( 9.7%) |
| 0.2 | 0.6108 | 0.7311(19.7%) | 0.7686(25.8%) | 0.7857(28.6%) | 0.7870(28.8%) | 0.7857(28.6%) |
| 0.3 | 0.5004 | 0.6255(25.0%) | 0.6876(37.4%) | 0.7168(43.2%) | 0.7343(46.7%) | 0.7295(45.8%) |
| 0.4 | 0.4643 | 0.5505(18.6%) | 0.6209(33.7%) | 0.6388(37.6%) | 0.6616(42.5%) | 0.6618(42.5%) |
| 0.5 | 0.4570 | 0.5053(10.6%) | 0.6003(31.4%) | 0.6182(35.3%) | 0.6509(42.4%) | 0.6504(42.3%) |
| 0.6 | 0.3987 | 0.4832(21.2%) | 0.5429(36.2%) | 0.5302(33.0%) | 0.5909(48.2%) | 0.6020(51.0%) |
| 0.7 | 0.3531 | 0.4355(23.3%) | 0.4889(38.5%) | 0.4690(32.8%) | 0.5006(41.8%) | 0.5064(43.4%) |
| 0.8 | 0.3137 | 0.3608(15.0%) | 0.4214(34.3%) | 0.4034(28.6%) | 0.4299(37.0%) | 0.4163(32.7%) |
| 0.9 | 0.2351 | 0.2614(11.2%) | 0.3301(40.4%) | 0.3315(41.0%) | 0.3519(49.7%) | 0.3345(42.3%) |
| 1 | 0.2159 | 0.2400(11.2%) | 0.3200(48.2%) | 0.3124(44.7%) | 0.3338(54.6%) | 0.3266(51.3%) |
| averag | 0.4639 | 0.5355(15.4%) | 0.5944(28.1%) | 0.5979(28.9%) | 0.6139(32.3%) | 0.6112(31.8%) |

図 2　テキスト構造を用いた検索結果の比較：WORD、WORD-Dtype、Role、DType

いずれかをデリミタとして、そのデリミタから'('までの距離が最短になる部分を同義語Yとして抽出した。また、ここで抽出された同義語を新たな出発語として、同義語抽出を複数回繰り返すことによって、より多くの同義語、関連語を抽出することができる。

今回は、同義語抽出を2回繰り返し、2回以上検索された語を同義語として、検索式に追加した。同義語抽出の繰り返しの効果は、以下の例で見ることができる。

**出発語:**「インターフェロン」
**1回目の同義語抽出：**
　インターフェロン --> IFN、IF
**2回目の同義語抽出：**
　IFN --> インターフェロン、interferon
　IF --> インターフェロン
**抽出された同義語：**　「IFN」、「IF」、「interferon」

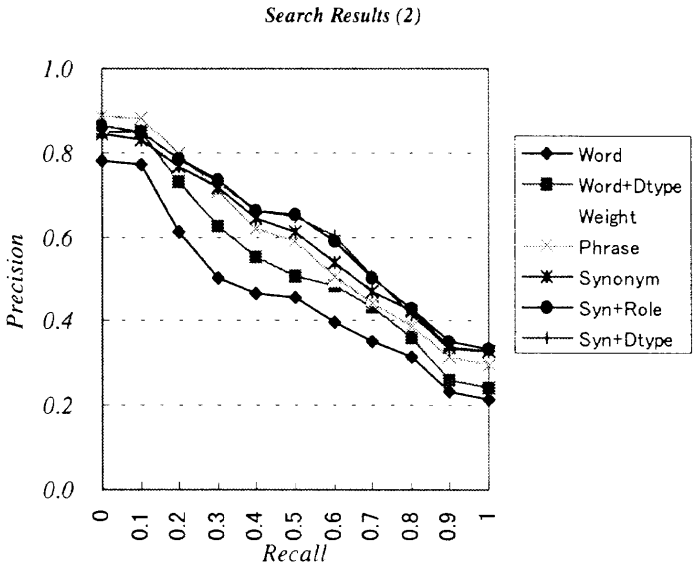
## 4　結果と考察

### 4.1　検索結果

3.3で述べた7通りの検索方式で検索した結果を図2に示す。収集した検索要求36件のうち、正解文書が0〜3件のものを除外した残りの25件の結果を図2、図3に示した。検索結果は、Salton と McGill[[46]に従い、ランク付きの検索結果ごとに再現率が0.0、0.1、0.2、...、1.0の11点での精度を算出し、各検索方式ごとにその平均を求めた。

前報[16]と同様に、基本型のWORDにデフォルト・カテゴリによる重み付けを加えたWORD+D-Typeは、ベースラインとしたWORDよりも、平均で15.4%精度が向上した。特に、上位〜中位での改善率が大きかった。

今回新たに提案した検索語の重みを個別に算出する




| | precision | | | | | |
|---|---|---|---|---|---|---|
| recall | Word | WordDtype (%change) | Weight (%change) | Synonym (%change) | SynRole (%change) | SynDtype (%change) |
| 0 | 0.7804 | 0.8483( 8.7%) | 0.8630(10.6%) | 0.8440( 8.1%) | 0.8629(10.6%) | 0.8616(10.4%) |
| 0.1 | 0.7732 | 0.8483( 9.7%) | 0.8630(11.6%) | 0.8307( 7.4%) | 0.8496( 9.9%) | 0.8482( 9.7%) |
| 0.2 | 0.6108 | 0.7311(19.7%) | 0.7612(24.6%) | 0.7662(25.4%) | 0.7870(28.8%) | 0.7857(28.6%) |
| 0.3 | 0.5004 | 0.6255(25.0%) | 0.6923(38.3%) | 0.7159(43.1%) | 0.7343(46.7%) | 0.7295(45.8%) |
| 0.4 | 0.4643 | 0.5505(18.6%) | 0.6188(33.3%) | 0.6453(39.0%) | 0.6616(42.5%) | 0.6618(42.5%) |
| 0.5 | 0.4570 | 0.5053(10.6%) | 0.5923(29.6%) | 0.6116(33.8%) | 0.6509(42.4%) | 0.6504(42.3%) |
| 0.6 | 0.3987 | 0.4832(21.2%) | 0.4966(24.6%) | 0.5410(35.7%) | 0.5909(48.2%) | 0.6020(51.0%) |
| 0.7 | 0.3531 | 0.4355(23.3%) | 0.4524(28.1%) | 0.4707(33.3%) | 0.5006(41.8%) | 0.5064(43.4%) |
| 0.8 | 0.3137 | 0.3608(15.0%) | 0.3980(26.9%) | 0.4239(35.1%) | 0.4299(37.0%) | 0.4163(32.7%) |
| 0.9 | 0.2351 | 0.2614(11.2%) | 0.3288(39.9%) | 0.3360(42.9%) | 0.3519(49.7%) | 0.3345(42.3%) |
| 1 | 0.2159 | 0.2400(11.2%) | 0.3121(44.6%) | 0.3289(52.3%) | 0.3338(54.6%) | 0.3266(51.3%) |
| avrag | 0.4639 | 0.5355(15.4%) | 0.5799(25.0%) | 0.5922(27.7%) | 0.6139(32.3%) | 0.6112(31.8%) |

図 3　検索結果(2)

方式を基盤とし、検索語の概念役割を分析したROLE、デフォルト・カテゴリを用いたD-Type、これらに自動抽出した同義語を加えて検索式を拡張したSYNONYM＋ROLE、SYNONYM＋D-Typeは、いずれも検索効率がWORD＋D-Typeよりもさらに向上し、ベースラインと比較した検索効率の向上は、11 points precisionの平均で、それぞれ、28.1%、28.9%、32.3%、31.8%であった。ROLEとD-Typeに有意な差はなかった。同義語を加えたSYNONYM＋ROLE、SYNONYM＋D-Typeは、それぞれROLE、D-Typeと比べて最上位ではやや精度が下がる傾向が見られたものの、全体としては検索効率が向上し、同義語自動抽出を用いた検索式の拡張は再現率向上デバイスとして機能している。

検索方式の基本となるWEIGHT、PHRASE、SYNONYMの結果は図3に示した。これらの、ベースラインであるWORDと比較した検索効率の向上は、11 points precisionの平均値で、それぞれ、25.0%、25.6%、27.7%であった。個々の用語の重みを個別に考慮するWEIGHTは、検索効率向上に大きな貢献があった。WEIGHTに複合語を追加したPHRASEは、上位ランクでごくわずかに精度を向上させた傾向がみられたが、WEIGHTとの有意差はなかった。SYNONYMは、WEIGHTと比べて、最上位ではやや精度が低下するが、中位～下位ランクで検索効率の向上が見られ、全体としては検索効率は改善されている。

WEIGHTが検索精度向上に大きく貢献しているため、図2に示したROLEとD-TYPE独自の貢献は3-8%程度であり、上位～中位を中心としたものであった。SYNONYMに概念役割による重み付けを加えたSYNONYM＋ROLEと、デフォルト・カテゴリによる重み付けを加えたSYNONYM＋D-Typeにおいて

も、いずれもSYNONYMよりも上位～中位で検索効率が向上している。

### 4.2 考察

#### (1) 構成要素カテゴリの利用

構成要素カテゴリを検索語と組み合わせて使用した検索では、前報と同様、デフォルト・カテゴリは、検索精度を向上させる効果が見られた。検索要求中の個々の概念の役割を分析するROLEについては、精度が向上する効果は見られたが、全体としてはデフォルト・カテゴリとほぼ同程度であった。どちらが有利かは検索要求ごとに異なっており、今後は、個別の概念の役割分析がより有効な場合を明確化するとともに、使用する適切な構成要素カテゴリのレベルを検討する必要がある。また、1文程度の情報要求文だけでは、各概念が論文の中で果たしている役割について十分な解析ができず、実際にはなにも構成要素カテゴリを付与しなかった検索語もあった。今後は、利用者とのインタラクションも含め、検索要求の収集法を再検討も必要である。

ROLEの解析については、同一概念が検索質問によって異なる役割で用いられる場合があるため、用語毎に対応する構成要素カテゴリをあらかじめ特定することはできない。しかし、検索要求文中の機能語と動詞を中心に、自動解析に用いることができる手がかりを得ることができた。今後は、これらの手がかりと論文への構成要素カテゴリの自動付与に用いたルールに加え、既存の検索要求文の係り受け解析の手法[47]を用いて、概念役割解析の自動化を行なう。

また、今回用いたROLEとD-TYPEの検索方式では、当該カテゴリのリージョンに限定した重みの算出になっていないので、今後は、ランキングの方式を、以下のように当該カテゴリの部分に集中した算出方式に変えて検討をする必要がある。

$Z_1$ = (region *ARTICLE* rankedby.$k$.$m$ $S_1$)
+ (region *ARTICLE* rankedby.$k$.$m$ $S_2$)
+, . . . ,
+ (*region ARTICLE* rankedby.$k$.$m$ $S_n$)

$Z_2$ = (*region* $C_1$ rankedby.$k$.$m$ ($S_1$ within region $C_1$))
+ (region $C_2$ rankedby.$k$.$m$ ($S_2$ within region $C_2$))
+, . . . ,
+ (*region C* rankedby.$k$.$m$ ($S_n$ within region $C_n$))

$Z_3$ = region *ARTICLE* including $Z_2$

$Z_4 = Z_1 + Z_2 + Z_3$

なお、構成要素カテゴリのテキスト(データベース中の各文)への自動付与については、50～120件程度の日本語論文群に対しては、表層的なパターンマッチングを主体とした人手作成ルールによる処理による実現可能性を示している[17]。また、D-TYPEで用いた程度の上位カテゴリ(A3、B、C1、Eなど)のレベルでは、日本語でも英文でもほぼ同じ精度での自動付与が可能であることを示唆する予備的結果を得ている。今後は、使用するカテゴリのレベルを検討するとともに、学習手法を用いたカテゴリ付与ルール構築の自動化と大規模なコーパスへの適用を検討したい。

#### (2) 同義語の自動獲得

テキスト構造を用いて自動的に抽出した同義語を用いたSYNONYMは、中～下位ランクに効果があり、検索モレを低減させる再現率向上デバイスとして機能した。テキスト構造を用いたことにより、関連語を抽出する箇所を特定することが可能になり、少ない処理量でテキストから検索に効果がある語を抽出することができた。

また、本稿で着目した「言い換え関係」は、学術論文では、各論文において主要な役割を果たしている専門用語に特に頻繁に見られた。日本語の学術論文の場合、専門用語は、日本語訳、英語の原綴、カタカナ、略語など、著者によって、同一概念が多様な形で用いられ、カタカナ表記にもゆれがある。全文検索では、この多様性が検索モレの原因の一つとなる。これは後述する「語彙的多言語データベース」問題であり、これに対処するためには、検索語とは異なる言語のものも検索できる「**言語横断**(*cross-lingual*、または、*cross language*, *cross-linguistic*)検索」手法が必要である。言語横断検索は、主として検索式を対象言語に翻訳して検索を行なうことによって異なる言語のテキストの検索を実現するが、そのためには、多言語シソーラスやレキシコン、対訳辞書などのトゥールが有効である。しかしながら、学術情報の検索では、常に最新概念を網羅する必要があり、トゥールの更新とメンテナンスは困難になる。それに対し、本稿では、論文内での「言い換え表現」によってランタイムにデータベースから

同義語を抽出することによって、これら専門用語表記の多様性、特に複数言語にまたがる多様性に対処する簡便な一方策を提案した。

なお、複数言語が含まれる「多言語データベース」には以下の5類型がある[26]。

(1) **並列型** (Parallel)： 完全対訳の複数言語レコードの組からなる
(2) **準並列型** (Equivalent)： ほぼ内容的に一致する複数言語レコードの組からなる
(3) **混在(非並列)型** (Non-equivalent)： 複数言語のレコードが、内容の対応する組としてではなく、個別に混在している
(4) **部分的** (Componential)： 1レコード中の特定部分のみが異言語で表記
(5) **語彙的** (Lexical)： 1レコード中に異言語の用語を含む

(4)は、たとえば、日本語論文で抄録と図表キャプションが英文のみで表記される場合が相当する。(5)は日本語テキスト中で、特定の語が英語で記述される場合などである。

(1)と(2)では、複数言語の対応したレコードを作成するコストはかかるが、検索では、英語の検索語は英語レコード、日本語の検索語は日本語レコードというように、組になった単言語検索であり、検索の技術としては比較的容易である。(3)～(5)の検索は、検索語と異なる言語のものも検索できる「言語横断検索(cross-lingual retrieval)」でなければ対応できない。

日本語の学術論文を全文データベースとした場合、一つの雑誌に英文論文と日本語論文が収録され((3)混在型)、日本語論文でも抄録・図表キャプションなど英文のみの要素があり((4)部分的)、専門用語などが英文のみで表記される((5)語彙的)ことが多く、主として日本で生産された情報だけを対象としたデータベースであっても、効果的な検索をおこなうには、言語横断検索の技法を用いる必要がある。今回提案した方法は、この言語横断検索へ応用可能な簡便な手法の一つともなる。今後は、より大規模なデータベースでの検討が必要である。

### (3) ランキング方式

検索効率の改善には、個々の検索語を独立して重みを算出するWEIGHT方式が有用であることがわかった。ROLE、D-Typeは上位～中位ランクで、SYNONYMは中位～下位ランクでの検索効率改善に有用であった。複合語を重複して追加するPHRASEは、特に顕著な貢献は見られなかった。今後は、PHRASEについて、ユーザ指定の重み付けkと近接演算の語間距離の調整、共出現条件との組合せなどを検討する必要がある。

### (4) データ特性と一般化

本稿で用いたデータベースは、C型肝炎という限定された主題領域の小規模なものであるが、医学全般を対象とする大規模なデータベースから一定の条件によって選定したレコードから構成されている。本報での検索結果は、全般に、他の検索研究と比べて検索効率が高い。ランク付きの検索は、一般に、テキストの自動分類をした後で検索を行なうと検索効率が向上するとされており[48]、本稿では特定の主題領域に限定したデータベースを用いたために全般に検索効率が高くなったと考えられる。

しかしながら、本稿で提案した各種の検索方式は、主題領域を越えて適用可能なものである。構成要素カテゴリは、医学、物理学、経済学、心理学、国語学など複数の主題領域の論文の分析を通じて設定したものであり、他の主題領域にも適用可能である。また、テキストのジャンルに応じて特徴的な機能構造は、論文以外のジャンルを対象としても分析され、自動翻訳、自動要約などの研究でも用いられており[49,50]、テキストの機能構造の原理は他の種類のテキストにも応用できると考えられる。同義語の自動抽出については、対象となるデータベースの規模が拡大した場合は、検索における上位文書のみを対象とするなど、対象テキストの範囲を限定することによって処理効率を維持することが可能である。今後は、より大規模なデータを用いた検証を行ないたい。

## 5 まとめ

各文に役割を示す「構成要素カテゴリ」を付与した小規模な日本語全文データベースを用いて検索実験を行なった。その結果、以下の結果を得た。

(1) ランキング方式を改良し、検索語の重みを個別に算出する方式に変更したところ、検索効率が向上した。
(2) 構成要素カテゴリと検索語を組み合わせることによって、検索効率が向上した。
(3) 構成要素カテゴリを用いてデータベースレコード中の特定箇所に着目して、「言い換え表現」の表層

的パターンマッチングによって、検索語の同義語をデータベースから自動獲得した。これを用いた検索は、検索効率が向上し、とくに再現率向上デバイスとしての機能を果たした。

(4) 今後の課題として、大規模データベースへの適用、検索式の役割分析の自動化などがある。

## 参考文献


[1] Oddy,R.N.; Liddy,E.D.; Balakrichnan B.; Bishop, A.; Elewononi, J.; Martin, E., "Towards the use of situational information in information retrieval." *Journal of Documentation*. Vol.48, pp.123-171, 1992.

[2] Liddy, E.D. "The Discourse level structure of empirical abstracts ; an Exploratory Study." *Information Processing and Management*. Vol.27, pp.55-81, 1991.

[3] Liddy,E.; Myaeng,SH., "DR-LINK." *SIG-IR Forum*. Vol.18, pp.1-20, 1994.

[4] Rama,D.V.; Srinivasan,P. "An investigation of content representation using text grammars." *ACM Transactions on Information Systems*. Vol.11,No.1,pp51-75, 1993.

[5] Miike,S.; Itoh,E.; Ono,K.; Sumita,K., "A Full-text retrieval system with a dynamic abstract generation function." *Proceedings of the 17th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Dublin, 1994, pp.152-161.

[6] Paice,CD.; Jones,P., "The Identification of important concepts in highly structured technical papers." *Proceedings of the 16th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval.* Pittsberg, 1993, pp 69-78.

[7] Jones,PA. "Automatic Abstracting and Indexing of Technical Documents: an approach based on concet selection." PhD dissertation, Lancaster University, August 1995.

[8] Endres-Neggemeyer, B, et al. "How to implement a naturalistic model of abstracting : four core working steps of an expert abstractor." *Information Processing and Management*. Vol.31, pp.631-674, 1995.

[9] Dillon,A. "Designing usable electronic text : ergonomic aspects of human information usage." Taylor & Francis. 1994, 195p.

[10] 神門典子.「構成要素カテゴリを用いた原著論文の内部構造分析」情報処理学会研究報告(92-FI-25). pp.39-46, 1992.

[11] 神門典子.「原著論文の機能構造の分析とその応用」図書館学会年報. Vol.41, No.2, pp.49-61, 1994.

[12] 神門典子.「複数領域における日本語原著論文の機能構造分析」*Library and Information Science*, No.31, pp.25-38, 1994.

[13] 神門典子.「情報メディアの構造」.慶應義塾大学文学研究科博士論文. 1995, 256p.

[14] 神門典子.「新聞の報道記事の構造：索引作成作業と検索との関連から」. 書誌索引展望, Vol.19, No.1, pp.1-17, 1995.

[15] 神門典子.「認識特性に基づくテキスト構造の分析：日英新聞記事を例として」. 学術情報センター紀要, No.8, pp.107-129, 1996.

[16] Kando,N. "Text-level structure of research articles and its implication for text-based information processing systems." *Proceedings of the 19th BCS-IRSG Annual Colloquium on Information Retrieval Research,* Aberdeen, Scotland, UK, 1997, pp.68-81.

[17] Kando,N. "Text-level structure: Implications for Information Retrieval and the Potential for Genre Analysis." Paper presented at the NLP Seminar, Sheffield University, Sheffield, UK, 1997,24p.

[18] Kando,N. "An approach for automatic template creation for English research articles." Paper presented at the IR Seminar, Glasgow University, Glasgow, Scotland, UK, 1997.

[19] Kando,N. "An approach for text information retrieval, browsing, and extraction using discourse-level structure." [Poster] Paper presented at *the 20th annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Philadelphia, PA,USA, 1997.

[20] Schamber,L.; Eisenberg M.B.; Nilan, M.

S., "A re-examination of relevance : toward a dynamic, situational definition." *Information Processing and Management*. Vol.26, pp.755-776, 1990.

[21] Schamber,L. "1. Relevance and information behavior". *Annual Review of Information Science and Technology*. Vol.29, pp.3-48, 1994.

[22] Barry,C. "The Identification of User Criteria and Document Characteristics: Beyond the Topical Approach to Information Retrieval." PhD dissertation, Syracuse University, 1993. available from: University Microfilms, Ann Arbor, MI.

[24] Wang,P. "Users' information needs at different stages of a research project: a cognitive view." Paper presented at *ISIC '96*. 1996, 19 p.

[25] Sackett,DL.; Haynes,RB.; Tugwell,P., "Clinical epidemiology : a basic science for clinical medicine." Boston, Little Browns., 1985, 370p.

[26] Kando,N. "Cross Linguistic Scholarly Information Transfer and Database Services in Japan." [Panel] Paper presented at *the 1997 ASIS Annual Meeting*, Panel on Multilingual Databases, Washington D.C.,USA, 1997. .

[27] Jing,Y.; Croft,WB., "An association thsaurus for informatin retireval." *Proceedings of RIAO '94*, 1994, pp.160-169.

[28] Xu,J.; Croft,B., "Query expansion using local and global dodument analysis." *Proceedings of the 19th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval*, Zurich, 1996, pp.4-11.

[29] Buckley,C.; Singhal,A.; Mitra,M.; Salton, G., "New retrieval approaches using SMART : TREC-4." *Proceedings of the TREC 4 Conference*. NIST Special Publication", 1996.

[30] Biber,D. "13. Intra-textual variation within medical research articles." *Corpus-based Research into language*. Edited by Oostdijk. Rodoph, Al Lanta, 1994, p.201-221.

[31] Buxton,A.B.; Meadows,A.J., "Categorization of the information in experimental papers and their abstracts." *Journal of Research Communication Studies*. Vol.1, pp.161-182, 1978.

[32] Fujii,H.; Croft,W.B., "Comparison of Indexing techniques for Japanese Text Retrieval." *Proceedings of the 16th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Pittsburgh, 1993, pp.237-246.

[33] Ogawa,Y.; Iwasaki,M., "A new character -based indexing method using frequency data for Japanese documents." *Proceedings of the 18th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Seatle, WA, 1995, pp.121-129.

[34] Chien,LF. "Fast and quasi-natural lnaguage search for gigabits of Chinese texts." *Proceedings of the 18th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Seatle, WA, 1995, pp.112-120.

[35] Lee,J.H.; Ahn,J.S., "Using n-grams for Korean Text Retrieval." *Proceedings of the 19th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Zurich, 1996, pp.216-224.

[36] Xiangji,Huang.; Robertson,S.E., "Experiments on large test collections with probabilistic approaches to Chinese text retrieval." *Proceedings of the 2nd International workshop on Information retrieval with Asian Languages 1997* (IRAL 97), Tsukuba, Japan, 1997, pp.129-140.

[37] Buckley,C.; Singhal,A.; Mitra,M., "Using query zoning and correlation within SMART: TREC-5". In D.K Harman, Editor, *Proceedings of the Fifth Text Retrieval Conference (TREC-5)*. NIST Special Publication 1996.

[38] 小川泰，松田透.「ランキング文書検索における

スコア合成法の評価」情報学基礎研究会研究報告 47-FI-14, p.95-100, 1997.

[39] Nie,JY.;Brisebois,M.;Ren,X., "On Chinese text retrieval." *Proceedings of the 19th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Zurich, 1996, pp.225-233,.

[40] Vines,P.et al. "Indexing for Chinese text retrieval." *Proceedings of the 1st Australian Document Computing Conference,* pp.85-89, 1996.

[41] Ogawa,Y.; Matsuda,Toru., "Overlapping statistical word indexing : a new indexing method for Japanese text." *Proceedings of the 20th Annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Philadelphia, 1997, pp.226-234.

[42] Dongwook,S.; Hyojin,N.; Sejin,N., "Information noise reduction using pruning." *Proceedings of the 2nd International workshop on Information retrieval with Asian Languages 1997 (IRAL 97)*, Tsukuba, Japan, 1997, pp.148-157.

[43] Yamada,Koichi.; Mori,Tatshunori.; Nakagawa,Hiroshi., "Japanese Compound Words Matching for Information Retrieval." *Proceedings of the 2nd International workshop on Information retrieval with Asian Languages 1997* (IRAL 97), Tsukuba, Japan, 1997, pp.158-164.

[44] "Livelink index engine query language reference." OpenText Corporation, 1997.

[45] 松本裕司ほか.「日本語形態素解析システムChaSen v.1.5マニュアル」奈良先端科学技術大学院大学, 1997.

[46] Salton,G.; McGill,M., Introduction to modern information retrieval, McGraw-Hill, 1983. .

[47] 松村敦、池田和幸、高須淳宏、安達淳.「構造化インデックスを用いた情報検索システム」アドバンスド・データベースシンポジウム'97論文集, pp.151-158, 1997.

[48] Lewis,D.D., Machine learning and automatic text categorization. Paper presented at the tutorial of *the 20th annual international ACM-SIGIR Conference on Research and Development in Information Retrieval,* Philadelphia, PA,USA, 1997.

[49] 島津美和子、熊野明、吉村裕美子、中村真理子, 文書の種類を考慮した機械翻訳システムの構成, 情報処理学会第49回全国大会講演論文集(第3分冊), p.239-240, 1994.

[50] Moens,M.F.; Uyttendaele,C., Automatic text structuring and categorization as a first step in summarizing legal cases. *Information Processing and Management*, Vol.33 No.6, pp.727-737, 1997.

学術情報センター紀要　第10号(1998)

研究論文

# ヤクート語語彙研究(3)：動植物名称
## －ヤクート語英雄叙事詩オロンホを資料として－

# A study on lexical items in Yakut language (3): Names of fauna and flora in Saxa epic poem *olonxo*

神戸市看護大学　藤代　節
Setsu FUJISHIRO
Kobe City College of Nursing

**要旨**

東・中央シベリアを中心として、広い地域に分布するチュルク系言語ヤクート語の動物及び植物を表す語彙についての研究である。ヤクート語にはその主要な話者であるサハ(あるいはヤクート)族が今日の居住域に至るまでに辿ってきた過程が、特に語彙において反映されており、そのことがチュルク諸語の中にあって、ヤクート語が特異な言語とされる所以でもある。この言語の特異性を明らかにするため語彙の中でもセットとして扱えるものを取り上げる。既に色彩名称と方向表現に関する研究を発表したが、本稿でこれに続いて同じくセットとして動植物名称を選び、研究の対象とした。色彩名称、方向表現についてと同じくサハ族の英雄叙事詩オロンホ『クース・デビリィエ』を主な資料とし、この中に見られるヤクート語の動植物名称の整理分析を試みた。

**ABSTRACT**

This paper is concerned with lexical items in Yakut language, which belongs to the Turkic language family. The main part of the Yakut language community consists of Yakut or Saxa people. Their language traces well the process of their immigration to Siberia. In this paper, the author presents the data of the set of lexical items "fauna and flora" from the olonxo "Kīīs Dābilijā", one of the most famous epics of the Saxa people and suggests that such approach to the study of lexicon is effective to uncover the complicated situation of languages in Siberia.

This paper is the third and the last serial of the study on the lexicon of Yakut language, which the author published in the preceeding numbers of this bulletin.

[キーワード]　ヤクート語、オロンホ、シベリア、動植物名称、チュルク系諸言語
[Keywords]　Yakut language, olonxo, Siberia, Terms of fauna and flora, Turkic languages



## 0　はじめに

シベリアの東部を中心に広範囲に分布するヤクート語は現在、(ヤクート)サハ共和国を中心に38万人の話者人口を持つチュルク系の言語である。その話者の中心はヤクート族あるいはサハ族[1]と称する人々でサハ共和国の中心的民族でもある。これらの人々の祖先は現在のシベリアの地に定着するまでにバイカル湖沿岸、あるいは中央アジア域から移動してきたとされる。ヤクート語はその長期間の移動を反映してか、他の

チュルク系の言語との隔たりが大きく、他のチュルク系言語と下位グループにまとめられることもなく、単独で一つの下位グループを形成する。チュルク系言語の発祥の地とされるのは中央アジアもしくはモンゴリア高原であるがヤクート語は現在27程度あるチュルク系の言語の中で唯一、極北の地に移動した。本稿はこのように同系統の言語からはなれて特異な発展をしたヤクート語の語彙についての研究である。ヤクート語は他のチュルク系言語との間に語彙において特に大きな隔たりがみられる。本稿に先行してこれまでに色彩名称を「ヤクート語語彙研究(1)」[2]で方向表現を「ヤクート語語彙研究(2)」[3]で扱ってきた。ここではさらに「動植物名」を対象とし、分析する。色彩名称及び方向表現と同様、動植物名は語彙研究を行う上でセットとして取り出しやすいものであり、かつ、借用などがあれば比較的直接に語彙に反映されやすい。したがって、ヤクート語が移動の途中に接触した言語の語彙を取り込んだり、シベリアに到着してからサハ族が接することとなる動植物名を語彙の指示内容をかえる、つまり、語彙の意味を変えるなどによるヤクート語内部での語彙体系の変化を追いやすい。このことは未だはっきりとしたことがわからないヤクート語の形成のプロセスについて語彙全般の研究から明らかにしようという目的を達するために少なからぬ貢献をなすものである。

資料とするのはサハ族の口承文芸のジャンルの一つ、オロンホである。オロンホは英雄叙事詩で、さまざまな長さがある。かつてはこのオロンホの語りに特に秀でたオロンホフト olonxohut(olonxo+「～する人」を表す接辞)と呼ばれるひとびとがおり、かつてはオロンホは文化的継承であり、また娯楽としてもかかせないものであった。いずれにしてもサハ族の生活に深くかかわるものであった。ここでは先の「語彙研究(1)」及び「語彙研究(2)」とおなじく、オロンホ『クース・デビリィエ』(「炎の乙女」以下『オロンホ』と略して表す)を主な資料とする[4]。オロンホ全般について、また、『クース・デビリィエ』については「語彙研究(1)」を参照のこと。

## 1　動植物名称

本稿で扱う動植物名称は『オロンホ』から抽出できる語彙としては量的にはさほど大きなものではない。しかしながら、どの民族についてもあてはまることであるがこれらの動物名、植物名のそれぞれの語彙は日常生活に密接な関係を持っている。例えば、極北シベリアでトナカイ飼育に古来もっぱら従事している人々の言語ではトナカイを成長過程を細かく分けて語彙上区別するなどトナカイ飼育に関連した語彙が発達している。また、日本語の語彙においては農業、特に稲作に関する語彙が発達しているのも同様の例といってよいだろう。また、これら特定の分野の語彙が当該の言語の歴史のある時点で発達していたとしても例えば今日の日本社会で農業に関する少なからぬ語彙が日常的な使用から消えつつあるようにその言語のおかれた状況にかなり影響されると考えてよい。産業形態の変化が言語に影響をおよぼしている。本稿で扱う動植物名称はこのように現実社会の有様を反映する傾向が顕著な語彙セットである。

サハ族にとって、また、サハ族が現在の地に至るまでのサハ族の祖先にとって家畜飼育、放牧、狩猟などの動物を扱う産業は基幹的であったことに疑いはない。さらに現代社会の都会の一部を除けば、人間は常に自然と対峙して存在しているのであるから一般に自然界の事物はオロンホのような口承文学作品に様々なモチーフとして登場してもおかしくない。また、一方ではこのような文化的継承において数ある自然界の動物の中からどの動物を表す語彙が取り上げられるのかはサハ族に限らず、その民族の実際の生活形態とともに文化的、歴史的背景を反映する。

『オロンホ』に見られた動植物名等を次章に掲げる。なお、厳密には動植物とは言えないものも含まれているが人間や擬人化された「悪魔」や「神体」などを除き、広義に生物と認められるものは含めた。『オロンホ』は全てで4969行であるがテキストから抽出したものは約250行である。後掲のデータの章に各語彙の生起する箇所をあげてある。転写、略号については7に掲げる。

## 2　『オロンホ』に見られる動物等の名称

### 2.1　「馬」「牛」「トナカイ(鹿)」

表Ⅰに掲げた動物の名称にはシベリアに固有の動物が比較的少ないことがまず目につく。現在のサハ族の居住地からすれば「クロテン」やその他の毛皮獣が日常には重要な意味を持つ動物と考えられるが実際にはあまり頻繁には現れない。多いのは「馬」に関連する語彙である。サハ族の南方起源について常に言及されるのが「馬」を神聖な動物と見なしている点である。現在のヤクート・サハ共和国においては南部では家畜としての馬の占める地位は第1位であるが、首都ヤクー

ツク市を中心とする中部域ではむしろ「牛」が家畜として最重要である。さらに北部域ではトナカイ飼育が産業の中心であり、トナカイが最も生活に密着した動物である。しかし、どのような産業的形態をとっているかにかかわりなくサハ族には馬が神聖な動物であることは現在のサハ族の居住地においてふつうに見られる。これはサハ族がもと南方に居住しており、オロンホの原型が創られた時期には「馬」が実際の産業生活

＜表Ⅰ＞

（獣類）

| | |
|---|---|
| 「馬」 | at, sïlgï |
| 「作業用の馬」 | kölö |
| 「雄馬」 | atïïr |
| 「雌馬」 | biä |
| 「子馬」 | kulun(生後初めての秋まで) |
| 「灰黄で尾とたてがみが黒い馬」 | ulaan d'aγïl at |
| 「馬つなぎ」 | särgä |
| 「ライオン」 | xaxaj |
| 「牛」 | ïnax |
| 「雄牛」 | oγus |
| 「ヘラ鹿」 | tajax |
| 「熊」 | ähä |
| 「狼」 | börö |
| 「クロテン」 | kiis, saarba |
| 「獣」 | kïïl |
| 「獲物」 | bullun |
| 「白い毛皮のもの」 | ürüŋ tüüläax |
| 「黒い毛皮のもの」 | xara tüüläax |
| 「走るもの」 | süürük |
| 「家畜」 | süöhü |

（鳥類）

| | |
|---|---|
| 「鶴」 | turuja |
| 「ソデグロ鶴」 | kïtalïk |
| 「鳴（しぎ）」 | bïld'ïrïït |
| 「郭公」 | käγä |
| 「雉鳩」 | öt |
| 「ヤマドリの雄」 | ärdäγäs |
| 「黒ガモ」 | andï |
| 「雷鳥」 | xabd'ï |
| 「ヨーロッパ大雷鳥」 | ular |
| 「ハシボソガラス」 | suor |
| 「ホオジロガモ」 | arïlïas(oruluos のa化バリアント) |
| 「ユキホウジロ」 | tulluk |
| 「鳥」 | kïïl |
| 「小鳥」 | čïïčaax |

（昆虫等）

| | |
|---|---|
| 「ウジ虫」 | čïärbä(<Rus. červ') |
| 「甲虫」 | čoxu |
| 「蛙」 | baγa |

（魚類）

| | |
|---|---|
| 「フナ」 | bïččïï |
| 「ヒメハヤ」 | mundu |
| 「カワメンタイ」 | sïalïhar |
| 「魚」 | balïk |

においても重要であったことを示唆している。オロンホにおいて「馬」を表す語彙やそのバリエーションが多くみられることは動物名称は実生活を反映するジャンルの語彙セットであるが文化的背景を色濃く映した『オロンホ』にはこの語彙に伝統的な「神聖な動物」のイメージが残されていると考えられる。

逆に今日重要な動物である「牛」に関連する語彙はほとんど現れず、「トナカイ」はわずかに「ヘラ鹿」に共通のイメージをみるのみである。今日、サハ族が飼育している牛はロシア産の牛であり、主に乳牛を食用としていたロシア人の嗜好からの影響が大きい。そのためか、ヤクートでは一般に「牛」を表す語彙としてīnax の他に koruoba があるがこれはロシア語 korova「牛」からの借用である。『オロンホ』には「実り無き雄牛」としてあまり芳しくないイメージでわずかに現れるのみである。

### 2.2 「白い毛皮のもの」「黒い毛皮のもの」

表Ⅰにある「白い毛皮のもの」、「黒い毛皮のもの」はサハ族の民間伝承などにみられる修辞的な表現で前者は「馬」を後者は「家畜一般」をさす。この点でも前項に言及した「馬」を特別なものとする文化的背景がみられる。

### 2.3 「鳥」・「昆虫」・「魚」

「馬」についてほどではないが「鳥」を表す語彙が『オロンホ』の中には比較的多い。しかし、現在のサハ族の狩猟の対象として重要な鴨 xaas はあまり現れない。また、魚類は『オロンホ』の主題と内容に左右される可能性も高いと考えられるが同じく、シベリアにおける漁猟の重要性からすれば貧弱である。balīk「魚」という総称を除けばいずれも淡水魚であり、その点では内陸部を北へ移動したという歴史を反映している。

### 2.4 まとめ

『オロンホ』の動物その他の名称を調べるとサハ族の現在の居住地に密接な関わりをもつ生物を表す名称があまり見られず、おそらく『オロンホ』が創られた時期の民族の産業形態及び文化的背景を表している名称がよく痕跡を留めている。中には「ライオン」のようにサハ族が北方への移動のはるか以前に持っていたのではないかとおもわれるような語彙も残っている。

## 3 『オロンホ』に見られる植物等の名称

### 3.1 「草木」

シベリアの諸民族言語において比較的語彙が貧弱なのは植物をあらわす語彙である。特に農業は中部シベリア以北ではロシア人の到着以降においてのみ知られるようになり、穀物や野菜の名称等はシベリアの諸民族言語において多くの場合、ロシア語からの借用語彙である。例えば、「大麦」、「小麦」、「ジャガ芋」、「パン」等々。現在のヤクーチヤでは旧ソ連邦政権時代に農業が奨励された影響で農耕に関する語彙が日常的に使われるがやはりロシア語からの借用語彙が多い。『オロンホ』においては農業に関連する植物語彙はほとんど見

＜表II＞

| （樹木） | | （草類） | |
|---|---|---|---|
| 「木」 | mas | 「草」 | ot |
| 「灌木」 | ïarγa | 「ルクオラ（鮮やかな緑の草）」 | lukuora |
| 「白樺」 | xatïŋ, čačir, tuos | 「草の茎」 | ot saγa, tutuluga |
| 「白樺製の桶」 | xolloγos | 「葉」 | säbirdäx |
| 「白樺製の手籠」 | atïjax | 「根元」 | tört |
| 「ハイマツ」 | bolbukta | 「尾状花」 | kïïstä |
| 「樅」 | xarïja | 「いばら」 | adïrïïn |
| 「ヤナギ」 | üöt, talax | 「スゲ」 | kïrïs, ača |
| 「枯れ木、枯れ枝」 | xappït amïnn'ïk | 「タンポポ」 | altan ot |
| 「針葉樹」 | kötöγö | 「ヤマナラシ」 | tätiŋ |
| 「倒れた木」 | sis tïalara, siŋ | 「（有用な）野草」 | sirämä |
| 「密林」 | xaraŋa tïalar | 「葦の茎」 | tïn'ax |

られないと言ってよい。サハ族が移動してきた過程で農業に携わっていた時期があった可能性は極めて低いと考えてよかろう。

ここにリストアップした語彙のみでは『オロンホ』の語彙構成全体について特徴的なことは言えないが、植物語彙の中には複合語的な語彙が見受けられる。例えば「タンポポ」は altan ot「(銅(〜金)の草)」となっている。二次的な派生を経た表現であることから植物に関連する語彙層が希薄であった可能性は認められよう。このことは現代ヤクート語の語彙にも共通する特徴である。

## 4　その他の言語資料との比較

### 4.1　異文化語彙の受容について

ヤクート語の方言とされる言語にサハ・ヤクート共和国の北西部に隣接するクラスノヤルスク地方タイムル自治管区に居住するドルガン族の言語がある。ドルガン語はドルガン族の形成に関与した複数の民族(エベンキ族、ヤクート族、ロシア人等)がヤクート語を基盤として発展させた共通語といえる。ドルガン語(現在の話者約6000人)はヤクート語(現在の話者約380000人)のタイムル方言とみなされることもあり、文法上、大きな差異はない。このドルガン語の言語資料において動植物名称がどのように現れているかを観察したい。

ヤクート語に非常に近い関係にあるドルガン語の言語資料として1996年に刊行された新約聖書『マルコによる福音書』[5]がある。聖書の翻訳において興味深いのは、シベリアの諸民族にとってキリスト教用語をはじめとする異文化語彙、すなわち外来の語彙がどのように翻訳されたかという点である。ドルガン語訳は現代語訳である。したがって、現在のドルガン族は「伝統的ではない外来の事物」についての知識を十分に持っている、あるいは既に日常的にそれらに接している場合が多く見られるという状況での翻訳である。それを考慮したとしても翻訳文の語彙にはこの言語がどのように「異文化・外来の事物」を表現しているかが反映されているといってよい。このような点で聖書翻訳はドルガン語に限らず広くシベリアの諸民族言語の異文化受容のタイプについて示唆的であると考えられる。さらにそれに対応するヤクート語を同じく聖書翻訳の資料から掲げる[6]。この資料は1898年にカザンにおいて翻訳された4つの福音書であり、ドルガン語とは方言差の他に1世紀近くに及ぶ時間差があるので一概には比較の対象とはならないがいずれもシベリアに本来的な語彙ではない。これらの受容を表IIIに併記する[7]。

本稿ではヤクート語の動植物名称が現在のサハ族の居住地に見られるものを必ずしも多くは含んでいない

＜表III＞

1.「マルコによる福音書」ドルガン語訳より

| | | | | |
|---|---|---|---|---|
| 「駱駝」 | Dol. | verbljud hüöhü | 駱駝/動物 | (cf. Rus. verbljud「駱駝」) |
| 「羊」 | Dol. | ovets hüöhü | 羊/動物 | (cf. Rus. ovec「羊」) |
| 「ろば」 | Dol. | osäl hüöhü | ろば/動物 | (cf. Rus. osjol「ろば」) |
| 「はと」 | Dol. | golub' čïïčaak-tar | はと/鳥-pl. | (cf. Rus. golub'「はと」) |
| 「にわとり」 | Dol. | pätux čïïčaak | にわとり/鳥 | (cf. Rus. petux「鶏」) |
| 「いなご」 | Dol. | saranča kurd'aga | ばった/虫 | (cf. Rus. saranča「ばった」) |
| 「いちじく」 | Dol. | smokovnik mac | いちじく/木 | (cf. Rus. smokovnica「いちじく」) |

2.「マルコによる福音書」ヤクート語訳より

| | | | |
|---|---|---|---|
| 「駱駝」 | Yak. | täbiän kïïl | 駱駝/動物 |
| 「羊」 | Yak. | baran | 羊　＜ Rus.baran「(雄)羊」 |
| 「ろば」 | Yak. | osjol | 驢馬 ＜ Rus.osjol「ろば」 |
| 「はと」 | Yak. | golub' | 鳩　＜ Rus.golb'「はと」 |
| 「にわとり」 | Yak. | pätux | 鶏　＜ Rus. petux「鶏」 |
| 「いなご」 | Yak. | akridalaax | いなご+-laax「〜を有する」akrida ＜ Rus.「いなご」 |
| 「いちじく」 | Yak. | smokovnica | いちじく　＜ Rus.smokovnica「いちじく」 |

ことを上で述べた。サハ族のシベリアへの移動過程での周囲の環境の変化にもかかわらず『オロンホ』においては動植物名称がよく残っている。このようなヤクート語の動植物名称の状況がある一方で、新しく異文化語彙として受容するという状況での動植物名称がどのように翻訳あるいは借用されているかを聖書翻訳には見ることが出来る。

聖書のドルガン語訳にみられる特徴の一つとして次の点を上げることができる：動物名や植物名についてドルガン語にとって外来のものはロシア語語彙からの借用語彙となっている。特徴的なのはそれぞれの借用語彙に「動物」、「鳥」、「虫」、「木」といった説明的語彙(「総称」)が続けられている点である。

上記１のような外来の事物を表す借用語彙の受容の仕方がドルガン語に特に顕著なのか、他の言語にも見られるのか、あるいは聖書翻訳に特徴的なのかは今後さらに多くの事例を詳細に検討して行かねばならない。本稿で扱った『オロンホ』に現れる動植物等の名称にはロシア語からの借用形は皆無とは言えないまでもほとんど見られなかった。個々の言語資料の性格もあるであろうが、少なくともサハ族にとっては『オロンホ』の動植物を表す語彙はヤクート語にとってより本来的な語彙であったということが言えよう。

## 5　おわりに

以上、セットとして扱える語彙として動植物名称を『オロンホ』の中にみてきた。オロンホはヤクート語の中の古い語彙も残していると考えられる。一方で、それらの語彙が現代標準ヤクート語の中での使用という点からすればおのずと限定された性質をもっていることからして、オロンホの語彙を現代ヤクート語の語彙と同列に比較してよいかについては議論の余地がある。また、本稿ではロシア語からの借用についても若干言及したがこの点についてさらに深く詳細に渡り、見直す必要があろう。

本稿で扱う動植物名称についての語彙研究をもってこれまで色彩名称、方向表現と２回にわたり調査分析の対象としてきた『炎の乙女』についての語彙研究の区切りとしたい。今後、ヤクート語の語彙について総合的に見直し、これまでの分析で得られた結果をさらに多くのヤクート語資料に応用的につきあわせ、チュルク諸語の中で最北に位置するヤクート語の語彙から民族の移動にともなう言語接触の軌跡を辿っていきたい。

## 6　データ

(　)の数字は『オロンホ』のテキストに附されている行の番号である。各動植物名は原則として『オロンホ』から抜き出した形で挙げてある。対訳文のロシア語もそれに応じて示した。複数の箇所での出現はまとめて掲げた。

# 『動物名等』

**(獣類)**

**「馬」　　at, sïlgï, etc.**

(**at**)

[140]　xaamïïlaax atï
(140)　коню с быстрой поступью

[143]　säliik atï
(143)　коню с резкой рысью

[146]　d'oruo atï
(146)　коня-иноходца

[354]　at tappat
(355)　что и коню тащить не под силу.

[2057] bu kännittän atïgar taxsan
(2057) После этого, к коню своему подойдя,

[2082] anallaax ata baraxsan
(2082) Добрый конь, ему предназначенный,

[2138] atïn sïŋaaγïn
(2138) /Чугдаан Бухатыыр/, поводья коня своего

[2179] anï atï uoran
(2179) Еще коня украдет -

[2373] akkïn bihiäxä xaallar,
(2373) коня нам оставь,

[4577] xara turaγas attaax,
(4577) темно-карим конем

(**sïlgï**)
[672] sonoγos sïlgï kïajan ujbat
(672) такие, что молодому коню не под силу держать,

[1774] aas maγan sïlgïm
(1776) молочно-белой лошади

[1976] köpsö at sïlgï kölöhün allïar diäri,
(1976) /в таких, что/ захудалую лошадь пот прошибет,

[4550] sïlgï süöhü ajïïhïta,
(4550) Пусть покровитель лошадей

[4697] köpsö at sïlgï kölöhün allïar diäri
(4698) такие, что захудалого коня пот прошибет,

(**その他**)
[2147] xantaraŋnatan ihän
(2147) Уздой коня придержав,

「作業用の馬」 **kölö**
[1988] ütüö kölötün säbilläätä
(1988) доброго коня своего снарядил.

[2183] ütüö kölölörün
(2183) добрых коней

[3270] siälläax salama ïjaammït
(3270) *саламой* из конской гривы увешанного,

**「雄馬」　atïïr**
[3407] xaxïdal xara atïïrdardaax
(3404) /где/ отощавшие черные жеребцы

[3662] kïtalïk älämäs atïïrdaax,
(3663) белопеними жеребцами владеющего,

**「雌馬」　biä**
[3404] biäγä turbatax
(3407) с кобылами не спаривающиеся,

[3663][4480][4491]　d'ällik älämäs biä
(3664)(4480)(4491)　вольная пегая кобылица

**「子馬」　kulun**(生後初めての秋まで)
[1403] ulaan-kulun kurduk,
(1403) жеребеночек светло-серый,

[2127] kulunnuu ojutan,
(2127) жеребенком скача,

**「灰黄で尾とたてがみが黒い馬」ulaan d'aγïl at(-taax)**

[921][3936][3999][4534][4572][4848][4846] ulaan d'aγïl attaax
владеющий буланым, с /белым/ оплечьем конем,

**「馬つなぎ」　särgä**
[199][1108][4258]altan särgä
медная коновязь поставлена,

**「ライオン」　xaxaj**
[407] xaxajdaax xald'aajïlarïn xaŋas öttünän,
(407) Южный косогоп со львами слева обойдя,

**「牛」　ïnax**
[3400] ïnaγï bappatax
(3403) с коровами не случающиеся,

**「雄牛」　　oγus**

[496]　kur oγus kuolajïn
(499)　вытянутым пищеводам откормленных быков,

[3403]　büörtük oγustardaax,
(3400)　туда долетело, где неплодовитые быки

**「ヘラ鹿」　　tajax**

[612]　buur tajaγï
(612)　матерого лося-самца

**「熊」　　ähä**

[614]　xardaŋ ähäni
(614)　грозного бурого медведя

**「狼」　　börö**

[999][1246][3124]sur börö buolannar,
в серых волков превратились,

[1206]　sur börölörö
(1206)　волки серые,

[1301]　sur börölör tus innilärigär kiirän,
(1301)　перед серыми волками

[2128]　börölüü xarbatan,
(2128)　волком несясь,

**「クロテン」　　kiis, saarba**

[1324]　üüs-kiis täriitinii mölböjüŋ diän
(1324)　как соболий мех, ранежьтесь,

[1853][2407]üüs kiis täriitä ölbürgäläax,
из соболиных шкурок-огузков сшитую,

[1862]　xara saarba täriitinän
(1862)　мехом черного соболя

**「獣」　　kïïl**

[2714][3062]Käj Suorun kïïla buolan,
в зверя необъятного гулкого неба

[4598]　tüört uon tüört kïïl
(4598)　у сорока четырех видов зверей

[4648] bili kïïl diäki körön turan,
(4648) в сторону того зверя глядя,

[4682] bili Öksökü kïïl
(4682) зверь Ексекю

「獲物」 **bullun**
[608] xara tïa bullun
(608) зверей из темных лесов

[610] küöx tïa bullun
(610) зверей из зеленых лесов

「白い毛皮のもの」**ürüŋ tüülääx**
[1092] ürüŋ tüülääγä
(1092) белошерстных,

[1205] ürüŋ tüülääγïn önöjön körbütä
(1205) на белошерстных своих глянул:

[1796] ürüŋ süürükpütün ölördülär,
(1796) наших белых бегунцов погубили,

[1874] ürüŋ-xara tüülääγï
(1874) /Хотя/ белошерстным и черношерстным нашим

「黒い毛皮のもの」**xara tüülääx**
[1204] xara tüülääγïn xajïhan,
(1204) на черношерстных своих обернулся,

[1797] xara süürükpütün xamïjdïlar.
(1797) наших черных бегунцов забрали;

[1874] ürüŋ-xara tüülääγï
(1874) /Хотя/ белошерстным и черношерстным нашим

「走るもの」 **süürük**
[1247] suban süürükpün sujdaan ärällär,
(1247) вольных бегунцов моих уничтожать стали,

[1248] ürüŋ süürükpün ölörön ärällär.
(1248) белых бегунцов моих умерщвлять стали,

[1285] ürüŋ xara süürüγün
(1285) на белых и черных своих бегунцов

[1562] ürüŋ süürüktärä
(1562) Белые бегунцы от этого

[1568] xara süürüktärä
(1568) черные бегунцы

[2466] xara süürääk xartalaax bïarïn,
(2467) Печень с кошкой-*харта* черных бегунцов,

[2467] ürüŋ süürük ürgünnäxtääxsüräγin,
(2468) жиром заплывшие сердца белых бегунцов,

[2514] ürüŋ süürükpün ölördülär,
(2514) белых бегунцов моих побили,

**「家畜」　　süöhü**

[395] ïtïk näriitä ïstaannarïn ann'ïnan,
(395) штаны из шкуры жертвенной скотины надев,

[396] sut täriitä sutuoralarïn ugunan,
(396) наколенники из шкуры павшей скотины напялив,

[3918] iitär süöhünü kürüölää,
(3918) скоту плодящемуся загон устрой,

**（鳥類）**

**「鶴」　　turuja**

(119) /там, где / журавль,
[119] turuja kïïl

(4052) журавля превратясь,
[4052] turuja kïïl buolan

[4357] turuja kïïl čonojo kötön
(4358) такой, что журавль, высоко /в поднебесье/ летая,

[4633] toγus suban turuja uolattarïn
(4633) он девять парней холостых, журавлям подобных,

[4649"] (suban turuja uol)
(4649") (холостой парень-журавль):

[4661] toγus suban turuja uolattarabït.
(4661) девять подобных журавлям холостых парней

[4692] toγus suban turuja uolattar
(4692) девять подобных журавлям холостых парней

「ソデグロ鶴」 **kïtalïk**
[124] kïtalïk kïïl
(124) /там, где / стерх,

[1096] kïtalïk kïïl
(1097) такой, что стерх,

[4354] kïtalïk kïïl kïrïja tötön
(4355) такую, что стерх, низко /над землей/ пролетая,

「鴫(しぎ)」 **bïld'ïrïït**
[174] bïstïbat bïld'ïrïït diän
(174) стаи несчетные куликов

「郭公」 **käγä**
[176] kärääbät käγä diän
(176) неугомонные кукушки

[4291] kärääbit käγätä käpsään barda,
(4291) онемевшие было кукушки закуковали,

「雉鳩」 **öt**
[178] örööböt ötön diän
(178) неутомимые горлицы

「ヤマドリの雄」 **ärdäγäs**
[4160] ärdäγäs ular tüöhün tüütün kurduk
(4160) на пестро-пятнистом, словно грудное оперение тетерева,

「黒ガモ」 **andï**
[4164] andï kïïl xabaγan tüütün
(4165) оперение турпана,

「雷鳥」 **xabd'ï**
[4436] xabd'ï saγa buoluox,
(4436) с куропатку величиной будет выглядеть;

「ヨーロッパ大雷鳥」**ular**

[4432] ular saγa buoluox,
(4432) с глухаря величиной представится,

「ハシボソガラス」**suor**

[4434] suor saγa buoluox,
(4434) с ворона величиной увидится,

「ホオジロガモ」**arïlïas**　(**oruluos** のa化バリアント)

[1300] arïlïas kus kïnatïn tïahïn kurduk, kuhuguraan,
(1300) словно крыльями утки-гоголя, со свистом воздух рассекая,

[4700] aγïs arïlïas maŋan kïrgïttar kälännär,
(4700) восемь девушек, подобных белым гоголям.

[4700] aγïs arïlïas maŋan kïrgïttar,
(4715) "Восеми моих девушек, подобных белым гоголям,

[4720] arïlïas maŋan kïrgïttar
(4720) девушки, подобные белым гоголям,

「ユキホウジロ」**tulluk**

[1739] tulluk kïïl taba täbimmätäx
(1740) где бы и пуночка поскользнулась, пройдя,

[4695] tulluk kïïl taba täbimmätäx
(4696) такому, что и пуночка поскользнулась бы, проводили,

[4872] tulluk kïïl taba täbimmätäx
(4873) такой гладкий, что пуночка поскользнется, вышла;

「鳥」　　**kïïl**

[4613][4959]öksökü kïïl buolan
птицу Ексекю превратившись,

[4641] üs bastaax öndölüjär öksökü kïïl
(4641) трехглавая огромная птица Ексекю

「小鳥」　　**čïïčaax**

[3931] kömüs čïïčaax buolan,
(3931) в серебряную птичку превратясь,

[4854] kömüs čïïčaaxpïtïn
(4854) "Серебряную пташку свою

（魚類）

「フナ」　**bïččïï**

[346]　bïččïï bahan

(346)　/даже/ голову мелкому карасю

「ヒメハヤ」　**mundu**

[403]　baspatax mundu miinin kurduk

(404)　что недоваренной ухе из рыбок-гольянов подобна;

「カワメンタイ」**sïalïhar**

[1323]　sïalïhar bïarïnïï sïmnaaŋ,

(1323)　как налимья печень, размягчитесь,

「魚」　**balïk**

[1919]　balïktaayar käläγäjdik oloron,

(1919)　молчаливее рыб мы жили.

[1996]　Ölüü D'iribinäj balïgïn

(1997)　рыбы Елюю Джирибинэй,

[2003]　balïksït ïal balaγanïn tahaaran

(2003)　похожими на рыбацкие юрты,

[3255]　üs luo balïk öhüötün

(3255)　/ее/ верхние матицы из трех рыб-*.tyo*

（昆虫類等）

「ウジ虫」　**čiärbä**（<Rus. červ'「蠕虫、ウジ虫」）

[409]　čiärbäläax Kiläjä Xaan kiriliästärin

(409)　по усыпанному червями крыльцу Килэйэ Хаан

「(水棲の) 甲虫」**čoxu**

[411]　čoxulaax Čuguja Xaan suollarïn

(411)　по усеянной водяными жуками дороге Чугуйа Хаан

「蛙」　**baγa**

[413]　baγalaax Čuguja Xaan aartïktarïn

(413)　по кишащему лягушками проходу Чагыйа Хаан

# 『植物名等』

## (樹木類)

「木」　**mas**

[61] oxton üünär mastaax,
(61) с падающими и вновь вырастающими деревьями,

[163] sihii üöt mastaradaax äbit,
(163) красные ивовые деревья есть у нее, оказывается;

[223] oxton köγüräääbät mastaax
(223) с падающими, но не убывающими деревьями,

[1347] kiil mas kurduk
(1347) подобно креневому дереву

[1349] čoruun mas kuruduk
(1349) подобно крепкому дереву

[1539] kiil mas kurduk,
(1539) подобно креневому дереву

[1980] baaraγaj mahïnan oŋohullubut
(1980) из могучего дерева сооруженного,

[2004] mas maaltaar,
(2005) [такие] деревянным опереньем,

[4237] ot-mas iččilärä
(4237) Духи-хозяува трав-деревьев,

「灌木」　**ïarγa (mas)**

[151] satïï ïarγa mastardaax äbit,
(151) низкий кустарник есть у нее, оказывается;

「白樺」　**xatïŋ, čäčir, tuos**

[156] xatïŋ mastaradaax äbit,
(156) березы есть у нее, оказывается;

[1110] araŋas čäčir saajïllïbït,
(1110) желтые березки-чэчир вокрук воткнуты;

[1991] xatïŋa xatïŋnaax,
(1990) сделанный из ствола березы

[1993] tuoha tuostaax,
(1994) из нароста креневого дерева

[2006] tuos tällämäättääx
(2006) оперенные берестяным опереньем,

[3095] türbälääx tuos kuruduk,
(3095) словно берестяные свитки,

[4302] käxtibit čačirä
(4302) пожухлые молодые березки

[4506] arïŋa tuohunan
(4507) из слоев бересты

[4707] aγïs xartïgas arïŋa tuohunan
(4708) из восьми слоев бересты

「白樺製の桶」 **xolloγos**
[3130] xolloγos bïhaγahïn saγa
(3130) величиной с половину берестяной бадьи

「白樺製の手籠」**atïjax**
[4895] atïjaxtaax uu kurduk
(4895) словно воду в берестяной чаше,

「ハイマツ」 **bolbukta**
[1477] d'aaŋï bolbuktatïn
(1477) будто горный стланик-кедрач,

「樅」 **xarïja**
[167] toŋ xarïja mastardaax äbit,
(167) мерзлые ели есть у нее, оказывается;

[2073] kömnöxtööx xarïja aŋaarïn saγa,
(2073) плетью, половине заснеженной ели равной,

「ヤナギ」 **üöt, talax**
[3836] ikkiännärin bili üöt ikki salaatïgar
(3836) затем к двум стволам ивы той,

「枯れ木、枯れ枝」**xappït amïnn'ïk**
[3840] xappït amïnňïkïnï
(3840) кучу сушняка-хвороста

「針葉樹」　**kötöγö**

[3844] kübä tïa kötöγötün killärän
(3844) хвою предгорного леса принесла,

「倒れた木」　**sis tïalara, siŋ**

[3802] sis tïalara
[3803] siŋnän bütüülärigär,
(3803) буреомом-валежником кончается,

「密林」　**xaraŋa tïalar**

[1569] xaraŋa tälar
(1569) по дремучим дебрям

（草類）

「草」　**ot**

[142] xampa n'aarsïn ottoox äbit,
(142) шелковисто-мягкая трава есть у нее, оказывается;

[145] Siräm küöx ottoox äbit,
(145) густая зеленая трава есть у нее, оказывается;

[148] Kokuora küöx ottoox äbit,
(148) изумрудная трава-*локуора* есть у нее, оказывается;

[344] хороŋ otu
(344) колкую траву

[4238] ot annïnan obugunaspït
(4238) под травами шустро бегающие

「ルクオラ（鮮やかな緑の草）」　**lokuora**

[1074] lukuorata turugurbut,
(1074) /В долине/, травой-*локуорой* богатой,

[4251] lokuora okkut torolujdun!
(4251) трава-*локуокра* щедро произрастает!

[4298] lokuora küöx otunan
(4298) травой-*локуорой*

「草の茎」　**ot saγa tutuluga**

[2154] ot saγa tutuluga suox
(2154) не имея подпорки даже с былинку,

「葉」　**säbirdäx**

[3992] xampa säbirdäyä xagdarïjbït,
(3992) нарядные листья увяли

[4242] säbirdäx annïnan sibiginäspit
(4242) под листьями щебечущие

[4253] xampa säbirdäxtärin ajgïrattïnar,
(4253) шелковисто-зелеными листьями зашумят!

[4301] xampa säbirdäyin anňïnna,
(4301) шелковистыми листьями убрались,

「根元」　**tört**

[1068] xarïalaax tördügär
(1068) за обгорелым комлем ее схоронясь,

「尾状花」　**kiistä**（「筆」の意）

[4303] kärä kiistätin kättä,
(4303) чудесные сережки одели;

「いばら」　**adïrïïn**

[661] adïrïïnnaax aartïktarï
(661) Тернистые пути-дороги

「スゲ」　**kïrïs, ača**

[1027] arïï sahïl kïrïstaax,
[1028] ača ot arajdaax,
(1028) с клиньями из травы-осоки,

[1075] ačata aatïrbït,
(1075) травой луговой избильной,

[1960] ača ot bïïstarajdaax,
(1960) с межами осики,

[4249] ača küöxxüt namïlïjdïn,
(4249) осока-трава густо стелется,

「タンポポ」　**altan ot**

[1029] altan ot bïïstarajdaax,
(1029) с межами из одуванчиков,

[1078] altan ot d'arγalaax,
(1078) одуванчиками усеянной,

[1959] altan ot d'arγaalaax,
(1959) /в тот/ с узорами из одуванчиков,

[4293] altan ot bïïstarajdaax
(4293) желтыми одуванчиками пестреющее,

**「ヤマナラシ」 tätiŋ**
[1066] uolattar subu tätiŋ
(1067) этой осины

**「(有用な)野草」sirämä**
[1077] sirämä siriädijbit,
(1076) травой злаковой изобильной,

**「葦の茎」 tïn'ax**
[1750] aarïktaax altan tajaγïttan tardïstan oloron
(1750) на медную трость с погремушками опираясь,

[1756] ajgïrastaax tajaγar tardïstan oloron
(1756) на трость с побрякушками опираясь,

**(その他)**
[171] loskuj ojuurdardaax äbit,
(171) есть колки у нее, оказывается;

[3525] biäs bïlastaax ot tügäγin saγa
(3525) /величиной/ с основание пятисаженного стога

[4300] xagdarïjbït tïata,
(4300) пожелтевшие леса-рощи

## 7 転写・略号・注及び参考文献

転写
а → a, б → b, в → v, г → g, д → d, е → e, ё → jo, ж → ž, з → z, и → i,
й → j, к → k, л → l, м → m, н → n, о → o, п → p, р → r, с → s, т → t,
у → u, ф → f, х → x, ц → c, ч → č ш → š, щ → šč, ы → ï, ь → ', ъ → ", э → ä,
ю → ju, я → ja, また、さらに ҷ → d', і → i, γ → ü, һ → h, ө → ö, ҽ → e, ҕ → γ,
ҥ → ŋ, ҥь → n'
長母音はそれぞれ aa, ïï, ää, ii, oo, uu, öö, üü と表記した。

**略号**

Yak．ヤクート語，Dol．ドルガン語，Rus．ロシア語

**注及び参考文献**


[1] 本稿では言語名については慣用に従い「ヤクート語」とする。

[2] 藤代　節，1996，「ヤクート語語彙研究(1)：色彩名称」（以下「語彙研究(1)と略す」），学術情報センター紀要第8号，学術情報センター，東京，pp.155-189.

[3] 藤代　節，1997，「ヤクート語語彙研究(2)：方向表現」（以下「語彙研究(2)と略す」），学術情報センター紀要第9号，学術情報センター，東京，pp.113-130.

[4] *Kïïs Dăbilijă*, *Jakutskij geroičeskij ăpos*（『オロンホ』），Nauka，Novosibirsk，1993.

[5] *Dolgan-haka tïlïnan tulmaas*（ドルガン語訳マルコによる福音書），Moskva:Biblijanï tulmaastïïr Institut，83pp.＋2.

[6] Markattan svjatoj evangelie（マルコによる福音書），*Svjatoe Evangeilie na jakutskom jazïke*（ヤクート語訳聖書），1898，Kazan'（Reprinted; 1994，Grecija : Izdanie Monastïrja Paraklita），pp.81-130.

[7] 藤代　節，1997，「ドルガン語訳新約聖書『マルコによる福音書』」，『環北太平洋の言語』第3号、宮岡伯人・津曲敏郎　編、京都大学大学院文学研究科、京都、pp.183-202.

学術情報センター紀要　第10号(1998)

研究論文

# 知識集約型工学の建築設計への応用

# Aplication of Knowledge Intensive Engineering to Architectural Design

学術情報センター　吉岡　真治
Masaharu YOSHIOKA
National Center for Science Information Systems

**要旨**

近年の資源の有限性に関する問題や環境問題などのために、製造業に携わる人々は、各々が作る製品をより多くの様々な観点から評価する事が求められている。これらの問題に対応するために、我々は、製品のライフサイクル(設計・生産・保守・リサイクルなど)に関連する情報や必要な知識を柔軟かつ統合的に利用可能とする知識集約型工学(Knowledge Intensive Engineering)を提案しており、その計算機上へ実現したシステムである Knowledge Intensive Engineering Framework (KIEF)システムを作成している。本論文では、まず、知識集約型工学および KIEF システムについて述べる。さらに、この KIEF システムの有効性を検討するために本システムの建築設計への応用を試みる。そのために、建築設計に関する知識をどの様に本システム上の知識として表現できるかについて考察する。最後に、現在 KIEF システム上に構築中の知識ベースについて述べ、KIEF システムの有効性について議論する。

**ABSTRACT**

Because of the finiteness of natural resources, environmental problems and so on, engineers are expected to evaluate their products from ever more various kinds of aspects. To support such claim, we proposed the concept of knowledge intensive engineering, in which various kinds of knowledge is used in a flexible and integrated manner in order to aim at generating more added-value. In addition, we also proposed Knowledge Intensive Engineering Framework (KIEF) system that forms a computational framework of knowledge intensive engineering. In this paper, I describe the concept of knowledge intensive engineering and the KIEF system. After that, I apply the KIEF system to architectural design to show the value of the system. To do so, I analyze knowledge that is used in architecture design and discuss how to implement the knowledge on the KIEF system. Finally, I show some results of this application and discuss the capability of the KIEF system.

[キーワード]　設計知識、CAD、知識集約型工学
[Keywords]　Design Knowledge, CAD, Knowledge Intensive Engineering

## 1　緒言

従来の製造業においては、「自然環境あるいは自然技術を技術によって人工的な技術製品に変換し、人間の生活環境をより快適に、安全に、そして利便的なものへ変化させる」ことを目的として多くの人工物を産み出し、我々の生活はそれにともない豊かになってきた。しかし、その一方で、これらの人工物の数が増大し、氾濫することにより、資源の有限性に関する問題、環境問題などの新たな問題を引き起こしている[1]。

これらの問題に対応するためには、従来のように、問題を細分化して、各々の問題に対して、その問題が属する領域固有の知識のみを用い、個別に最適化を図るような古いタイプの問題解決の方法ではなく、様々な領域で用いられる知識を総合的に利用するような問題解決が望まれる。そこで、筆者らは製品のライフサイクル(設計・生産・保守・リサイクルなど)に関連す

る情報や必要な知識を柔軟かつ統合的に利用可能とする知識集約型工学(Knowledge Intensive Engineering)[2]を提案しており、その計算機上へ実現したシステムであるKnowledge Intensive Engineering Framework (KIEF)システムを作成している[3]。このKIEFシステムでは、製品に関する知識を表現するために必要な概念に関する定義であるオントロジカルな知識ベースとモデル・ベース推論の技術を基に、情報や知識の共有を行う。

一方、建築設計では、建築基準法が性能規定型の基準を用いるような改正が検討されていることに伴い、建築物の性能を積極的に評価する枠組が求められており、[4]などにおいてその影響が議論されている。そこで、本研究では、この建築設計の支援システムを、KIEFシステムの応用システムとして作成することにより、KIEFシステムの妥当性の検証を行う。そのために、本論文では、まず、最初にKIEFシステムについて概観した後に、建築設計に関する知識の分析とKIEFシステムへの実装可能性について議論し、最後に現在KIEFシステム上に構築している知識ベースについて述べ、KIEFシステムの有効性および必要とされる改良点について考察する。

## 2　知識集約型工学

設計や生産などの製品のライフサイクルに関わる様々な作業は、製品に関する様々な情報から各々の作業に必要な情報を抽出し、その情報を基に、新しい製品に関する情報を作り出す過程と考える事ができる。知識集約型工学では、これらの作業において必要とされる情報をモデルとして表現し、その各々のモデルを統合的に管理することにより、これらの作業を支援しようとする枠組である(図1)。

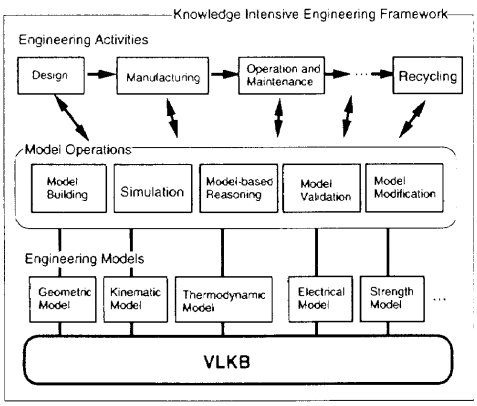



図 1　知識集約型工学

この様に様々な領域を扱うためには、非常に大規模な知識ベースが必要になる。これに対し、知識ベースの規模が大きくなると知識ベース中の各々の知識の整合性の管理が困難になるという問題がある。そこで、この知識集約型工学を計算機上で実現したKnowledge Intensive Engineering Framework (KIEF)システムでは、各々の詳細な知識ベースについては、ある程度、適用領域を絞ることによってサイズを限定した整合性のある知識ベースシステムとして整理することにする。また、これらの知識ベースを組み合わせ、大規模な知識ベースとして運用するために、これらの知識ベースで用いられている概念に関する定義であるオントロジカルな知識ベースを作成する。さらに、このオントロジカルな定義を元に各々の知識ベースシステムを統合することにより、全体として大規模知識ベースシステムとしての運用を行う。そのため、このKIEFシステムには次の2つの要素から構成される。

- 大規模知識ベース
  異なる作業で用いられる各々のモデルにおいて表現される概念に対してオントロジカルな定義を与える。
- モデル管理機構
  上記の知識ベースに基づき、異なるモデル間のデータ交換や情報の整合性を管理する。

KIEFシステムでは、前者に対し、物理的な概念に対してオントロジカルな定義を与える工学知識ベース[5]を、後者に対してプラガブル・メタモデル機構による統合モデリング環境[6]を利用する。

以下では、これらの要素について概観する。

### 2.1　知識ベースシステム

工学における学問領域は、機構学、材料力学、電子工学等の幅広い学問領域に細分化されているが、各々の学問領域の中心となる知識は一般的な物理学の知識として扱われているものが大半である。そこで、我々は、物理法則や物理現象に関する知識を中心とした知識ベースの構築を行っている[5]。この知識ベースシステムでは知識を次の3つのレベルに分けて整理する。

**概念辞書**　実体、関係、物理現象、属性、物理法則といった概念に関するオントロジカルな定義

**フィジカル・フィーチャ**　機構とその機構を支配する物理現象の組合わせにより表現される機構ライブラリ

**モデル・ライブラリ**　物理現象などに関して、様々な

モデリングシステムでの表現方法を記述したライブラリ

この中で、フィジカルフィーチャは、概念辞書に表現された概念の組み合わせとして表現され、設計対象はこのフィジカル・フィーチャを組み合わせることによりモデル化される。また、モデル・ライブラリは概念辞書において定義された概念と対応づけらている。このモデルライブラリは、フィジカル・フィーチャの組み合わせとして表現されている設計対象を各々のシステムの上でモデル化する際に、対応する要素を集めることによりモデル構築の支援を行うことができる。図2は回転伝達機構についてこの3つのレベルの知識間の関係を例を用いて示したものである。

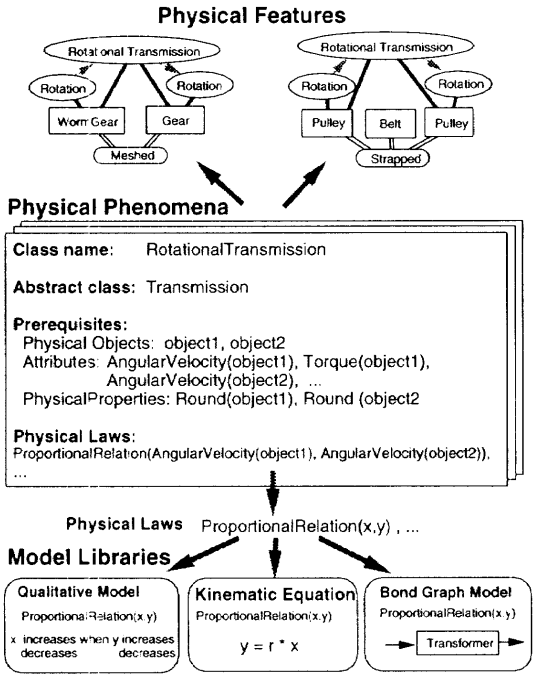



図 2　概念辞書、フィジカルフィーチャ、モデルライブラリの関係

## 2.2　プラガブル・メタモデル機構

プラガブル・メタモデル機構とは、設計対象を表現するために必要な様々な概念とその概念間の関係を表現したモデルであるメタモデルを中心として、複数の設計対象をモデリングするツール(以降ではモデラと呼ぶ)を統合するシステムである。

このプラガブル・メタモデル機構では、各々のモデラに関する知識を記述し、プラガブル・メタモデル機構と各々のモデラとのデータ交換のインターフェースを作成することにより、既存のモデラを統合することができる。このモデラに関する知識の定義を表1に示す。

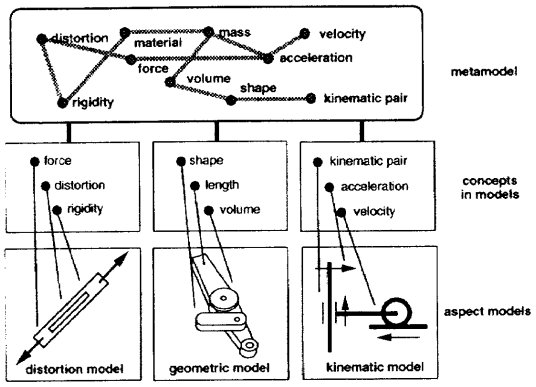



図 3　メタモデル

表 1　モデラに関する知識

| 名前 | 内容 |
|---|---|
| 関連する概念 | 概念のリスト |
| モデリングに利用する概念 | 概念のリスト |
| モデリングを操作した結果得られる概念 | 概念のリスト |
| 属性データの変換方式 | 関連する属性を表現するグラフと変換の手続き |

プラガブル・メタモデル機構はこの知識を用いることにより、各々のモデラにおけるモデリングの支援を行う。このプラガブル・メタモデル機構において、モデリングの作業は定性的な概念間の関係のモデルを作る定性モデルの構築の段階とそこで作成した定性モデルを元に各々のモデラ上でモデルを作る段階に分けることができる。

1. 定性モデルの構築
   (a) モデルの抽象化
   モデルの抽象化とは、メタモデルとして表現されている設計対象をモデラが扱うことの出来る概念と対応づける操作である。プラガブル・メタモデル機構においてこの操作は、次の2段階の操作として実現する。まず、最初の段階はメタモデルに表現されている概念から、関連する概念の記述を利用することにより、フィルタリングを行う。次に、このフィルタリングを行ったモデル

に対し、モデラに関する知識に記述されているモデラで用いる概念と対応づける。例えば、ロボットのアームをはりモデラで強度解析をする例を考えると、最初のフィルタリングの段階の操作は、ロボットのアームを駆動するモーターの電機の流れなどといった現象をモデル化の対象からはずすことに相当する。また、つぎの段階は、ロボットのアームという概念をはりモデラが扱うはりという概念に対応づける操作に対応する。

(b) モデルの簡略化

モデルの簡略化とは、計算コストなどを考えて、設計対象の一部をモデル化の対象からはずす操作である。プラガブル・メタモデル機構においてこの操作は、メタモデルに表現されている概念の内、モデル化の対象からはずすものを選択する操作と、モデラが扱える概念以外の概念を無視する操作として実現する。先ほどのロボットのアームの例で考えると、アームの自重が軽い場合にそれを無視したり、ロボットの本体の部分については、その接続部分以外を無視したりする操作に対応する。

2. 各々のモデラ上でのモデル構築

(a) データの交換

モデラに関する知識の概念間の対応関係と変換手続きを利用することにより、メタモデル上に表現されている属性情報などのデータを獲得し、モデル作成に利用する。また、必要とされるデータが存在しない場合は、プラガブル・メタモデル機構がモデラ知識の利用した結果得られる概念に関する記述を利用する事により、適切なモデラを選択し、必要なデータの計算もしくは入力を促す。先ほどのロボットアームの例で考えると、はりの長さを計算するのにソリッドモデラもしくは2次元CADなどの形状に関する情報を扱えるモデラを選択し、データが存在しない場合にはデータの入力、そしてデータの変換を行う。

また、プラガブル・メタモデル機構はこれらの全ての操作をモデリング作成の履歴として保存する。この情報は、設計者は各々のモデラを利用して得られたデータに関するモデル化の仮定であり、この仮定の正当性を評価することは、結果の正当性の検証に役立つと考えられる。また、データ交換の際の関係も同様に保存し、この関係を用いることにより、データ変更の際の整合性管理に利用する。

### 2.3 KIEFシステム

KIEFシステムでは、オントロジカルな知識ベースとして概念辞書を用い、概念間の関係を表現する知識ベースとしてフィジカル・フィーチャ、プラガブル・メタモデル機構に接続するモデラで用いる知識(モデル・ライブラリを含む)を用いる事により、大規模な知識ベースを取り扱う(図4)。

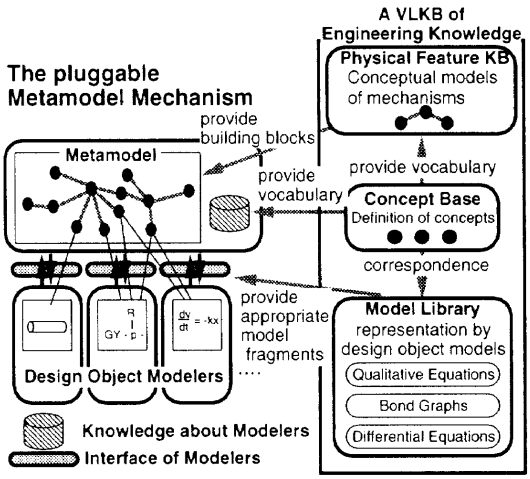



図4 Knowledge Intensive Engineering Framework

このKIEFシステムを利用した設計対象モデルの構築および評価のプロセスは次のような手順になる。

1. フィジカル・フィーチャの組合わせによる初期メタモデルの作成

機構ライブラリであるフィジカル・フィーチャを組み合わせることにより、設計対象の構造および、そこで考慮する物理現象に関するモデルを作成し、それを初期のメタモデルとする。

2. フィジカル・フィーチャを用いた物理現象の導出

フィジカル・フィーチャはある構造が存在した際に起る物理現象に関する記述とみなし、初期に作成したメタモデルでは考慮していなかったが生起する可能性のある物理現象を導出する。

3. 複数のモデラによる評価

プラガブル・メタモデル機構を利用し、複数のモデラを利用し、その結果をメタモデルに反映させることにより、複数のモデラに記述されている知識を適用し、設計対象を評価する。

現在の、KIEFシステムは、VisualWorks2.5の上に構築されており、機能として与えられる設計仕様から1.の段階のフィジカル・フィーチャの組合わせを支援する機能設計支援ツールFBSモデラ[7]、定性推論に基づく物理シミュレータ、数値解析に基づく物理シミュレータ、はりモデラ、2次元形状モデラ、ソリッドモデラ(DESIGNBASE)などが接続されている。

## 3 KIEFシステムの建築設計への応用

建築設計は、建築基準法が従来の仕様規定型基準から性能規定型基準へ改正されようとしているのに伴い、従来の法律で規定されている問題点を洗い出して問題解決をする設計から、積極的に問題点を洗い出し、その性能を明らかにするような設計へと移行することが求められている[4]。この様な状況に対応するためには、本研究で扱っているような知識ベースシステムを用いた設計支援が有効であると考えられる。そこで、本研究では、この応用により、KIEFシステムにおいてどの様な知識が記述でき、どの様な点を改良する必要があるのかについての検討を行う。

### 3.1 建築設計に関する知識

本研究では、実際の建設会社の研究者へのインタビューを通じて知識獲得を行っている。また、インタビューした研究者が主に扱っている分野が建築設計の中の建築基礎の設計であるために、その分野に関する知識を重点的に扱うことにした。

建築基礎の設計の方針に関する知識は、日本建築学会により[8]の様にまとまった形で本となり出版されている。しかし、インタビューの結果、この様な形式の知識には、次のような問題点があることが指摘された。

- 知識の適用基準に関する記述が不備である。
  設計上に必要な様々な解析方法が紹介されているが、どの様なときにどの様な解析方法を適用すれば良いのかに関する記述が不足している。特に、一つの問題に対して複数の解析手法が存在する場合に問題になる。
- 知識のブラックボックス化が問題である。
  解析手法の背景にある知識間の関係などが不明確でブラックボックス化しているために、無意味に一部だけを詳細に解析したり、入出力の結果を適切に評価できない場合がある。
- 検討項目の洗い出しが煩雑である。
  この枠組では、フローチャートにしたがって全ての検討項目について検討して行くことを進めているが、実際の設計では全てを検討する必要性がないために、この様な手順を踏むことは煩雑である。

### 3.2 建築設計に関する知識のKIEFシステムへの実装

先に述べた問題点を解決するためには、知識ベースの構築にあたり、次の項目について考慮する必要であると考えられる。

1. 知識間の因果関係を明確に表現する。
   各々の知識について、単独の現象についての知識として捉えるだけでなく、現象間の因果関係を明確に表現することにより、各々の知識間の関係を明らかにする。
2. 設計対象に応じた検討項目の洗い出しをする。
   現象の因果関係の知識を用いることにより、設計対象に応じた検討項目を選びだせる枠組が必要である。
3. 洗い出した検討項目に対して解析手法を提示する。
   検討すべき現象に対して解析手法を提案する。この時、複数の解析手法が存在するときには、そこから選択を行うための指針を与える必要がある。

これらの項目についてKIEFシステムでの実装可能性を順に検討する。まず、1.の因果関係の記述であるが、これは、KIEFシステムにおけるメタモデルによる設計対象を表現する概念間の関係の表現の枠組やフィジカル・フィーチャを用いた現象の因果関係の表現を利用することにより記述が可能であると考える。また、2.の検討項目の洗い出しについてはフィジカルフィーチャを用いた物理現象の導出の枠組を用いることにより実現可能であると考える。

しかし、3.の解析手法の提示については、KIEFシステムにおいてそのままの形で扱うことができない。そこで、この知識を実装するためにどの様な知識表現の枠組が必要かを考察する。この解析手法の提示の操作の実装については、解析手法に関する知識を統合するための部分と解析手法の選択を支援する部分の大きく

2つの部分に分けることができる。この内、前者については、各々の解析手法をプラガブル・メタモデル機構に接続する外部モデラとして扱うことにより実装が可能である。後者については、メタモデルとして表現されている概念間の関係に基づいて解析手法を提案するための枠組が必要になる。この知識を記述するためには、解析手法に関する知識を記述する枠組であるモデラに関する知識の記述の枠組を拡張する必要がある。

以上の考察に基づき、今回の建築設計に関する知識をKIEFシステムの各々知識ベースの適切な部分に振り分けると次のようになる。

**概念辞書**　建築設計に関連する実体、物理現象などの概念を知識として登録する。

**フィジカル・フィーチャ**　物理現象の因果関係を記述し、知識間の因果関係の明確化や検討項目の洗い出しに利用する。

**モデル・ライブラリ**　各々の解析手法を物理現象と対応づけて整理する。

**モデラに関する知識**　解析手法の特徴などを明記することにより、解析手法の選択を支援する。

### 3.3　建築設計支援システムのための知識ベース構築

前節に述べた知識ベースに関する分析を元に建築設計支援システムのための知識ベース構築を行っている。そのために、建築に関連する物理現象として、以下のような物理現象に関する概念定義と現象の因果関係を記述したフィジカルフィーチャを作成した。

**自然現象**　地震、地震の伝播、雨、風、地下水位の変化、…

**地盤の変化**　地割れ、沈下、地滑り、液状化、地盤の支持力の喪失、…

**構造物の崩壊**　杭の破壊、基礎スラブの破壊、梁の破壊、…

**建物の変化**　建物の傾き、建物の振動、…

また、各々の現象を評価するための解析手法について、プラガブル・メタモデル機構におけるモデラに関

表 2　解析手法に関する知識の整理

| スロット名 | 内容 | |
|---|---|---|
| 現象名 | 現象の名前 | |
| 評価方法<br>(繰り返し可) | 手法の説明 | 手法に関する記述 |
| | 入力パラメータ | 評価手法で必要とするパラメータ |
| | 出力パラメータ | 評価の結果得られるパラメータ |

表 3　地震に関する記述例

| 現象名 | 地震 | |
|---|---|---|
| 評価手法 | 手法の説明1 | 既定値としての設計加重を利用(簡易法)<br>想定する地震の強さにより2レベルを設定<br>地表面最大加速度レベル1 250Gai、レベル2 500Gal<br>地表面最大速度レベル1 25cm/S、レベル2 50cm/S |
| | 入力パラメータ | なし |
| | 出力パラメータ | 地表面最大加速度、地表面最大速度 |
| | 手法の説明2 | 設計用入力地震動の利用<br>従来の波形記録を利用 |
| | 入力パラメータ | なし |
| | 出力パラメータ | 波形データ(地盤および構造物の加速度、速度、変位の時刻歴) |

する知識として記述するために、表2に示すフォーマットで整理した。また、地震について記述した例を表3に示す。

### 3.4 KIEFシステムに関する考察

上記のような記述をすることにより、[8]に記述されている建築基礎設計に対応する知識については、ほぼ知識ベースとして記述できた。しかし、一方で、KIEFシステムの改良の必要性が明らかになった。ここでは、この改良の方針を述べる。

・モデルの詳細度の取扱いの必要性
今回知識ベースとしてモデル化した様々な解析手法においてはその手法に応じた詳細度で建物や地盤をモデル化する。例えば、地震の伝播を考える際には、地盤をいく層もの地層の重なりとしてモデル化することが必要であるが、地下水位の変位を考えるときには、これらについて細かくモデル化しない。この様な解析手法に応じたモデルの詳細度の変更を可能とし、データの一貫性を管理する必要がある。

・解析手法の選択の支援
また、解析手法の特徴に関する記述については、表3に示した解析手法の説明をテキストで表現するにとどまっており、この部分の支援の枠組についても考察する必要がある。

## 4 結言

本論文では、まず、最初にKIEFシステムについて概観し、建築設計に関する知識の分析とKIEFシステムへの実装可能性について議論を行った。また、現在構築中の知識ベースについて述べ、KIEFシステムにおいてその多くの知識が記述できることが確認された。また、記述できない形式の知識を明らかにすると共に、それらの知識を扱うための改良の指針について議論した。

これからの展望としては、今回議論したKIEFシステムの改良の指針に基づく改良と改良し、本システムを建築設計支援システムとして完成させるとともに、実際の設計者などに使ってもらうことにより、以降の改良の指針を得る事などがあげられる。

## 謝辞

本研究の一部は、国際研究プログラムIMS(Intelligent Manufacturing Systems)内のコンソーシアムGNOSISの一環として行われた。

## 参考文献

[1] 吉川弘之, テクノグローブ, 工業調査会, 1993.

[2] Tomiyama, Tetsuo., From general "design theory to knowledge-intensive engineering." *Artificial Intelligence for Engineering Design, Analysis and Manufacturing (AIEDAM),* Vol.8, No.4, pp. 319-333, 1994.

[3] 吉岡真治, 冨山哲男.知識集約型工学のための計算機フレームワークの研究, 第7回設計工学・システム部門講演論文集, pp. 105-108. 日本機械学会, 1997.

[4] 日本建築学会建築法制委員会(編), 性能規定の法定化による社会へのインパクト, 日本建築学会, 1997.1997年度日本建築学会大会(関東)建築法制部門研究懇談会資料.

[5] 関谷貴之, 石井理貴, 冨山哲男, 工学知識ベースの汎用性向上の基礎研究.1995年度人工知能学会全国大会講演論文集, pp. 407-410, 1995.

[6] 吉岡真治,冨山哲男.設計支援のための統合モデリング環境の研究－プラガブル・メタモデル機構の提案－, 人工知能学会誌, Vol.13, No.2, pp.312-319 1998.

[7] Umeda, Yasushi.; Ishii, Masaki.; Yoshioka, Masaharu,; Tomiyama, Tetsuo., "Supporting conceptual design based on the Function-Behavior-State modeler". *Artificial Intelligence for Engineering Design, Analysis and Manufacturing (AIEDAM),* Vol.10, No.4, pp.275-288, September 1996.

[8] 日本建築学会(編), 建築基礎構造設計指針, 日本建築学会, 1988.

学術情報センター紀要　第10号（1998）

研究論文

# 軍の戦闘能力を左右したのは何か

# What Determind the Fighting Power of the Armed Forces

中京大学情報科学部　山田　尚勇
Hisao YAMADA
School of Computer and Cognitive Sciences
Chukyo (Tyûkyô) University, Toyota, Japan

**要旨**

本稿では、まずごく初歩的な数理的分析を通して、かつて日本の思想を風靡した精神主義の非合理性を検討し、大東亜戦争における日本の敗北の真の一因を明らかにする。その上で、そうした思想を醸し出す一端となった、日本語の表記法の問題点を検討し、この問題についてはわれわれの多くがいまだに不合理な思考の枠に囚われていることを指摘する。

**ABSTRACT**

In this note we first examine through elementary mathematical analysis the irrationality of spiritualism once dominated Japanese thinking, and bare a real cause of our defeat in the Pacific War. Thereupon we examine some problems in our writing system which condoned us to foster such spiritualism and point out that even today we are often under the spell of irrational thinking about our own writing system.

[キーワード]　漢字、教育、軍隊、数学、精神主義、戦略、戦力、縦書き、物量主義、分散逐次投入。
[Keywords]　Kanzi, education, armed forces, mathematics, spiritualism, strategy, fighting power, vertical writing, materialism, divided serial deployment.

**目次**

An *ex parte* judgment is always suspect.
-Rex Stout

## 1　いまごろなぜ戦争の話しか

これからしばらく、戦争において敵や味方の戦力を左右したのは何であったのかについて、主観を離れ、ごく初歩の理詰めを通して、客観的な分析を試みてみることにする。

この平和な世の中に、何を物騒な、と思われるかたもいらっしゃると思うし、それに過去の悲惨な戦争体験から、もう戦争の話しは聞くのもいやだというかたたちも少なくないと思う。しかし、以下に続く、やさしく初歩的な理屈をたどれば、かつてのわが国の戦争指導者が、いかに数理的な分析力に弱く、それゆえわれわれ国民を、負けることの分かっている戦争に巻き込むことになってしまったかが見えてくるであろうし、同時に、われわれ自身がこのくらいの評価能力を具えていなければ、また同じような目に遭わされかねないという、反省と自戒とを込めた思索であることがお分かりいただけるものと思う。

言うまでもなく、勝つことが明らかなら戦争をしてもよいのだなどというつもりは毛頭ない。しかし、巷における回顧談の中には、負けることが明らかだったから、戦ったのがいけなかったという論旨の話しが多いことは事実であるし、また自衛隊の予算増額要求の背後にも、いざという時に負けるわけにはいかないからという理由づけがしばしば用いられている。

その一方で、国際摩擦の解決は、本来理を尽した政治的交渉によるべきものであるにもかかわらず、その基礎となる国際的折衝に充分な能力を持つ人材をその任に当たらせるような人事が行なわれているとは言えないし、また教育も、そうした人材を広く育てあげ得るような配慮の乏しい、画一的なものでしかない。本稿の目的の一端は、そうしたお門違いの思考がいまだに存在する実情を指摘することにもある。

すなわち本稿は、日本の教育システム一般が、暗記と、与えられた問題に模範回答を与えることに主力を集中していて、いかに初歩的な科学的知識さえ自から活用する能力を身につけさせることにおいて、充分成功していないかについて筆者が書きつづった、連作の一つとして新たに書いたものである。

そうした文章の性格上、いくらかの数理的思考力を要求しているが、そのほとんどは小学生の算術のレベルのものであり、また残りも、たかだか中学の数学の範囲内のものである。一般に分析となると、少々の数学はつきものである。しかし、ここでの数学は、無味乾燥どころか、なま身で息づいていて、しかもやさしいと思われることは受け合いである。したがって、少し時間をかけて、1行1行を考え考え読んでいただければ、森羅万象の中にひそんでいる、簡単な数理的構造の一つが透けて見えてきて、ついには快感と満足とを感じて下さることと思う。

日本が第2次大戦に敗れたとき、筆者は(旧制)中学の3年生であった。物資窮乏のもと、4年生になってからの数学の教科書は、表紙もなく、綴じてもなく、ただ印刷された全紙を折って裁断したもののいく折りかを束ね合わせたものであったと思う。

ともかくそのときの教科書は、もうとっくに失ってしまったし、またひょっとすると、いまでは日本中を探しても、もうどこにも残っていないほどのものかもしれない。そう考えると、いまとなっては心情的に、

失ってしまったことが誠に惜しいという気がする。

何しろ1946年といえば、軍事教練のために軍から配属された将校が毎日のように「人生18年」を訓示し、われわれが20歳までも生きられるかどうか分からないと聞かされつつ、その日その日を精一杯生きたという時代が、やっと終わったところである。したがって知識を吸収するのにも、誰もが目一杯の努力をしていたときの教科書が、まだそのまま印刷されていたのである。国語や歴史などの教科書では、世の中の情勢や価値観が変わって都合の悪くなった部分は、墨で黒く塗りつぶされていたが、数学ではそんなことはなかった。その代わり、内容は超特急の駆け足で展開した。

今考えるとその教科書は、外見に比べて内容のほうはしっかりとしていた。当時の中学4年生はいまの高校1年生に相当する年齢であるが、その教科書には、すでにやさしい微分方程式までもが含まれていた。しかもその数学の科目は、一部の生徒が選択したというようなものではなく、全員が履修した科目であった。（ただし、筆者は実物を見たことはないが、すでに1930年代初頭の、中学レベルの工業学校の数学教科書の中には、微積分学の初歩を取り入れたものがあったという。）

微分方程式というと、ふつうにはまず微分積分を学んでからでなければ理解できないものとされている。したがって、ここまで読んでこれから先の内容を推測し、もう怖じ気づいてしまったかたがいらっしゃるかもしれない。しかし、実はこれから説明することは、自分で一歩一歩を考えれば、少なくとも考え方としては常識的で、まともで、やさしいものである。しかも自分では数学が不得手だと思っているかたがたにも少しでも興味を持っていただけるようにと、説明にはエピソードをはさむなどして、可読性を高めるようにくふうをしてみたので、まあだまされたと思って、ゆっくりともう少し読み進めていただきたい。そうすると、いままでぼんやりと、定性的にしか理解していなかったようなことがかなり定量的になり、これまでは見えていなかった構造がはっきりとした形をとり、自分で驚かれるかたも、けっこういらっしゃることになると思う。

そしてそれは、たとえば物が落ちるという現象は万人が日常目にしていながら何ら不思議に思われていなかったのに、近代力学の祖ニュートンは、伝説によると、りんごの落ちるのを見て万有引力の法則を思いつき、それを力学体系に組み込むことによって、天体の運動が簡単に説明できるようになったというような話しを思い出させてくれることになるのではなかろうか。しかし言うまでもなく、ニュートンの理論と異なり、こちらには特に新しいことはない。

## 2　敵・味方のつぶし合いを数値で比べようとすると

さて、すぐ本題に入り、いま初めにいる敵（enemy）の戦闘単位の数を、その頭文字をとって $e_1$ で表わすとしよう。同様に、味方（friend）の戦闘単位の数を $f_1$ としよう。この単位としては何でもいい。兵士のこともあろうし、戦車のこともあろうし、戦闘機のこともあろうし、軍艦のこともあろう。ただ敵と味方で同じものをもってくればいい。

いまこの敵 $e_1$ と味方 $f_1$ とが、奇襲攻撃を仕かけるのではなく、1939年のノモンハン事件や、1944年のレイテ海戦のように、正々堂々と正面衝突して戦うことを考える。戦闘とはお互いに敵の勢力をつぶし合うことだから、戦闘が始まれば、敵・味方とも、その数は時間とともに減っていくことになる。したがって、ある時間が経ったときには、敵の数と味方の数は、それぞれ $e$ と $f$ とに減ってしまっていたとする。この $e$ と $f$ とはまだ分かっていない数で、しかも時間とともに変わるから、数学で言う変数であるが、ここではそんなにむずかしく考えなくてもよい。ただ、ある時間における敵と味方の数を一般的に表わす記号である。

戦闘が続いているのだから、ある適当な短い時間のあいだに、この二つの数はさらに少し減るであろう。もちろんその減り方には偶発的なものがあるのは当たりまえである。しかし、平均的に見ると（あるいはもっといかめしく言いたければ、確率的に見て）、その減りぐあいは、観察している時間が長くなれば、ほぼそれに比例して大きくなる。

だから、ある観察時間中に敵と味方の数の変わった量を、それぞれ $\Delta e$ および $\Delta f$ で表わすとしよう。ここの $\Delta$（デルタ）は一般に小さな変化という意味を表わす記号で、$e$ や $f$ が少し変化するのだから、それぞれを $\Delta e$ および $\Delta f$ と書くわけである。（これらはそれぞれ $e$ の微分、$f$ の微分と呼ばれるものだが、何もむずかしいものではなく、ただの記号、つまり省略した書き方にすぎない。）さて、お互いのつぶし合いなのだから、大まかに言って味方の減り分 $-\Delta f$ は、そのときの敵の勢力、つまり残存数 $e$ に比例するだろうし、敵の減り分 $-\Delta e$ はそのときの味方の残存数 $f$ に比例すると考えるのは妥当であろう。ここで、$e$ や $f$ やは正の数だから、増加

ではなく、減っていくということが分かるように、$-\Delta e$、$-\Delta f$ と、マイナスがついているのである。

しかし、これらの減り分がそれぞれ相手の残存数に比例するとは言っても、敵と味方ではその比例の仕方が同じとは言えない。たとえば第２次大戦中の日本兵１人１人は実に勇敢で強かったという定評があるし、アメリカの戦車の１台１台の火力は実に強力だったということがある。したがって、たとえば兵士とか戦車とか、いま特定の種類の戦闘単位に限ったときの、敵と味方とのあいだの単位あたりの戦闘力の比率(ratio)を考え、味方は敵の $r$ 倍強かったものとして、味方の単位にそれを掛けてやる。この $r$ は一つの数であり、味方が敵よりも本当に強いときには $r$ は１より大きい($r>1$)し、逆だと１より小さい($r<1$)し、同格なら $r=1$ となるだけで、$r$ は一般性を持った記号である。この $r$ は、一般的に「術力」などと呼ばれたりするが、たとえば兵士の時にはその戦闘能力の対敵「練度」の頭文字の $r$ と思ってもよい。戦車などの兵器の場合いだと、能力的に互格(互角)に戦える台数の比となる。

したがって、ある同一観察時間内における敵方の兵力損失量と味方の兵力損失量との比を、いま割り算の形で書くとすると、$r$ が比例定数となるから、上に述べた関係を算術的にまとめれば、

$$(-\Delta e)/(-\Delta f)=r\cdot f/e$$

という式ができる。ここで・は掛け算の記号で×と同じであり、また/はただの割り算の記号で、ふつうには、たとえば $-\Delta e$ が分子として横線の上に、そして $-\Delta f$ が分母として横線の下に書かれた形で示されるのだけれども、印刷を簡単にする都合上、慣習に従ってこうした記し方を使うことにする。

あとはただの算術演算を使い、上の式ではまず左辺のマイナス同士を消し合わしたあと、それぞれの分母を＝の反対側にまわしてやって

$$e\cdot\Delta e=r\cdot f\cdot\Delta f \qquad (1)$$

ができる。ここで $r$ は、値はまだ未知のままでよいが、しかし定数であって変化はしない。

この式の両辺が敵・味方の勢力の釣り合いを示していることは見てすぐ分かるであろう。すなわち、左辺の敵の数 $e$ が大きければ、右辺の味方の減りぐあい $\Delta f$ が大きくなり、また味方の数 $f$ と敵の減りぐあい $\Delta e$ についても同じである。そして味方が敵の $r$ 倍強い兵士から成り立っていれば、その分だけ $\Delta e$ が大きくないと、この式は釣り合わないから、右辺には $r$ が掛かっていることが分かる。

## 3　もっと使いやすい式にするには

このままの式では、$\Delta e$ とか $\Delta f$ とか、一見得体の知れないものがあって、すぐには計算に使えない。だからここで微積分学のごく初歩の知識を少しばかり借りなければならないが、そうすると(ここでは読み飛ばしたとしてもあとで困らない、以下の初歩的なステップを経て)いま求めている、あとの関係式(2)が出てくることになる。しかし、めんどうがらずに、式の導出を読んでいただければ、世の中の仕組みの数理的な構造が、少しは透けて見えてくることになると思う。

まず、観察時間を短かくしていくとともに、($f$ と比例している) $\Delta e$ の値も小さくなる。しかし、戦闘の始まりから「いま」にいたるまでの期間に行なう観察の回数は、観察時間の短縮に反比例して大きくなる。したがって $\Delta e$ のおのおのの値を観察回数分だけ集めたものの総和は、実はある一定の極限値に収束するのである。(そうした計算操作はニュートンの区分求積法などと呼ばれる。)

そのようにして各項を寄せ集めたものの総和を示すために $\Sigma$ という記号を用い、左辺をまとめて $\Sigma\,(e\cdot\Delta e)$ と書き、右辺についても同様にすると

$$\Sigma\,(e\cdot\Delta e)=\Sigma\,(r\cdot f\cdot\Delta f)$$

という式になる。さらにこの観察時間を短くしていった極限だということを示すために、数学の記法に従って $\Delta$ を $d$ と書き、$\Sigma$ の代わりに積分記号 $\int$ を書くのだが、$\Sigma$ がそれぞれの括弧の中の積の項の寄せ集めだということを知っていれば、括弧が省略されても別に困らないから、以下では特に括弧を書かないことにすると、積分方程式

$$\int e\cdot de=r\cdot\int f\cdot df$$

というものができる。ここで $r$ は全ての項に共通だから、前に括り出すことができるのである。

問題点をここまで抽象化すると、もうその演算の細部の意味づけを理解することは、そう直観的でなくなるので、あとはごく初歩の積分学に頼って計算を実行する(つまり積分する)ことになるのだが、ここでは式の両辺のおのおのが変数 $e$ あるいは $f$ の１次式、それもこれ以上簡単にはならないという、裸の変数 $e$ と $f$ とだけから成っているのだから、微積分学では全く初歩の計算として知られているように、その積分ももっとも簡単な２次式が出るだけである。すなわち、式の左右は

$$e^2/2=r\cdot f^2/2+C$$

となる。ここで $C$ は、(この場合いは)戦闘過程の始ま

るまえの状況によって決まってくる、積分定数と呼ばれる、固定した数(constant)である。

これをすっきりした形にするために、左右それぞれを2倍にし、$2C$ を改めて新しい定数記号 $P$ と書き換えれば、

$$e^2 = r \cdot f^2 + P$$

となる。ここで $P$ は戦闘の準備(preparation)の頭文字を当ててみただけのことで、それ以上の意味はない定数である。

この定数 $P$ の値を求めるためには、戦闘開始時の状況に辻褄を合わせてやればよい。すなわち、その時点で敵と味方の数はそれぞれ $e_1$ と $f_1$ とだったから、$P$ の値は

$$e_1^2 = r \cdot f_1^2 + P$$

を成り立たせなければならないことになる。

あとは算術で、これを $P$ について解くと

$$P = e_1^2 - r \cdot f_1^2$$

となり、これを上の $e_1^2$ の式に代入すると

$$e^2 = r \cdot f^2 + e_1^2 - r \cdot f_1^2$$

となるから、整理すると

$$e_1^2 - e^2 = r(f_1^2 - f^2) \qquad (2)$$

という一般的な関係が求まる。これならすぐにでも $e$ もしくは $f$ の一方に数を代入すれば、算術的に他を計算できる形になっている。

いま兵士を戦闘単位の例としてこれを復習してみる。敵に比べて戦闘練度(つまり強さ)が $r$ 倍の兵士 $f_1$ 人が、敵 $e_1$ 人と戦闘を開始したとすると、この式は、時間とともに変化する敵兵士の数 $e$ と味方兵士の数 $f$ とのあいだに存在する関係を示している。

したがって、対敵練度 $r$ の味方兵士 $f_1$ 人が敵兵士 $e_1$ 人と死闘を繰り広げ、互格の戦闘ができた、つまり最後にはちょうど両軍とも全滅したとすると、その時には $e=f=0$ の条件が成り立つということだから、これらを上の(2)の式に代入すると

$$e_1^2 = r \cdot f_1^2$$

となる。つまりこの両軍の同時全滅が起こるためには、味方の対敵練度が

$$r = (e_1/f_1)^2 \qquad (3)$$

であることが必要である。別の言い方をすると、これは

$$s = e/f = \sqrt{r} \qquad (4)$$

のとき、すなわち数において $s$ 倍優勢な(superior)数の敵と互格に戦えるためには、味方兵士1人1人の対敵練度(つまり強さ)は戦闘単位数の比 $s$ の自乗にまで高くする必要があるということになる。あるいは逆に言うと、敵の数量的優越度 $s$ に対して味方の対敵練度 $r$ を大いに高めても、$r$ はたかだかその平方根倍の効き目しかないということになる。

これらの式は、一般に $r$ や $s$ が1よりも小さい場合いにも、もちろんあてはまるものである。

## 4　日本軍はこんなことも知らなかったのだろうか

初めに述べたように、ここで使った数学は、筆者が中学4年生で学んだ常微分方程式の初歩、それも式(1)は変数 $e$ と $f$ との関数を、それぞれ式の右と左とに分けた形をした、特にやさしい、変数分離型と呼ばれる種類のもので、その中でも、それらの関数がただの $e$ と $f$ 自身だけという簡単なものだから、これ以上に簡単なものはないという簡単な例である。つまり上で述べた議論は、当時の中学4年生なら、ちょっと考えれば、自分で簡単に導き出せたたぐいのものである。

したがって、戦場において戦うことを使命とし本分とする軍人としては、兵卒はいざ知らず、少なくとも陸軍士官学校(陸士)あるいは海軍兵学校(海兵)、さらには陸軍大学校(陸大)や海軍大学校(海大)を出ていた将官や参謀たちであったなら、このくらいのことは当然心得ていたものと思われるであろう。

かれら上級軍人の学歴を示す肩書きがどう呼ばれていたのかを筆者は知らない。しかし、日本の陸士や海兵に対応する、アメリカのウエストポイント(マッカーサー将軍やアイゼンハワー将軍の出身校)やアナポリス(ニミッツ提督やハルゼー提督の出身校)の卒業生は「Bachelor of Engineering(工学士)」の称号を与えられているのだから、日本の将官や参謀たちとて、中学4年生の数学の知識ぐらいはあったはずであり、戦略・戦術のプロとして、これぐらいのことを心得ていたのは当然のことと思われる。しかも、たとえば戸部・他(1984)[21]によると、日本の陸海軍中枢の人事における評価システムは、年功序列と、陸士・海兵、さらには陸大・海大の卒業成績の順位をもとにしたものであり、特に海軍での兵学はすべて理数系であったから、理数科目に強い、それも独創性を発揮するよりは、暗記と記憶力を強調した教育によって突出した、いわゆる学校秀才型の者に有利であったという。

にもかかわらず、これから見ていくように、こうした戦力評価に関して、理数科の観点からのかれらの理解がいかにおそまつなものであったかを思わざるを得ない。おそらくそれはかれらが、書物に書かれた理論を教師が講義し、学生がそれを丸暗記して点数をとる

といった、実世界における体験とは縁の薄い、いわゆるスコラ派哲学の方法論による教育を受けていたからであろう。しかしそんなことでは、現実の世界における自然現象や社会現象はおろか、かれらに直接関わりのあった戦略・戦術における諸現象の本質を理解し、それらに向って能動的に対処して行けるだけの思考力がなかなか身につかなかったことであろう。

残念ながらその悪弊は、今日においてもさして改善されているとは言えない。現在においても、進学のための試験問題は、諸現象の中から理論や法則性を見いだすのではなく、たとえば大学の共通入学試験のように、単に孤立した記憶によって答えを示せればよいという、いわゆる求答主義に基ずいたスピード技能を主として要求している。したがって、そうした制度のもとでは、ひとり入試突破技術の熟練工のみを残し、真に問題の本質とまともに取り組んでじっくりと自分で考えるといったことに関心を持っているような人材は、入試ごとに落ちこぼれていってしまうような結果をもたらさざるを得ない。そのことは、数学や論理といえば抽象的な丸暗記ものと心得ていて、それらを身近な具体的な問題と結びつけて理解することができず、しかも思考に対する情熱が燃え尽きたとしか思えないような大学生が、理系の中でも多数派を占めている現状からも容易に推察できる。

大東亜戦争時代の大半の期間、筆者は台湾台北市の松山飛行場に近接して住んでいた。そしてときどきは、飛行場との境を流れる小川を越えて陸軍航空隊の基地に遊びにいき、搭乗員たちから「ちかごろ敵さんの弾が、5倍も10倍もよく当たるようになりやがった」といったような話しを聞かせてもらうことまでできた。

戦後になって、それは砲弾から発射した電波が目標から反射されるのを測定して距離を知り、最近接時に自動的に爆発させるものながら、機密保持のために距離を時間にすり換え、VT(variable time)信管と、意味を不明にした符牒で呼ばれていた近接信管のせいだと分かった(NHK 1993[3]参照)。しかもアメリカ側のテストによると、実は命中率は一気に20倍にもあがっていたこと、そして日本軍の搭乗員の勘はほぼ正確にそれを感じ取っていたことを知った。しかるに日本の兵器の専門家のほうは、戦後までその存在を全く知らなかったという、情報不足とイマジネーション欠落の状態だったようである。(筆者が1954年にアメリカの大学に留学したときに修士論文を指導して下さった教授は、戦争中はこの近接信管の評価テストに携わっていたという偶然の一致までもあった。)

さらに筆者は、1944年10月12日から14日にかけての、米国第38機動部隊に対する台湾沖航空戦のときや、その後の沖縄戦のときには、飛行場わきの小高い岡をかすめて出撃する、数多い陸軍機を見送った。あとで聞いたことによると、そのうち、旧式の機種のものの多くは神風特別攻撃隊機とのことであった。

1945年3月に大本営が台湾を作戦地に指定したので、筆者が満15歳になった6月には、中学3年だったわれわれ級友のほとんどが召集を受け、陸軍2等兵として学徒兵に編入された。そのとき筆者は支給された軍装シャツの襟につけるべき、黄色い小さな星が一つついた赤い襟章を何の注意もなく渡され、上下をさかさまに縫いつけて下士官に小突かれたことなどを、いまでも思い出す。

そうした個人的な経験があったので、戦後になって大東亜戦争の指導者的な地位にあった人たちなどによる戦記ものや、戦略戦術の分析ものなどを、いままでにかなりの数を読んだ。しかしそれらの中で、上に述べたような、戦力についての定量的な分析や、それに基ずいた作戦の分析や評価に、いまだお目にかかったことがない。しかもそれは、われわれ読者がそれだけの数理的理解能力がないと見くびってのことではなさそうである。

それに比べて、VT信管による敵弾の命中率を、命のかかった勘によって5倍から10倍も良くなったと感じ、しかもそれが敵兵の技能力によるものではないと思っていた搭乗員の本能的判断力は、大したものであったと言うべきであろう。なぜならアメリカ側のテストが示したように、VT信管によって砲弾の実効命中率が20倍にも上がっていたとしたならば、それによる長期的累積効果は、式(4)によって$\sqrt{20}\fallingdotseq 4.47$倍の味方の損率として感じられたはずであるし、また戦闘のさなかにある者の体験による主観的評価値はもっと大きくなり、命中率の向上分に近い感じとなったはずだからである。

とにかく、彼我の戦力に関する日本軍の中枢の判断がいかに甘かったかを明らかにするために、以下では過去において軍の指導者によってなされたいくつかの代表的な発言などを例として取りあげ、上で計算した簡単な式と対照して分析することにより、それらの妥当性について少し考えてみることにしよう。

## 5　百発百中の砲の威力

まず第1に取り上げるのは、日本海軍が1905年に帝

政ロシアのバルチック艦隊を完全に撃破した日本海海戦の直後、東郷平八郎司令長官の与えた訓示の中にあり、かつてはまことに有名であった、「百発百中の砲一門、よく百発一中の砲百門を制す」という文言である。

特に1930年のロンドン軍縮条約によって、日本海軍の艦船保有数がアメリカおよびイギリスのおのおのに対して約60パーセントに制限されたあとは、この文言が神格化されて、日本海軍における血の出るような猛訓練の指導原理となった。したがって、まずはこれを分析してみることにする。

いま単純に1艦に砲1門があると考えて、敵砲 $e=100$ 門と互格に闘える、対敵練度 $r=100$ の味方の砲数（すなわち艦数）を、式(2)によって計算する。両艦隊が互格であるとは、戦闘において両軍が同時に全滅するということだから、$e=f=0$ を代入すると、(2)は $100^2=100\times f_1^2$、すなわち $f_1=10$ となり、100門を相手にして互格に闘うには、東郷長官の訓示が述べたように、わが方には1門ではなく、実は10倍の10門が必要になる。

事実、もしも命中弾1発で軍艦が沈むとすると、このときには両軍の最初の一斉射撃で、敵艦10隻が沈むが、味方も1隻沈み、敵艦90隻と味方艦9隻が残り、依然として10対1の比率が保たれる。そのあとの計算は少しややこしくなるが、しかし大まかな確率的思考をしてみると（つまり同じ条件で何度も戦闘があったとしての平均値をとるということをすると）、この10対1の撃沈率は最後まで続くと考えても、そう不自然でないと思えるであろう。したがって統計的には、敵の100門の砲と渡り合えるためには、味方には、1門の砲ではなくて、10門の砲が要ることが分かるであろう。

とは言っても、事はそれほど簡単ではない。

一隻の軍艦に100門の主砲は積めないから、いま仮りに敵・味方とも全艦が戦艦「大和」や「武蔵」なみに主砲10門ずつ（実際は9門だったが）を積んでいたとすると、100門を積む敵艦の数は10隻になる。もし味方のほうが百発百中であり、しかも各回に目標を最適に分散して狙えるとし、かつ1発の命中弾で対手が沈むものと仮定すれば、敵味方の同時射撃のときには、敵艦10隻を沈めるのに、味方は1隻でよい。（実際には砲塔の構造上、10門が別々の相手を射つことはできないのだが、いまはそこまで立ち入らない。）それに対し、敵の100門はいちどに命中弾1発しか射てないから、仮りに味方艦の数がもっと多かったとしても、敵は味方を1隻しか沈められないことになる。つまり味方の1隻は、やはり敵10艦としか互格に戦えないということになる。

しかし、ここで用いた仮定があまりにも事実に反することは明らかだから、つぎにもう少し現実的な計算をしてみる。

すなわち、いま条件を少し変えて、もし1艦に $n$ 発の命中弾があって、はじめてそれが沈むものとし、かつそれまでは戦闘能力を完全に維持できるものとすると、各門が各射撃においていつも同じ目標を砲撃することをすれば、味方1艦はちょうど $n$ 射撃後に、敵10艦を全部沈められることになる。しかし、敵の100門は味方の1艦だけを目標にした全射撃を $n$ 回集中してやっと $n$ 発の命中弾が出るのだから、もしも攻撃目標を分散すると、損傷を与えることはできても、1隻も味方艦を沈没させられない。したがって、またもや味方の1艦は、やはり最適な戦術を使っている敵の10艦と互格ということにしかならない。以上とて、まだ命中率が100パーセントとか、対敵練度比が100などといった、極端な条件を仮定した上での計算であるが、それでやっと、「1艦の砲、よく10艦の砲を制し」得ることになる。

ロンドン条約によって数量的に劣勢に立たされた日本海軍が、いよいよの猛訓練によって対敵練度を上げ、その補いをつけようとしたことはよく分かるし、また、のちにはそれが高じて、過度の精神主義に走ることになったのもうなずけないこともない。しかしすでに上のいくつかの具体的な計算例の結果が示すように、100倍というような超練度をもってしても、東郷長官の訓示がともすると与える錯覚のように、百発百中の砲1門では百発一中の砲100門と互格には戦えないことは明らかである。それどころか上のように、各艦に積んだ主砲全部がそれぞれ独立した目標を射てる自由度があるというような、事実に反した仮定をした場合いでさえ、100倍の対敵練度で互角に戦える敵の数量的優勢さは、やはり式(4)が示しているように、対敵練度の平方根倍、すなわち10倍にしかなっていない。

## 6　バルチック艦隊が打ち破れた理由

いま日本海軍の実質的な練度らしいものを、日本海軍の砲術の神様といわれ、第2次大戦末期には重巡洋艦「利根」の艦長としてレイテ沖海戦を戦い抜いた（生出　1988[7]参照）黛治夫大佐が与えているデータ（黛 1972[30]、p. 273）によって計算すると、日本艦隊がバルチック艦隊を打ち破った日本海海戦のときには、日本艦隊の総合対敵練度は15.6もの高さを示していたようである。さらに、両軍の戦闘単位数として、艦数で

はなく、口径12センチから30センチまでの砲数で比べると、日本が358門、ロシアが245門であったという(p. 157)。すなわち敵は味方のほぼ正確に3分の2の砲数しかなかったことになる。

おおまかな計算ではあるが、(2)式にこれらを代入して計算すると、ロシア側の砲が全滅した(つまり$e=0$になった)ときに、日本側にはまだ約352門余の砲が残っていることになることが簡単に分かる。すなわち、日本側の損失はたったの1.5パーセントほどで、98.5パーセントは生き残るという、一見まことに驚異的な計算結果になる。

したがって、日本海海戦における日本側の圧勝には、よく言われているように東郷長官の巧みな操艦戦術が効を奏したこともある程度は事実であろうが、むしろ先に簡単な計算で求めた式(2)による検討が示してくれるのは、$r=15.6$という、信じられないほどに高い対敵練度にあったかに思えるかもしれない。

しかし、黛の与えた$r=15.6$の内容をさらに詳細に分析すると、その中身には、日本艦隊の砲弾が相手のものに比べて、ずっと近代的であったことによって、$r$が2.6倍にもなっている効果が含まれているから、15.6をそれで割ると、真に兵士の術力だけによる練度は、もっと妥当な$r=6$という値に下がってしまう。

そこで、もし日本側の砲弾が近代化していなかったと仮定して、はじめからこの$r=6$で式(2)を計算しなおすと、相手が全滅したときの日本側の残存砲数は343.7門となり、やはり約4パーセントの損失が出るにすぎないことが分かる。

さらに、ここで日本海軍の対敵練度を極端に下げて、ロシア側と全く同じ、すなわち$r=1$だと仮定して式(2)を用いると、敵が全滅した$e=0$のときには$245^2=358^2-f^2$となり、これから$f=261$が求められる。すなわち、日本海軍の兵士の練度も整備も相手側と同等で、かつ東郷司令長官の優れた操艦戦術もなかったとしても、ただ358門対245門という、大砲の数量的優位に頼るだけで、相手を全滅させたときに、日本側の砲はまだ261門、つまり約73パーセントが生き残っていることになる。

すなわち、ここで強調しておきたいのは、戦闘能力としては、兵士の対敵練度よりも、戦闘単位の員数のほうが圧倒的に、あずかって力があるということである。1867年に始まった明治維新ののち、新政府の政策に対する不平士族による反乱は、1874年の佐賀の乱を初めとし、神風連の乱、萩の乱、秋月の乱(ともに1876年)と続き、最大の西南の役(1877年)が最後となったが、いずれの場合いにも、戦いが専門であるはずの士族、すなわちさむらいたちの反乱軍に対して、戦士としてはしろうとの農民によって新たに編成された政府軍が勝利をおさめ得たのは、その点を数理的に明らかにしようとした文献はみつけることはできなかったものの、やはり数における優位が大きくものを言ったものと思える。

したがって、凡庸の兵力でも、数さえ揃えば敵に圧勝できるということこそ、日本海海戦の真の教訓となるべきはずのものであったのである。それなのに日本の軍部が、兵力比較におけるこんな簡単な計算を理解しようともせず、日本海海戦の大勝利を兵士の練度と東郷長官の操艦戦術に帰してそれらを神格化し、精神主義一辺倒にまで高めてしまったところに、その後40年にわたる、日本人全体の歩まされた悲劇の道の出発点があったのである。

たとえば、まず1921年(大正10年)にワシントンで開催された軍備縮小の国際会議で合意されたワシントン条約、およびそのその見直しとして1930年(昭和5年)にロンドンで開催された軍縮会議におけるロンドン条約が日本に押しつけた条件は、英・米・日の艦隊保有量の比を5対5対3にすることであったから、もし英米を同時に相手にまわして闘うとすると、その兵力比は10対3になる。これを先に求めた式(3)に代入すると、英米を相手に互格で戦えるためには、装備の技術力や指揮官の戦術力などいっさいの優劣を含めたわが方の対敵練度は、11.1以上という、有り得べくもない数値でなくてはならないことになる。

ワシントン会議のときの随員であり、後に海軍軍令部次長となった加藤寛治中将などは、ワシントン条約の条件を不満とし、主力艦保有量の対米最低率は70パーセントはなくてはならないという、当時の日本海軍の定説を強く主張していた。アメリカのみを仮想敵国とする限り、艦船数が70パーセントあれば、互格に戦えるために必要な、日本の対敵練度は式(3)から、$(1/0.7)^2\fallingdotseq2.04$と、約2倍強ですむ。上の黛の調査が示すように、この程度の対敵練度であるならば、猛訓練によって実現は可能であろう。しかしながら、英米を同時に相手とすることになると、彼我の数量的格差は2対0.7となり、それを補うために必要な対敵練度は、式(3)によって、$(2/0.7)^2\fallingdotseq8.16$となる。これはほとんど実現の可能性のない、厳しいものである。

すなわち、この数値に比べると、凡庸あるいは凡庸

以下のロシアの海軍力を相手にした、東郷長官による猛訓練と海戦に対する細心の前準備とをもってしても、黛の指摘するように、わが方の実質的な対敵練度はやっと $r=6$ でしかなかったのである。したがって、政治や思想を抜きにした、純粋に軍事力だけの立ち場からすれば、ロンドン条約の条件は、日本としては絶対に呑むべきではなかったし、また意に反してこれを受け入れたあとは、「戦わば敗るるは必至」と肝に銘じ、あらゆる犠牲を払ってでも平和に徹するべきであったのである。

大東亜戦の初頭、太平洋全域において日本軍が圧勝したのは、これらの地域の日本軍が兵員数および装備において圧倒的に優勢であった上に、ひよどり越えと桶狭間との戦いをいっしょにしたような奇襲作戦がもたらした当然の結果であったことが、ここに述べた簡単な計算によって明らかである。したがってその後、いざ両軍が相応の戦力を投入し、正面切って対峙するに及んでは、もはやそのような一方的な勝利はあり得べくもなく、上のような理論に沿った形で戦闘の大勢が推移することになったわけである。

そうした事実を理解せず、初戦の表面的な勝利に驕り酔い痴れ、相手方の情報を充分に収集することもせず、希望的判断によって作戦を遂行しようとした日本軍が、その後大敗に大敗を喫したのは理の当然であったろう。そこには、もはや精神主義などの入り込む余地はなかったのである。

## 7　日本海海戦後にたどった道

東郷長官の訓示に戻ると、そもそも練度100倍というのは、いかにもむちゃな設定である。それで先に述べた黛大佐が彼我の射撃演習の成績などを基にして算出した資料に探ると（黛　1972[30]、1977[31]）、日本海軍の射撃の命中率、つまり対敵練度は、アメリカなどの場合いの約3倍だったとするのが妥当のようである。いま式(3)にこの $r=3$ を代入して計算すると、日本海軍は数量的に約1.73倍の相手と互格に戦えたに過ぎなくなってしまう。

したがって、この3倍という対敵練度を基にして、日本海軍は3倍の数の敵と互格に戦えただろうとした黛の推定は、砲術的には正しい数値を用いつつも、戦略的あるいは戦術的には自己の戦闘能力に過大評価を与えていたことになる。事実、1944年10月にフィリピンで展開された「捷1号」作戦の一端を担った、レイテ海での対艦隊砲撃戦では、黛艦長の率いる重巡洋艦「利根」こそ異例の善戦をしたものの（生出　1988[7]）、日本海軍自慢の戦艦「大和」にいたっては、口径46センチの巨砲を約100発も射って、命中弾は1発もなかったのだという（小倉　1997[10]）。

もちろん戦闘の結果は戦闘単位の数や対敵練度だけによって決まるのではなく、戦略目的、情報収集力、組織の特性など、ほかにも多くの要素がからんでいる。しかし、いまではかなり明らかにされているように、それらのほとんどについても、日本軍は欠陥に近い弱点をかかえていたと言える（千早　1982[19]、戸部・他1984[21]）。そのことについては、すくなくともこれらの2書をお読みいただければ充分であろうから、ここではこれ以上は立ち入らないことにしよう。

問題なのは、日本軍においては合理的、科学的、客観的な戦闘能力の評価が重要視されず、いたずらに主観的で自己過信的な精神主義が横行していたことである。

すでに述べたように、筆者は大東亜戦争の戦略・戦術的分析や戦記ものなどを、いままでにかなり読んできた。しかし戦闘能力については、本稿で述べたような初歩的な定量的分析でさえ、それらの著作の中には、まず見かけたことがない。ただ、わずかな例外として、千早(1982)[19]に出てくる一つのエピソードがあっただけである。

それは、こともあろうに、東郷平八郎長官の時代に日本海軍の戦術家と言われた佐藤鉄太郎中将が、質量が $m$ で速度が $v$ の運動物体の持つ運動エネルギーが $K=(1/2)\cdot m\cdot v^2$ であるという初等力学の法則を、理論的根拠も与えないままもじり、本稿の用語を用いて書くと

$$\text{戦力}=(1/2)\cdot\text{戦闘単位数}\cdot(\text{練度})^2$$

すなわち、味方の戦力(power)は

$$p(f)=(1/2)\cdot f\cdot r^2$$

となると称していたことである。敵の戦力のほうも、同じく $p(e)=(1/2)\cdot e$ となるのだから、この式によるときは、両軍が互格に戦えるということは、$p(e)=p(f)$ から

$$(1/2)e=(1/2)f\cdot r^2$$

が得られ、これから

$$e/f=r^2$$

となり、これを式(4)と比べてみると、佐藤の主張は味方の練度の効果を4乗にも過大評価していたことになる。

したがって、仮りに日本海軍の対敵練度として、先

に見た黛の調査による評価値 $r=3$ をとると、4乗の過大評価ということは、$r=3$ の代わりに $r=3^4=81$ と考えるということであり、黛の調査の与えた評価値の27倍もの大きな値を与えてしまうことになる。日本海軍の戦術家と言われた将官が、このような、自軍の術力の過大評価値をまじめに信じていたものとすれば、日本海軍が全体として自信過剰になってしまっていたのも無理はないと思える。

## 8　量の不足を質で補うことのむずかしさ

言うまでもなく、すべての将官・士官が佐藤のこの説を信じていたわけではない。このエピソードを紹介している千早は海兵と海大とを卒業しており、(海兵でなのか海大でなのかは不明であるが)その「級友で数学に長けた一海軍士官は、これに関して疑問を持って研究した結果、

戦力≒戦闘単位数・$\sqrt{}$練度

の方が正しいと主張した」と書いている(千早1982[19]、ただしこの式の用語は本稿のものに合わせてある)。

すなわちこれは $p(f)\fallingdotseq f\cdot\sqrt{r}$ ということであるから、これを用いれば、敵・味方が互格の戦力を持つということは $e=f\cdot\sqrt{r}$ だから、彼我が互格の戦力となる員数比は $s\fallingdotseq e/f=\sqrt{r}$ となり、これは式(4)と同じであるから、彼の主張のほうが正しいことが分かる。

しかしこれを読んだとき筆者は、「冗談じゃない」と思わずつぶやいてしまったのである。その理由は、かつての中学4年生が自力ですぐにでも出せるほどの答えを、千早自身はただ引用するに留まっており、しかもそれを海兵あるいは海大出身の、日本海軍のエリートの一人である、千早の級友が、その専門であるべき戦術について「研究した結果」に得た結論として、それを紹介しているという点にある。だから筆者は、それほどまでに自から考えることをしない、あるいはそれができない、指導者たちによって、日本国民はあの無謀な戦争を戦わされたのであろうか、と考えてしまい、その結果、本稿を書く気になったのである。もっともそのときには、こんなに長いものになることとは思わなかったのだが。

いかなる反復作戦においても、1回当たりの損失が10パーセントを越えるような作戦は長つづきがしないというのが、欧米における戦術の目安となっていたようである。事実、第2次大戦中のドイツ空軍による英国本土の爆撃は、毎回それ以上の損失が重なったために、ついにドイツは空爆作戦を中止せざるを得なかったという。

もちろん敵側にも損失は出るから、ただ味方側の被害が10パーセントと言うだけでは彼我のあいだの勝敗は何とも判断できない。したがって、いま一回の作戦において敵の損失のほうが味方の損失よりも10パーセント大きいとし、かつ単純に、味方の対敵練度 $r=1$ という状況を考えてみると、式(2)において

$$e_1^2-(0.8\cdot e_1)^2=f_1^2-(0.9\cdot f_1)^2$$

となり、これから作戦ごとに必要な、味方の数量的な最低対敵優越度 $f/e$ を求めるには、

$$(1-0.8^2)\cdot e_1^2=(1-0.9^2)\cdot f_1^2$$

から

$$f_1^2/e_1^2=(1-0.8^2)/(1-0.9^2)\fallingdotseq 1.895$$

つまり

$$f_1/e_1\fallingdotseq 1.376$$

となって、少なくとも37.6パーセントほど、味方は毎回敵よりも数量的に優位でなくてはならない。

もし同じ数の兵力をもって、いつもこの優位性を生み出そうとするなら、式(3)と突き合わせてみることにより、これは対敵練度を含めた味方の対敵術力をいつも1.895倍以上に保っておかなければならないことを意味する。しかし実状としては、兵隊の教育制度がきちんと整備されておらず、また後述するような文字のむずかしさなどもあって、教育の能率の悪かった日本軍は、戦争が長びくにつれて順次失われた、優秀な兵士一般の補充がままにならず、兵器の技術レベルの落差に加うるに、兵士の対敵練度も急速に落ちていったのである。

## 9　硫黄島作戦および沖縄作戦における日本軍の総合術力の概算

第2次対戦後、アメリカの某雑誌が大戦中の名将10人を選定したことがあるが、その中に日本からは硫黄島作戦(1945年2月19日－3月17日)のときの守備兵団の第109師団長栗林忠道大将と、沖縄戦(1945年4月1日－6月21日)のときの第32軍司令官牛島満大将の二人が入っている(岡田　1972[8])。

この二つの対戦は、陣地を構築し、航空機による援護もほとんどないまま島を防御した、劣悪な兵備の日本軍と、大艦隊の巨砲や艦載機の大編隊による猛烈な援護射撃や援護爆撃を受けつつ上陸作戦を敢攻したアメリカの移動部隊との闘争であり、ここまでに見てきたほかの対戦とは違って、火力、装備、戦闘態勢、防

御手段などが大幅に異なる兵力間の戦闘である。

したがって、これらにおける戦闘はいままで本稿で述べてきた、主として式(2)や(3)などの簡単な式を用いての分析の限界をはるかに越えた、複雑な数理的分析が必要となる性格のものである。

それを承知の上で、この二つの戦闘における日本軍の総合術力をごく大雑把に検討し、栗林および牛島両大将の、指揮官としてのアメリカ側の評価が正当なものであったことを確認してみよう。

まず硫黄島作戦であるが、アメリカ側の兵力は上陸部隊だけに限ってみても６万人の大部隊であった。これを迎え撃って玉砕した日本側の兵力は陸海軍合わせて約１万９千人で、アメリカ軍に与えた兵員の損失は、戦死１万３千、戦傷３万３千、計４万６千人であった（岡田　1973[8]）。

これらの値を式(2)に代入して単純に計算してみると、アメリカ軍に対する日本軍の総合対敵練度・術力として、$r=4.11$という数字が求まる。これは地上を移動して来る攻撃軍に対して、堅固な陣地を構築し、劣悪な装備とはいえ、大砲をもって攻撃を敢行している守備軍の$r$としては、かなり高いにしても、不可能ではない値であると思われるし、またこの高い$r$の値を反映して、栗林大将が10名将の１人に選ばれることになったものと考えてよい。それでも６万対１万９千という数量的格差はいかんともしがたく、遂には事実上全兵力が失われてしまったのであった。

つぎに沖縄作戦であるが、こちらにおける日本側の情報収集能力の不備や、二転三転した大本営の作戦計画のまずさなどについては、たとえば戸部・他(1984)[21]や岡田(1972)[8]を参照していただくとして、兵力的には、この戦闘におけるアメリカ軍は上陸部隊のみを数えても約23万8700人、それに対する日本側の守備隊は正規軍約８万6400人であった。そして戦闘が終了したとき、アメリカ軍の戦死者は約１万2300人、日本軍は約６万5000人であった。

戦闘状況は硫黄島におけるときと似たものであったが、やはり大幅に単純化し、これらの数字を式(2)に用いて計算すると、日本軍の対敵術力は$r=17.2$と出る。この異状な高さがあって、牛島司令官が10大名将の１人として数えられることになったのだと考えられる。同時にこの術力値の高さが当時のアメリカ軍に日本軍の戦闘能力を過大評価させることになり、それに対抗すべく、アメリカを日本本土に対する原子爆弾の使用に踏み切らせる一端ともなったという。

しかし忘れてならないことは、上記の$r$の高さは見かけのことであって、沖縄作戦では正規の日本軍戦死者のほかに、沖縄の島民のほとんど全員が戦闘に巻き込まれ、戦死者だけでも約10万人も出ている。その中には中学１年生以上、老人に至るまで、数多くが補充兵として何らかの戦闘に加わっていた。もしこうした人たちを日本軍の兵力の一部として数えるならば、最終的には日本軍の対敵術力が上の数字よりもかなり下がることは明らかである。

本質的には正規軍の任務である戦闘に、しろうとを強制的に巻き込んだ作戦の責任は、究極的には大本営にある。しかしながら、それを実行に移して多くの沖縄の一般住民を死に至らしめた責任の一端は牛島大将にもなかったとは言えまい。その意味において筆者は、牛島の指揮に対して、アメリカ側の分析ほど高い評価を与えたくない。むしろ筆者は、あとで述べる、1944年以後のフィリピン方面軍司令官として戦って敗れた、山下奉文（ともゆき）大将のほうを高く評価する。

## 10　長篠における織田・徳川連合軍と武田軍との戦い

ここでちょっと寄り道をし、時代を16世紀に戻して、武田勝頼を滅亡に導くきっかけとなった長篠（ながしの）の合戦をみてみることにしよう。

戦国時代の末期、1575年（天正３年）の５月、当時最強の騎馬軍団の誉れの高かった、甲斐の国（山梨県）の勇将武田勝頼（当時30歳）の軍勢は、南下して徳川家康（当時34歳）の配下奥平貞昌を三河の国（愛知県）長篠城に攻めた。これを救わんと、家康の兵8000は、織田信長（当時42歳）の兵３万とともに、長篠城の西、設楽原（したらがはら）に布陣して武田の軍勢を迎え、激戦半日にしてこれを破った。

現場での実地調査を経ることもなく書き継がれてきた、従来の歴史書による定説では、信長の鉄砲隊3000が、武田の騎馬隊に対し1000人ずつ周期を合わせた３交替で、当時の新兵器鉄砲による一斉射撃を行なったことが、この戦勝の理由とされてきた。しかしながら近年の詳しい研究では、この説には疑問点が多く、いまでは否定されかけているらしい（たとえばNHK 1997[4]）。それには、たとえば(1)織田鉄砲隊が1000名の長い砲列を敷いたとされる場所の中央には丘がせり出しており、見通しがなくて一斉射撃の合図が伝えられず、また号令も届かなかった上に、(2)当時の火縄銃は操作性が悪く、技の巧拙による操作時間差が大きくて、一斉射撃は戦闘力を大幅に下げてしまうものであっ

たこと、また、(3)設楽原一帯は水田地帯で、合戦のあった5月には田の泥が深く、騎馬隊が自由に駆けまわることは不可能であった、などが挙げられている。

その代わりとして、信長が家康の家臣酒井忠次に命じ、長篠の合戦の前夜に、鉄砲隊500を含む4000の兵を、暗夜の山道を迂回させて、長篠城攻略の要地、とびの巣山を守っていた武田の軍勢3000を急襲し敗走させて、武田勢の退路を断ったことなどが織田側の勝因として挙げられている。

しかしわれわれは、先に求めた計算式を使って、この合戦を少し検討してみよう。

まずおのおのの兵力であるが、武田の軍勢は1万5千、そのうち3000がとびの巣山を守っていたので、設楽原には1万2千がいた。それに対するに織田・徳川方は、織田軍3万、徳川軍8000、計3万8千であったが、この内4000がとびの巣山の急襲に割かれたので、設楽原には3万4千が布陣していたという。つまり長篠の合戦は織田・徳川連合軍3万4千、武田軍1万2千のあいだで戦われたことになる。

武田の強力な騎馬隊に対する、織田の新兵器鉄砲隊の対敵練度(術力)$r$が分からないので、いま仮りに武田側を$r=1$の味方とし、式(2)を用いて、武田勢全滅のときの織田軍の残存兵力を計算してみると、

$$34000^2-e^2=12000^2$$

から$e=31812$が求まり、これは織田側の初頭の兵力の

$$31812/34000=93.56\text{パーセント}$$

になる。すなわち、織田軍は6.5パーセント足らずの損失で武田軍を全滅させられる勘定になる。

逆に、こうした兵力数の格差がありながら、武田軍が織田軍と同格に戦える、つまり両軍の同時全滅に持ち込めるためには、織田軍に対する武田軍の術力は、式(3)を用い、

$$r=34000^2/12000^2=8.03$$

の高さでなくてはならない。しかしいくら武田勢の騎馬隊が強かったとしても、新兵器を持つ織田勢の鉄砲隊に対して、これだけ高い対敵術力は、なかなか発揮できなかったであろう。

設楽原における長篠の合戦は奇襲戦ではなく、また特に地の利の差もなく、両者ともに堂々と陣を張っての正面対決であったのだから、武田勢の敗北は戦いの始まるまえから予測できたことであったと思われる。

織田側の進攻の目的は必ずしも武田勢を打ち負かすことではなく、武田勢による長篠城の囲みを解くことにあったのだから、この兵力の格差を見せつけられた武田勢は、陣を畳んで引き揚げるか、少なくともまだ甲斐に残してあった1万の兵の増強を待つべきであった。それでも3万4千対2万5千の戦いでは、武田側の対敵術力が少なくとも$r=34000^2/25000^2=1.85$はなければ、同格に戦えないことになる。しかしこのぐらいの対敵術力ならば、あるいは何とか発揮できたかもしれない。

一般に当時の合戦のあとには、特に敗北した側は雑兵の逃亡者が多く出るので、残存兵力数は必ずしも戦闘による兵力の直接の損失を反映していない。したがって、長篠の合戦が終わった時点での両者の実質的な兵力損失量について調べてみることはできなかったが、もしそれらが分かれば、やはり式(2)を使って、当時の騎馬軍団と鉄砲隊との術力の比$r$がいま計算できるはずである。

アメリカ軍の推定値によると、第1次世界大戦のときの手動式ライフル戦での使用弾数対敵兵殺傷数の比率は7000発に1件、第2次大戦時の自動小銃と手動小銃の混用では2万5千発に1件、朝鮮戦争時の自動小銃戦では5万発に1件だったという。

戦闘状況も戦術も、また銃の操作性もこれら現代の戦いとは全く異なっていた長篠の合戦における鉄砲の威力を、こうした数値と比較してみるのは大して意味のあることではないかもしれない。それでも長篠の合戦における火縄銃の効力は、それに対応する戦法の立て方によっては、それほど大きなものにさせずに済ませることができたと思える。したがって長篠の合戦の正面攻撃における織田側の大勝利は、やはり両軍のあいだの兵力数の格差によるものと考えるのが正しいのではないだろうか。もちろんこの合戦のあと、鉄砲は戦闘にますます導入され、それとともに戦略、戦術、戦法が大きく変わっていったのは歴史的事実ではあるが。

## 11　兵力の分散逐次投入は敗北への道

いままで現代戦の実例としては海軍のことを書いてきた。しかし海軍に劣らず、日本陸軍にもおなじような数理的理論の軽視はあった。その中から、ここでは戦闘にあたって、持てる兵力を全力投入せず、小出しに逐次投入して敗退した例をとり、その理由付けを考えてみよう。前節で述べた長篠の戦いにおける武田勢の用兵も、実はその一例といえるものであった。

この兵力の分散逐次投入という作戦は、日本のもう一人の神格化された軍人、陸軍の乃木希典大将が、日

露戦争(1904年～1905年)のとき、旅順港外のロシアの堅陣、203高地において、多大の犠牲を払ってこれを落としたときに使われ、それ以来日本陸軍の手本のようになったものである。しかし忘れてならないのは、このときの日本軍の勝利は、当時ロシア極東軍の総司令官であったアレクセイ・N・クロパトキン元帥(1848～1926)が回想録(1902)[13]のなかで、「日本軍の砲弾と山砲は、我が軍のよりもはるかに優れていた」と述べているほど、まさった装備があった上でのことであった。

戸部・他(1984)[21]の分析によると、日本の敗北に決定的な影響を与えた、日米両軍間の本格的な正面衝突は大東亜戦争中に6回あった。すなわち、1939年のノモンハン事件(陸軍)、1942年のミッドウェー作戦(海軍)、1942年のガダルカナル作戦(陸軍)、1944年のインパール作戦(陸軍)、1944年のレイテ海戦(海軍)、そして1945年の沖縄戦(陸軍)である。

このうち、少なくともノモンハン事件とガダルカナル作戦の二つでは、持てるだけの兵力を一気に投入せず、何回かに分けて小出しにしたために、兵力を不必要に消耗してしまったと言ってよい。

作戦の具体的な詳細は、例えば上記の戸部・他(1984)[21]によってかなり詳しく本筋をたどることができるが、さらに詳細を知りたければ、いまでは数多い資料が刊行されている。ここでは先に求めた式(2)に具体的な数値を与えた例によって、兵力の分散逐次投入が全戦力に与える効果を理論的に考察する。さらにもう少し詳しく一般的な取り扱いについて見たいときには、(中学レベルの算術的演算を用いて求めた)付録にある諸式を参照して欲しい。

いま敵・味方のおのおのが同じ対敵練度の戦闘単位を3万ずつ用いて戦うものとする。この単位としては何でもよいのであるが、ここではたとえば歩兵の数を念頭において具体的な心像を描けばよい。また、この3万ずつという数は、実戦において妥当な規模として選んでみただけで、敵・味方が同数であれば、計算は全く同じ結論を与えてくれる(詳しくは付録の式(p)を見よ)。

さて、敵が全兵力を一気に投入してくるのに対して、味方は3万の単位を1万ずつ3回に分けて逐次投入するとする。またその第2次および第3次の投入時期は、味方の残存兵力が各投入時の兵力の2分の1になったときとする。

すると当初の戦闘は、式(2)において $e_1=30000$、$f_1=10000$、$r=1$ で始まる。したがって、味方の兵力が半減したときの敵の兵力は $f=5000$ として求められる $e$ である。すなわち、

$$30000^2-e^2=10000^2-5000^2$$

から、簡単に $e=28722.8$ となり、敵兵力の損失が1277単位ほどなのに対して味方の損失は5000であるから、すでに味方は敵の3.9倍強の損失をこうむっている。

ことの重大さに驚き、味方はここでさらに10000の兵士を投入して兵力を増強することを決定する。すなわち、この時点では式(2)において新しく $e_1=28722.8$、$f_1=15000$ となる。そこで、さらに味方の兵力が半減して $f=7500$ となるまで戦ったとすると、その時の敵の残存兵力は、式(2)によって、$e=25617.4$ となり、味方の新たな損失7500に対し、敵兵力の新たな損失は3105.4にすぎなく、この第2回の戦闘における味方の損失は、敵に比べて2.42倍ほどになる。

そこで味方は、残る10000の兵力を投入して最終決戦をいどむことにすると、式(2)で新たに $e_1=25617.4$、$f_1=17500$ となり、味方が全滅するまで戦った結果として、敵はまだ18708.2の兵力を残していることが求まる。つまり味方は30000の兵力を全部消耗してしまったにもかかわらず、敵に与えた損害は11291.7だけだから、敵にはまだ初期兵力の62.4パーセントが残っていることになる。

初めから全力を投入すれば互格に戦えた相手であるのに、兵力を3分して逐次投入するということをしただけで、戦いの勝敗に、理論上これだけの差がつくということは、かつての中学4年生の数学力をもってして、こんなに簡単に示すことができる。にもかかわらず、こうした兵力の分散逐次投入は、大東亜戦争を通して、日本軍、特に陸軍において顕著に繰り返され、それでなくてさえ乏しかった兵力を、みすみす失ってしまったパタンであった。戦術・戦略のプロであるべき軍部に、なぜそうした愚かさを防ぐことができなかったのだろうか。

特にそれが長期戦を苦手とし、ことあるごとに短期決戦を志向していた日本軍の行動であっただけに、いよいよもって腑に落ちないものがある。

## 12　合理的思考の欠如

筆者は日本軍内の犯罪を処罰する法律についての知識は持ち合わせていない。しかし日下(1993)[12]によれば、かつての日本の陸軍刑法には、負けるに決まった戦闘を始め、その結果、現に負けた指揮官は軍法会

議にかけて処断されるということが明記されてあったという。おそらく海軍の刑法にも、同じような条文があったことであろう。

にもかかわらず、実状として、そのような処置がとられたことがなかったということは、誤った判断をした責任者を誰も処分できないほど、組織としての日本軍が無法集団だったからではなかったと思う。むしろ問題は、当時の指導者たちが、ここに述べてきたような初歩的な数理的分析でさえも、自から実行するだけの能力と自主性とを持ち合わせていなかったので、それに考えが至らなかったのだと思う。

言うまでもなく、かつての軍の指揮官たちすべてが、そのような判断力に欠けていたわけではなかったであろう。たとえば、大東亜戦争における陸戦の天王山と言われた、1942年夏から秋にかけてのガダルカナル島作戦において、遙か離れたニューブリテン島ラバウルにあった軍司令部から、ルンガ飛行場(ヘンダーソン飛行場)南の米軍の堅陣を攻撃略取すべしとの命令を受けて、第2回総攻撃を実施する立ち場にあった二見秋三郎参謀長は、彼我の戦力の差のあまりにも大なるをもって、作戦の無謀さを主張したがゆえに更迭されている。さらに、現地派遣の精鋭4個大隊から成る、川口支隊の支隊長川口清健少将も、その後になって、同じ理由によって敵陣への正面攻撃を避け、迂回作戦を提案したがために、またもや罷免された。しかし結果的にはかれらのほうが正しく、この第2回目の総攻撃も再びわが軍の完全な敗北に終わったのであった。

その後1944年のインパール作戦にあたっても、現地の惨状を知らない中央からのあまりにも無謀な命令に正面きって反抗した指揮官のあったことは、インパール抗命事件として知られている(高木　1976[16])。

また1944年末から1945年の日本の降服にいたるまでの、フィリピン群島における日本陸軍の奮闘については、圧倒的に優勢なアメリカ軍の制圧下にあって、すでに現地から日本への情報伝達の手段がほとんどとだえていた。大本営は日本軍劣勢との情報が日本に伝わるのを恐れて、当時フィリピン群島に在住していた日本の非戦闘員、特に婦女子までをも、日本に送還することを拒否した(岡田　1972[8])というようなことさえあって、内地では戦闘の実情もよく分からなくなっていたから、戦闘の報道らしいものは全くといってよいほど当時の記録に見られない。しかし陸軍は50万ないし60万人の兵力を投入し、戦争終結時までにはその80パーセントを失なうという、一方的な負け戦であった。

それでもその裏には、自からはいち早くサイゴンに逃げ出しながら、兵にはアメリカ軍との正面対決による決戦を指示した、南方軍総司令官寺内寿一元帥のフィリピン方面での戦闘方針を受け入れず、持久戦によってアメリカ軍を引きとめて日本本土への攻撃を遅らせるという作戦をとった、フィリピン防衛の責任者、第14方面軍司令官山下奉文大将の適確な判断があった(岡田　1972[8])からこそ、日本陸軍の残り20パーセントも生きながらえることができたのである。ちなみに、とても制度化していたとは思えないが、日本軍では戦局が不利になると、機会があればあとを部下にまかせて、指揮官が戦線を離脱することは珍しい例ではなかったようである(内藤　1976[22]参照)。

このように、日本軍の大敗のほとんどは、合理主義と自主的判断を排し、精神主義と教条主義によって、兵力の逐次投入作戦を強行させた軍司令部がもたらした悲劇であった。にもかかわらず、そうしたかれら司令部の命令が、陸軍刑法の定めていた処罰に該当する行為であることを意識した者が、中枢にさえいなかったということは、やはり陸軍には、戦略・戦術を客観的、数理的に分析する能力が欠けていたからだとしか思えない。

## 13　作戦計画の不在

再び海軍に戻ると、すでに述べたように、東郷司令長官はロシア艦隊に対して圧倒的な砲員数を保有して日本海海戦を戦うことができた。したがって、その作戦参謀としては、一部のすきもない綿密周到な正攻法的戦術で知られた、秋山真之少佐を頼みにすることができた。

しかし、最後には日米戦争に臨まざるを得なかったものの、連合艦隊司令長官山本五十六大将は、すでに本稿で検討した如き、彼我の戦力の員数的格差による不利を熟知していたからなのであろうが、米内光政大将、井上成美少将とともに、アメリカを刺激する日独伊三国(軍事)同盟に最後まで反対し、さらに対米参戦にも終始反対したことで知られている(阿川　1965[1]、1986[2]、生出　1983[6]、宮野　1982[33])。

そのいっぽうで山本は、圧倒的な彼我の戦力の格差を補うためには、参謀としても秋山少佐の前例のような緻密な作戦の専門家では不充分とみたのであろうか、日米戦争に当たっての先任参謀として、人が思いつかないような奇想天外なアイディアを追い、かつ山カン的な思いつきを珍重するような奇才、黒島亀人大佐を

頼りにせざるを得ないと、最後まで信じていたという。

もともと山本は軍政にかけての第一級のプロではあったが、用兵の中心となる作戦は専門ではなく、また時の連合艦隊参謀長宇垣纏少佐も作戦の専門家ではなかったので、「山本司令部には山本をふくめて、プロの作戦家は1人もいない」まま、ハワイ作戦を敢行したとさえ言われるが(生出　1983[6])、山本としては戦力的に劣勢の日本海軍を指揮するのであり、しかも生来のバクチ打ち的な性向もあって、戦略の策定にあたっては奇策に頼らざるを得ないことを痛感していたのであろうか。

なお軍令大権と軍政大権の与えるものの違いについては、たとえば内藤(1976[22])を参照していただきたいが、要するに軍令は戦う兵術に関わり、軍政は戦うに必要ないっさいの支援に関わる。アメリカやイギリスと異なり、日本軍においてはこの2者が対等の立ち場でなかったことが、しばしば大きな問題となった。具体的には、たとえば戦訓によって航空母艦信濃の脆弱性を知った技術官による、早急な改装の進言を、軍政の専門家でありながら用兵上の立ち場から退け、2年後にむざむざと海の藻屑となさしめたのは、連合艦隊司令長官山本五十六大将その人であった。また駆潜艇の兵装について、技術官らが8センチ砲の不可欠を主張したにもかかわらず、用兵上のことは「技術官の発言するところにあらず」ときめつけて、のちほどやはり無用と分かることになった4センチ銃を採用させたのは、最後の連合艦隊司令長官となった豊田副武少将であった。

それでも山本には、兵力の分散逐次投入の不利についての充分の心得があったであろうから、海軍軍令部の強硬な反対があったにもかかわらず、1941年の日米開戦にあたっては、連合艦隊司令長官の職を賭け、動員可能な兵力を総結集してハワイの真珠湾攻撃を強行している(阿川　1965[1]、生出　1983[6])。

また1942年8月のガダルカナル島争奪戦のときには、当初陸軍は4分の1個師団(約2500人)で作戦に充分であると考えたのに対し、山本は5個師団(約5万人)に加うるに、海軍も航空機を主力とした全兵力を集中することを大本営に具申している。にもかかわらず、結果的には中央部の方針により、兵力はかなり分散逐次投入され、その結果陸軍は大敗を喫したのである。

しかし、そうした山本の作戦も、正攻法では勝ち目がないと判断した賭け事師のギャンブルであったと言うべきで、かれのその後の一連の作戦は、プロの作戦家から見ればアマチュアの愚策に終始したという批判は当たっているようである(生出　1983[6]、千早　1982[19]など参照)。

しかも生出(1983)[6]によると、山本は水から石油が作れるといった非科学的なことを信じていて、同僚や部下を手こずらせたというから、上に述べた山本の判断も、真に数理的な裏付けをもったものであったというよりは、やはりバクチ打ちの勘でしかなかった可能性が大きい。

ちなみに時の首相兼陸軍大臣東条英機大将も、やはり水から石油が作れるということを信じていたという。溺れる者は藁をも摑むの心理が働いていたのだろうか。同じ東条内閣のもとでは、原料の不足もあって鉄材の不足に悩まされた結果、熔鉱炉も使わず、アルミニウムを砂鉄といっしょに燃やして鉄の生産をするという、エネルギーの根本原則を全く無視した製鉄工場の建造が、時の国の指導者たちによって真剣に取りあげられ、もう少しで実現するところであった(中谷　1947[23]参照)。また松の木や根から、飛行機用の燃料や潤滑油を生産するといった、エネルギーの原理および材木資源の保全を全く無視した国を挙げての努力(内藤　1976[22])も、現実には同じように役に立たないものであった(中谷　1946[24])。そうした愚かな騒ぎのために、重要な原料生産計画が混乱させられたことがあったのは言うまでもない。

そうした中で、大東亜戦争の開始時の真珠湾攻撃に始まり、1943年4月18日ブーゲンビル島バラレ近くにおいて航空機上に戦死するまでに、彼が遂行させたすべての海戦において、山本は艦隊による戦力を過小評価し、航空戦力のみを過大評価するという誤ちを犯したがために、アメリカの陸海空軍の緻密な協同作戦のまえに、日本軍はいたずらに航空戦力のみを消耗してしまい、のちには丸裸となった艦隊戦力が、実力を発揮する機会を与えられることもなく、滅亡させられる道を拓いたのである(生出　1983[6]、千早　1982[19]参照)。

### 14　自からを縛った戦艦無用論と陸海軍間の不協和

そうした、航空戦力と艦隊戦力といった異種の戦闘単位を、グループ別にして分散逐次投入したときの彼我の兵力の相対的消耗比の計算には、第11節で述べた計算よりはずっと詳しく、各種のパラメタ(媒介変数)を導入した上での数理的検討を要することは明らかである。にもかかわらず、第11節で示した、単一戦闘単

位の分散逐次投入が不利であることの数理的分析を理解されたかたは、ここでもやはり同じような原理が働いて、分散逐次投入が損失の増大を伴うであろうことを、容易に感じ取ることができることと思う。

したがって、新興兵器としての航空機による戦術に魅惑され、いたずらに戦艦無用論を唱えて、両者による総合戦力の威力を理解し得なかった山本五十六連合艦隊司令長官、およびその幕僚は(生出　1983[6]、1988[7])、やはり自からの理解の能力の圏内にあった数理的洞察力を駆使して考えるという自主性に欠けていたと言わざるを得ないであろう。

さらに考察の枠を広げると、大東亜戦争に関して日本では、開戦まえから陸軍と海軍とのあいだに種々意見の相異があり、開戦後も陸海軍のあいだで、真に渾然一体となった協同作戦というものは、ほとんど見られなかったと言ってよい。戦略的に見れば、これは明らかに兵力の分散逐次投入の変形に過ぎない。したがって日本は、ただでさえ乏しい国力を構造的に幾重にも分散逐次投入することによって、ますます自からの戦闘能力を弱めていたと言える。

それももとを正せば、日本には対米戦争を遂行するだけの戦力がないとして、海軍内で開戦に反対をしていた情報通の米内光政大将、山本五十六大将、井上成美少将らに対し、あらゆる機会をとらえて圧力をかけ、ついに海軍を屈服させ、戦後に井上をして「長い闘争の相手は国内にいたのだ」と歎かせた(生出　1983[6])ほど、戦争のみに軍人の生き甲斐があると考えていた、たとえば参謀辻政信大佐のように陸軍の中枢にいた者(杉森　1963[15])、および、ずるずるとそれに同調していった海軍の中枢部(生出　1983[6]、宮野1982[33])に、あの無謀な戦争への責任があるということであろう。

今日のように各種の科学兵器の発達した時代においても、戦争の最終にして決定的な詰めは陸上における敵陣および領土の占領によるものである。その他のいかなる軍事行動も、また兵器も、この最終目的をできるだけ速かに達成する手段にしかすぎない。したがって海軍の本質とは、敵国の占領に先立ち、敵国の海岸を封鎖し、敵国の基地や生産設備を破壊し、敵国の商船を撃沈して、敵国の経済に打撃を与え、戦闘能力を失わせることにある(内藤　1976[22]参照)。

言うまでもなく、敵国の海軍も同じ意図のもとに行動し、攻撃を仕掛けてくるから、時に及んでは味方の海軍は敵側の意図を砕くべく、海戦において敵側の戦闘能力を破砕する必要は避けられないが、あくまでもそれは副次的なことである。

したがって、島国であり、かつ基本的な資源を海外の産地に頼らざるを得ない日本は、必ずや海上交通を敵海軍におびやかされるであろうから、日本海軍は船団護衛、対潜水艦戦、対機雷戦、防空戦など、通商の防衛を中心に捉えた作戦計画、戦闘計画を策定すべきであった(新見　1995[25]参照)。

しかし、当時の海軍は、その使命の本質を理解せず、狭い視野に立って、ひたすら敵海軍との砲撃戦を主務と心得た、大艦巨砲主義に走り、作戦全般における航空機の重要性さえも本当に理解していたとは言えない状況であった。

かくして日本海軍の永年にわたる努力の結果は、「ある特定の参考書のみを全力をつくして勉強し、内容ことごとく脳裡に暗んずる自信下に受験し、問題すべてヤマを外れた」(福井　1956[27])ことになってしまったと評されても仕方がなかったと言えよう(内藤1976[22])。

残念なことに、今日といえども、われわれはそうした「井の中の蛙」的な思考形態から解放されているとは言えない。国による検定教科書、大学入試センターの試験などで統制されている今日の教育行政は、国際社会においては国として盲点だらけの民衆を育てることになっているからである。

そのほか日本軍には、自信過剰の先入観、情報収集の不徹底、入手情報の分析軽視、味方情報の防諜軽視、科学兵器に関する無関心、潜水艦戦略への無理解、航空機搭乗員教育制度の不備、彼我の兵站能力の重要性に対する理解の欠如、命令系統の不徹底など、挙げればきりのない精神的、知的なたるみがあったことも(内藤　1976[22]、千早　1982[19]、生出　1983[6]、児島1987[14]、奥宮　1989[9]、NHK 1993[3])、日本軍の敗因を必要以上に積み重ねたことは明らかであるが、詳しいことは本稿の主旨からだいぶ外れるので、これ以上はここで立ち入らないことにする。

ただひと言つけ加えるならば、第2次大戦は国と国とのあいだの総力戦であり、おのおのの国がその軍隊のほかに、国のあらゆる生産力を動員して戦い、かつその生産力をもお互いにつぶし合った戦いであった(エンソー　1956[5])。したがって、ここに述べてきたような数理的考察は、もっと一般的には、国の軍事力、人口、工業生産力などのすべてを総合して員数化し、かつおのおのの対敵練度や術力を考慮に入れた上での

戦闘力の計算となってくる。

そう考えると、人口がアメリカの約2分の1しかなく、工業生産力もあれこれ平均すると数分の1であり、さらに、労働者1人あたりの生産性が年とともに上昇していったアメリカに比べると、もともと技術力がなく、しかも戦争とともに人的資源の浪費が進んでいったために、かえって生産性が落ち込んでいった日本の事情(神原　1946[11])は、総合戦力という点から見ると、構造的には兵力の分散逐次投入に似た、きわめてあやふい戦力の管理が行なわれたものだったと考えてよい。したがってその結果は、兵站的にも、戦場において兵力を分散逐次投入することを余儀なくされることに直接につながり、戦略的に日本は、はじめから大きな敗因をかかえた上で、あの戦争に突入したのであった。

## 15　なぜ数理的思考が不得手になるのだろうか

以上に述べてきたような、ものごとの判断にあたって数理的な分析的考察力が乏しく、ともすると均一集団主義、情緒主義、精神主義に走るという日本の傾向は、実は何も軍人だけに限られていたことではない。それは日本で指導的地位にあった者たちのあいだにおける一般的な傾向であったし、また国際的に比較してみたとき、今日といえども日本の指導者層に特に目立っている特徴であると思う。

そうした傾向は、日本文化の長い発展過程において徐々に培われてきたものであろう。特に日本では、いわゆる理系の出身者と対比して、文系の出身者に数理的思考を毛嫌いする傾向が顕著である。

たとえば筆者の経験に限ってみても、アメリカの大学では情報科学系の大学院に、法律や経済、心理学、社会学などを専攻した、いわゆる文系の学士学生が入学をしてくることはさして珍しくないが、日本ではまことに例外的な珍しさである。そしてその根底には、日本の文系の学生のほとんどに、数理的な思考力と知識とが不足していて、理系の授業について行けないことがあると言っても言い過ぎではない。

こうした文系と理系とが2極化して解離する傾向を持つに至ったことには、日本の文化の長い歴史の中に、いろいろと原因が見いだせることと思われる。そのすべてを検討しようとするのは本稿の主旨を越えることになるし、また筆者のよくするところでもない。しかしながら、日本とアメリカの両国で大学教育を受け、その後アメリカにおいてほぼ15年間も大学教育に携わり、あるいは大学と企業において研究を行なったことのある筆者の経験から、ここではその原因の一つとなっていると思われることについて、ひとこと述べてみる。

それはわれわれが用いている文字が、われわれの教育過程、そしてその後になっては、われわれの思考形態に与えている影響である。

第1に、漢字は表意文字であるとよく言われているが、それが正しくないことについては、すでにいくつかの実験によって認知科学的に調べられてある(たとえば　Horodeck 1987[29])。とにかく、表語文字である漢字は語の増加に合わせて作られてきたので、その数が多く、したがって充分な読み書きができるようになるまでにあまりにも時間がかかりすぎる。

たとえばアメリカなどでは小学校の5年生にもなれば、大人の用いる百科事典をそのまま使って、自から知識の獲得や学習が可能であるが、日本では高校卒でも、まだ使えない百科事典は多い。それどころか、一人前の知識人でさえ、たとえば「あご」「わき」「へそ」「くるぶし」「かかと」といった、日常なんでもないことばの漢字が書けなかったり、書き誤ったりすることは珍しくない。アルファベットを用いている国では、知っていることばのほとんどは、すぐさま文字で書き表わせるのに、日本ではそれはむしろ珍しく、大学の教授でさえ、手もとに辞書なしでは文章を書くのに不安な者が少なくない。

こうした、文章による情報伝達の手段が複雑でむずかしく、したがってその能力の獲得が非能率であったことが、戦前戦中にかけての兵士の教育に大きな負担となっていて、それがわが国の戦力に対してもかなり不利に働いていた(保科　1949[28])。

ことばのそうした複雑な表記体系を依然として使っている結果、いまどういうことが起こっているかといえば、まず学習が受け身になり、自から積極的に調べて考えるという習慣が身につきにくい。かつ、小・中学校での学習の多くの時間が、ただ漢字の読み書きといった暗記と再現の能力を身につけることに費され、いかに情報を分かりやすく明瞭に伝達するかという文章能力が充分に習得されない。またつね日ごろ、文字習得に付随して暗記が重要視されている結果として、数理的な分野においても分析を自から行なうというよりは、与えられた問題の模範解答を丸暗記して答えを書くという、受け身の数理的能力の習得に片寄り勝ちである。そのことは、たとえば国際的な数学コンテストで、日本は中・高生までは高い得点を得るが、解答

に創造性が必要とされる大学レベルのコンテストでは、日本の成績は国際的にかなり下がってしまうという実状に反映されている。

さらに、数多くの漢字を学習させられているうちに、それ自体が目的と化し、むずかしい字をやたらに使うことをもって尊しとするような考えに染まり、文字とは情報内容をやさしく明確に伝えるために使う道具であるといった考えを、かえって幼稚だと考えるようにさえなってくる。しかもそれらの人たちに限って、義務教育の範囲内の数学を活用することはおろか、それを用いた記述にさえついて行けない、あるいはついて行こうとしない者がかなり多いようである。

### 16　日本軍の戦力を削いだ表記法

そうした思想が見え透けていたのが、大東亜戦争の終りにいたるまでの日本の軍隊の文書であろう。軍の上層部でやりとりされた、たとえば(軍閥)内閣で決定した戦争指導大綱のような文章でさえ、やたらに漢語をちりばめただけの、情緒的、抽象的、観念的なものであり、内容についてはほとんど具体的な指示のないものであって、もし筆者の学生がこのような指導文を書いてきたとしたら、とても合格点は与えられないというようなものであった。また当時は、平均すると漢字の500字も充分に読み書きできない兵隊が多かったにもかかわらず、兵士に対する伝達文にさえ、やたらとむずかしい漢字・漢語がたくさん使われ、情報の疎通にさしつかえた例が、戦後になってあれこれと報告されている(山田　1997[37]参照)。

当時、兵器の部分名称だけでも4000以上に達しており、日常語や和語を用いればなんでもないものが、ことさらにめんどうな漢字を並べた「漢語」が用いられ、兵隊たちはもちろん、将校たちさえもそれに苦しめられていた。たとえば試みに「チューラ」「ホーテツキ」「トボクキ」「キューガクレーキャクユソクトウ」などのことばがすぐ分かり、かつ漢字で書ける者が、今の大学卒の人たちの中にどれだけいるであろうか。いまここにこれらの意味だけを書いておくと、それぞれ「ボルト」「ミシン」「くつブラシ」「ピストン冷却油ポンプ」である。試みに漢字を書いて当てはめてみていただきたい。当時は漢字500字でさえ、満足に読み書きできない兵隊が多かったのだから、これでは情報の伝達が思うにまかせなかったのも無理はない。

そうした経験を踏まえて陸軍省では、やっと大戦の末期にいたって、兵器の名称や日常活動に使われる用語のための漢字の総数を600字以下に制限する研究を始めたが、それも自分では戦場におもむかなかった軍国主義者や、当時の敵国の文字であった漢字に固執した超国家主義者の強い反対にあって、思うにまかせられなかったという。

第2次大戦中、1942年にウェーク環礁を占領した日本海軍の設営隊員が、ショベルとモッコを用いた人海戦術で滑走路の補修を始めたところ、捕虜の1人が、そのへんにころがっている土木工事機械を使わせてくれれば、自分1人で充分だと豪語したので、半信半疑でやらせてみたところ、みごと1人で300人分の作業をこなしたというのは有名な話しである。

そうした教訓を受けて、日本でも捕獲機を参考にして、急ぎ土木工事機械の開発を始め、1943年にはなんとか試作機が実用に供せられだした(内藤　1976[22])。

そのとき、すでに兵隊の識字能力の低さが身にしみていたので、戦前のように耳で聞いても何のことかよく分からないような、音読みのむずかしい漢語の代わりに、和語を用い、ブルドーザーを押均機(おしならし)、キャリオールスクレーパーを鋤取機(すきとりき)、パワーショベルを掬揚掘削機(すくいあげくっさくき)、グレーダーを敷均機(しきならしき)、トレンチャーを溝掘機(みぞほりき)などと呼ぶことにしたという。

ひとたび説明を受けてあれば、これらは耳で聞いてすぐ分かる用語であったにもかかわらず、かなを使わず、依然として漢字で書かれていたために、多くの兵隊には読むことができず、やはり難解な用語とされていたようである。そんな調子だから、せっかく製造されたこれら土木機械も、日本軍はせいぜい人力作業の補助ぐらいにしか使わなかったそうである。

漢字語の欠点は、当初はそれが和語の漢字表記であっても、漢語を貴しとし、和語を幼稚とする偏見の強い日本では、いつのまにかそれが音読みされるようになってしまうことである。身近な例をとってみても「もみじ(紅葉)」が「コーヨー」になり、「やまずみ(山積)」が「サンセキ」になり、「いれふだ(入札)」が「ニューサツ」になってしまったようなことは、たえずおこっている。もし第2次大戦がもう10年か15年続いていたとしたならば、おそらくこれらの用語もいつしか押均機は「オウキンキ」、鋤取機は「ジョシュキ」、掬揚掘削機は「キクヨークッサクキ」、敷均機は「フキンキ」、溝掘機は「コウクツキ」と、耳で聞いて何とも訳の分からないものに変わってしまったであろうことは想像に難くない。

なお、軍においてことさら難解な用語が用いられたのは、兵に地図が読めれば戦場から逃げ出せ、また水兵に海図が理解できれば軍船で反乱を起こす可能性があったから、統率のための無意識のくふうだったのではないかという黛(1998)[32]の指摘には驚かされた。

しかし考えてみると、第２次大戦後になるまで、われわれは外国のラジオ放送を聞くことが法律によって禁止され、当然それを許すラジオの販売者、所有者は罰せられていたのだから、「寄らしむべし、知らしむべからず」といった差別的な思考の枠が為政者の常識であった時代の中国において、民衆を無知に留める目的で作り出されたといわれる、ことさらむずかしい漢字をそのまま後生大事に受け入れた日本文化内でも、将校が兵の読めない表記法に固執したのは、そう驚くべきことではなかったのかもしれない。

その当然の帰結として、国や軍の統率はピラミッド型の中央集権主義によるものとなり、局所戦の状況に応じて指揮権が柔軟に行使できる、今日で言えば分散処理的作戦がならず、戦略が硬直化していたという黛の指摘には同感である。さらにそうした用兵思想が通信技術の開発をないがしろにさせ、その結果日本軍は世界における戦略戦術のパラダイムの変化についていけなくなり、それが大きな敗因の一つとなったことは明らかである。

そしていまだに問題なのは、日本の政治においてはあい変わらず同じような考え方が支配的で、情報の公開や地方分権などの政策に消極的であるということであろう。

## 17　日本語を歪める漢字偏重

実は文字に関しても、そうした思潮は今日といえども大して変わっていない。嘘だと思うなら、最近の特許明細書に出てくる、難解にして、読み方の分からない用語の羅列を眺めてみることをおすすめする(高橋1995[17]参照)。

たとえば「嵌載」「遊転」「串装」「介設」「枢着」「弾発」などの表記は辞書にもなく、また学術用語集にもない。おそらくこれらは「カンサイ」「ユーテン」「カンソー」「カイセツ」「キューチャク」「ダンパツ」などと読ませるつもりで書かれたのであろうが、耳で聞いたのでは何のことか分からないばかりでなく、目で見てさえ、まだ何のことか分からない。

したがって、国際的に流通しなくてはならない性格を持つ公文書として、こうした勝手気ままな用語法を野放しにしている日本の特許制度は、国際的に批判の的となっているということである。

もう一つ例を挙げておくと、テレビ朝日(東京ではチャンネル10)での田原総一郎司会の番組み「サンデープロジェクト」では、1998年１月11日には日本の経済予測に関する討論が取り上げられていた。討論者の中には自由民主党の若手代議士渡辺喜美があったが、討論の中で彼が「シワク」「シワク」と連発していた。はじめは何のことか分からなかったが、そのうちに脈絡からして「思惑」のことだと分かった。文教族なる集団を組んで、国の国語政策に、国際的かつ長期的には必ずしも良いとは言えないものがあるにしても、かなり強い影響を与えている自由民主党の議員の一人にして、このありさまである。

もともと「おもわく」の「く」は、「いわく」「おそらく」などの「く」と同じ、未然形接続の接辞であったものが、「おもわく」の名詞的用法が進んで、これに「思惑」という湯桶読みの当て字が用いられるようになったものとされている。しかし、いくらことばが時代とともに変わるとは言っても、「シワク」はあまりにもひど過ぎないだろうか。これも漢字偏重の産物の一つである。

ついでに書くと、同じ番組みに出席していたもう一人の討論者に某経済研究所の主任研究員紺谷典子があった。彼女は討論の中で「ゴウワン」ということばを使っていた。辞書にはないが、前後から推してこれは「強腕」であったと思われる。もともと日本語には「すごうで」ということばがあり、漢字では「凄腕」と書かれるし、また漢語としては「辣腕」がある。しかし「凄」や「辣」はあまり使われる字ではないので、これらを避け、「強」の辞を用いて新語を作ったものであろう。「すごうで」は「すご腕」で充分であるのに、何でも漢字にしないと幼稚と思う潜在意識が、「ゴウワン」のようなことばを作らせたのではないだろうか。

ここで注意したいのは、このことばが文字に書かれたのではなく、口による討論に出てきたことである。その過程としては、おそらく頭の中にこうした漢字がたくさん転がっており、言わんとする意味を表わすのに、聞き手による理解度を考えず、ただそれらしい字を拾って並べるという安易な処理が行なわれたものであろう。これでは口頭による情報の伝達がむずかしくなるのも無理はないであろう。

話しかたが速すぎて放送の内容が分からないという苦情が、NHK には全国で年間に数万件も寄せられてい

るという。その内訳は若者よりも老人に多いと言うものの、その本質は老人の聴力の減退にあるのではなく、耳で聞いて分かる言葉を使うというルールを無視した造語法の乱用によって、こうした新語の乱発が行なわれ、ただでさえ多い同音異義語がますます増えているからではないのだろうか。

上に述べたような、漢字で表記された和語の音読みへの転化を防ぎ、あるいは漢字語を読みやすくする一つの方法は、戦後の一時期実行されたように、日本語において変化する語尾の部分は、少数の例外(たとえば「すくない(少ない)」と「すこし(少こし)」のようなもの)を除いて、すべてかなとしておくことである。本稿においては一つの試みとして、国語審議会によって示された、送りがなの新しい「目安」ではなく、こうした表記法を徹底して使ってみた。

前節で述べた土木工事機械の例にあっても、もし「押し均し機」、「鋤き取り機」、「掬い揚げ掘削機」(この場合い、長く慣用されていた掘削については立ち入らないことにする)、「敷き均し機」、「溝掘り機」と書けば、読み方もすぐ分かるし、その音読み化も防げるであろう。

しかしながら、国民の真の国語力のことをわきまえない者たちの横車と、少しばかりの書く手間(これはワープロの出現によってすでに解決ずみ)と紙面上のスペースをけちるという近視眼的な考えから、ふたたび送りがなを大幅に削る表記に戻してしまったのは、日本語の将来にとって、大きな問題を復活させることになってしまったと言ってよいと思う。

日米戦争の初頭を飾った、ハワイの真珠湾攻撃は、いまでは日米両国の専門家によって、戦略的にも戦術的にも失敗であったという評価が固まっている。戦闘機乗りが上手だということで戦前に有名となり、年功によってこの真珠湾攻撃実施の航空参謀になった源田実少佐は、実戦においては自から考える能力の足りない、平凡以下の参謀だったようであるが(生出1983[6],1988[7]、戸部・他　1984[21])、その彼が戦後参議員議員に立候補したときに送られて来たあいさつ状は、やたらにむずかしい漢字・漢語の羅列から成っていて、筆者は呆れたのを覚えている。

その彼が当選したのはどんな理由によるのか、筆者には明らかではないが、すでに民主主義の時代になっていたのに、その文書には広く大衆に話しかけ、趣旨を徹底させるというような気配が微塵もなく、戦時中の独りよがりをそのまま引きずっているものであった。これなども極端な漢字教育が、必然的に落ちつかせる先としての弊害の一つであろう。

### 18　使用文字の型が思考過程に与える影響

そのほか、近年になって科学的に少しずつ明らかにされだしたこととして、漢字の使用がわれわれの思考過程にある種のひずみを与えるらしいことがある。

むかしから漢字についていろいろと言われてきた、いわゆる表現力の豊かさなども、その裏を返すと、ともすれば事実を誇張するという、一種の歪曲効果であると考えられるが、手塚(1987)[20]などは、それを文科系の目でかなり正確にとらえている文章である。

使われている文字の型の違いが思考の型に与える影響については、科学的にも心理学的実験に基づく心理物理学や認知心理学の研究によって、1970年代から少しずつ明らかにされてきているし(新しいところではたとえば　田中　1997[18])、また筆者自身もそれらについては何度か解説を試みてあるので(たとえば　山田1991[35]、1997[37])、ご関心のおありのかたがたは、ご参考にしてしただきたい。これをごくつづめて言えば、漢字を常用にしていると、思考が客観的・合理的でなくなる傾向を示すようになるということである。

しかもその代償として、むかしから言われて来た、漢字が持つとされるさまざまな利点の主張のほうには、実は科学的根拠が乏しく、それらは、最近アメリカで特に脚光を浴びている、一般に人間の示すいろいろな個性というものが、生後10年ほどのあいだの体験によって大脳の中で作り上げられた、強固な神経結線の働きによるものであるという研究結果の一例としての個人差にすぎなく、文字そのものの性質ではないと考えられるようになってきている。したがって、漢字を用いるということには本質的な利点はなく、むしろそれが客観的、抽象的、合理的な思考に与える不利な点のほうが問題になってくると考えられる。

その上、漢字の読み書きを重視し、音声言語をないがしろにして来たこれまでの教育は、ただに日本人の国語能力のみならず、外国語の教育にまで好ましからぬ影響を与えている。すなわち、たとえば今日の世界においては事実上の共通語である英語の教育においても、平易で明瞭な英語の会話や文章作りの能力を身につけることをさせないで、アメリカの知識人でさえ理解に苦しむような、やたらとむずかしい文法構造や単語を含む文章の読解力のみに力を入れさせているという、今日の英語教育の実状は、とかくむずかしい漢字

を使いたがり、いまでは起源さえ一般の人たちによく知られていないような、ことさらむずかしい表現の文を貴しとする国語教育の思想を、そのまま英語教育に反映させているからに過ぎないのではないだろうか。

この問題はなにもいまに始まったものではなく、すでに古くから議論されてきたことであるが(たとえば　藤村　1940[26]参照)、国際社会においてはほとんど役に立たない、そうした英語教育の現状に業をにやした中央教育審議会(中教審)は、大学の入学試験から英語を外してもよい、というような勧告案を最近まとめたようである。

しかしながら、いよいよ国際化し、かつ情報の伝達がますます重要になる、これからの国際情報社会において、国の運命を左右する指導的地位につくべき者が、国際共通語の駆使能力も身につけないままで大学の教育を修められるような教育制度を認めるということは、本末転倒の処置であると筆者は考える。「敵は本能寺にあり」、まず改善されなければならないのは、今日の社会に合わなくなってきている、日本語の表記法のほうであろう(山田　1997[37])。

さらに、いまでさえ入学試験の科目が少なく、社会の中堅としての大学卒業者の知識が片寄っていて、たとえば文科系の出身者に理工系の基礎知識が欠けているという現状である。いま以上に入学試験の科目を減らすのはどんなものであろうか。むしろ改められるべきなのは、試験結果から、個人の教育適性を判断する評価法のほうであると考えるべきであろう(山田1994[36]参照)。

### 19　縦書き・横書きの問題と文書コストの問題

日本における文科系と理科系間の断絶の度合いを強めているものの一つに、縦書きと横書きの問題も考えられる。

かなは縦書きの文化の中で発達してきた文字であるから、むかしのままの字体では、手で書くのになかなかきれいな横書き文字にはなりにくいことは事実である。しかしワープロの普及した今日では、その点は大きな問題ではなくなってきている。

いっぽう読むという観点からすると、かつて筆者が詳しく調べ上げてあるように(山田　1987[34])、縦書きと横書きとのあいだに、認知心理学的には本質的な読みやすさの優劣はなく、個人がそれを感じるとするのは、先に述べたように、幼いときの経験によって作り上げられた、人間に内在の諸現象の、もう一つの現れにすぎなく、やはり文字そのものの性質によるものではない。

にもかかわらず、縦書きで育ってきた文化人たちの中には、横書きは読みづらいと、あたかもそれが文字の本質的な性質のように考え、横書きの読みに慣れることを拒否する者さえある。またジャーナリズムやメディア一般も、世にさきがけて良識と理念とを形作っていくべき立ち場にありながら、同じ内容の本を縦書きと横書きとで出版するのでは、売れ行きに数倍の差が出るという資本の論理で、そうした世情に迎合しすぎていないだろうか。

あたかもその態度は、かつて1930年代に軍部の独裁政治が台頭しつつあったときに、正面きってそれを批判せず、むしろそのお先棒かつぎの役を果たしたこと(宮野　1982[33])を彷彿させると思えないだろうか。

公文書に関しても、1949年4月に内閣から各省の大臣にあてた通達(閣甲第104号)によって、公文書の横書き化の実施が命令されている。それから約50年、いまでは公文書の横書きはほぼ定着したが、例外として、たとえば法の遵守にもっとも忠実でなければならない法務省管轄下の法令が、従来からの法律によって縦書きが定められているという口実のもとに、いまも依然として縦書きのままにしてあり、これを横書きに改めるに必要な法の改正は、半世紀もほったらかしにされたままになっている。そのほかの例外としては官報の縦書きがある。

こうした無言の共同謀議とも言える態度が官民の一部にあることによって、結果としてどういうことが起こっているかというと、文科系の知識人が、横書きでないと不便きわまりない数理的な取り扱いを含む横書きの文書をまじめに読まなくなり、したがってそうした文書が、かれらに理解が可能な形で書かれなくなるということが起こり、それがまた、ますます横書きの文書が読まれなくなるように働くという悪循環が断ち切れず、結果として日本文化の中で文科系と理科系との人びとのあいだの解離を助長していることは、ほぼ間違いないであろう。

いくつかの銀行などで行なわれた総会屋に対する不正利益提供行為は、かなりまえからのことだったにもかかわらず、1997年になってやっと表面化したのであるが、それを反省した結果としてある銀行では、今後役員会は詳細な議事録を作成保存することを決めたことを、6月17日になって発表した。これは、いつの役員会でいかなることが決定されたかがあとになっても

明確にできるようにするためであるという。

しかしながら、速記とタイプライタの導入によって、効率よくしかも経済性をもって議事録が作成できるようになったアメリカの企業一般では、もう100年ほどもまえからそうした議事の詳細の文書化は実行に移され、それによって経営が客観化し、明朗化している。もし日本でもそうした文書化が徹底していたならば、今回のような不祥事はかなり起こりにくかったことであろう。遅まきながら日本でも、ようやく企業の経営が少数の役員による密室内の決定に近い形から解放されて、客観的で合理的な開かれた経営体制へと動き出したと言うべきであろう。

とは言っても、世界一複雑で非能率的な表記法を用いている日本の文書では、そうした議事録の作成にはアメリカに比べて数倍の費用と、それに近い余分の時間がかかっているのが現状である。いまのままの表記法で、はたしてアメリカにおけるような詳しく正確な議事録が常に即時に作成できるものだろうか。もしそれでもそれを作成し続けようとするならば、それは企業全体としてのオフィス関係の経費の増加につながるであろうし、さらに一歩進めて、アメリカではふつうに行なわれているように、単に役員会ばかりでなく、社内の全ての会合の詳しい議事録をも整備するということになると、経営費に占めるオフィスのコストがかなり上昇し、国際的な経済競争において、重い負担となりかねないであろう。かつては日本軍の戦力への負担となり、いまとなっても国際的な経済競争の足を引っ張っている表記法の問題に、われわれはもっと考慮を払うべきではなかろうか。

## 20　これからの課題

本稿においては、戦闘の勝敗を決する第1の要因が数量的優勢さにあることの一端を、もっとも基本的な初歩の数量的分析によって示した。もちろん実際の戦闘には数量的な格差のほかに、勝敗を決する数多くの要因が複雑に絡み合っていることは言うに及ばない。しかしここに述べたことは、それら他の要因を論じるまえに、まず確かめておかなければならない基本である。

本稿で取り上げた実例は、主として大東亜における作戦だけであった。しかしながら、第2次世界大戦全般における物量の勝利をつとに主張したエンソー(1956)[5]の分析は、いいかえれば、ここで述べた対敵兵力の数量的優位差の勝利の記述にほかならない。

そうした分析のつぎに、なぜわれわれが過去においてそのような戦略や戦術の検討と理解から目をそむけるような非合理的な態度をとり続けるようなことが起こったかの原因の一端として、われわれが用いている、ことばの表記法の及ぼしているであろう影響について、ごく簡単に述べてみた。

現在、文字についてそうした事実が科学的に解明されだしているにもかかわらず、一部の論客やジャーナリズムがことさらそれに目をつむり、科学的に根拠のない、主観による思い込みにすぎない従来の文字論にいまだに執着しているとしか思えない態度をとり続けているのは、やはりかつての精神主義の枠をそのまま引きずっているかのように見える。そのようなことでは、前節にちょっと触れたように、ひとたび社会の大変動が再来すれば、またもやかれらが非合理主義的な思潮のお先棒をかつぐの恐れなしとしないであろう。

敗戦直後の1945年9月27日に天皇が連合軍最高司令官マッカーサー元帥を訪問したときに、天皇は「平和を欲していたが、開戦を止められなかった。それは国民が戦争を望んでいたからである。もし自分が反対したら、彼らは自分を退任させ、他の天皇を立てたであろう」と話したと、当時のアメリカ大衆の大雑誌「ライフ」の特派員が報道し、アメリカの国民を沸かせた。マッカーサー司令部はそれに対して何の反論もしなかったので、これは真実だとされたが、果たしていまの日本のジャーナリズムについて、再び日本国民をそうした世論にまで引きずって行くことはないとの確信が持てるであろうか。

ともあれ、いにしえの剣聖の言に「型に入りて型より出でざるはくらく、型に入らざるはすなわちあやふし」というのがあるという。本稿に述べた戦略および戦術の分析は、その「型」のもっとも基本的なものの一つであると考えてよい。にもかかわらず、こうした観点から実際の戦闘の経過を数理的に分析し、さらにその結果と現実のデータとの相関をとって述べている文献は見当たらないようである。

今回本稿をまとめるにあたり、井上成美海軍大将について少し調べていたら、宮野(1982)[33]の第2章に、本稿の式(2)で$r=1$のときの簡単な応用の例が記述されているのがみつかった。しかしながら、その記述の前後を含めて注意深く読んでみると、宮野自身は理論の理解なしにそれを書いているようである。しかも宮野の記述に従うならば、かつての海軍大学校の学生たちも、井上の設問に応えてその理論を導き出すことが

できなかったとされている。

さらに宮野には、百発百中の砲1門と百発一中の砲百門との比較についての引用もあるが、その記述は、やはり宮野が本稿で述べたような理由を理解した上で書いているようにはみえない文章である。

本稿で述べた数理的な議論は、理系の高等教育を受けたものなら問題なくついて行ける程度のものであり、また高等教育を受けたものなら、たとい文系であってもこのくらいはついて行けるべきであると思えるほどやさしいものである。しかるに日本の現状においては、それとはかなり隔りがある文系理系間の2極化が存在する。その上、本稿では簡単にしか触れられなかったが、表記法の複雑さのゆえに生じている事務の能力の悪さと不透明さ、かてて加えて資源の分散逐次投入の変形である、お役所仕事のタテ割り構造が原因となっている非能率さなどが依然として存在していて、これからの社会においてますます中心的役割りを担っていく情報活動全般を動きの鈍いものとしている。おまけに、すでに石油資源問題において明らかになったように、資源の豊かな国ぐにの工業化に伴って、21世紀の国際政治においては資源国粋主義とも呼ぶべき経済思潮がいっそう力をふるうようになることはまぬがれないであろう。

そうした停滞した現状や未来の動向がいま明らかになりつつあるがゆえに、これからの国際社会における日本の将来に対して、筆者はかなりの不安を覚えてしまうのである。

## 21　おわりに

19世紀なかば、「戦争と平和」などの名作を書いた、世界的文豪レフ・トルストイは、歴史についても独特の哲学的見解を持つにいたったという。すなわち「歴史上の一切の大事業は、すべて眼に見えぬ捕捉しがたい大きな「力」によって左右されるのだ。神を信ずる者はこれを原理と呼ぶし、信じない人々は歴史上の法則と断定するけれども、とにかく、一切の権力も、かかる権力を有すると信ずる人々も、この捕捉しがたい大きな「力」の、盲目の機械にほかならない」（原久一郎・訳）と述べている。

そうした考えは、大筋において正しいと筆者も思う。しかし、この諦観とも言うべきもののかなり大きな部分は、現象の裏に働いている数理的な原理を理解し、それがかれの言う「力」の重要な構造であると知ることができれば、かなり薄れる性質のものではなかろうか。ましてや本稿で扱ったことがらなどは、歴史上の大事業などと言うよりは、むしろ大事業を構成していた下部構造にすぎず、したがって数理的扱いに、よりよくなじむ性格のものであろう。

そして、これら下部構造を統制している数理的な原理を理解することは、とりもなおさず、たとえば大東亜戦争のような上部構造を動かしていた「力」の一部を、よりよく理解することになるのであるから、こうした初歩的な数理的原理を自から用いて、ものごとを考察する能力を持たなかった者に国の指導をゆだねたことは、まことに危険なことであったということを、われわれは決して忘れてはならないと思う。

ちかごろは学問における西欧的な分析的手法が行きづまり、これからは東洋古来の多面的な合成的手法の時代であるといったような主張が、一部の文化人のあいだで人気を博する傾向がみられる。筆者はその主張自体には反対する者ではない。むしろ総合的な視点を持った、幅の広い文化人の増えることには大賛成である。しかし、ここに述べてきたような、かなり初歩的な数理的分析能力も身につけないまま、いたずらに合成的見地のみを主張することは、やはり「あやうい」道であると考える。

筆者の学問的関心は、広くは情報科学であり、現在特に関心があるのは人間の能力の認知科学的評価法である。本稿で述べてきた理論らしいものは、筆者は特に勉強したことはないが、おそらくオペレーションズ・リサーチ（OR）と呼ばれる分野に属するものの、初歩のまた初歩ではないかと考えていたところ、はたして黛（1998）[32]によれば、第1次大戦後イギリスのフレデリック・ランチェスターという、航空機および内燃機関の技術者によって類似の分析がなされており、1980年代の日本では「ランチェスター戦略」という名称で、特定市場への資源の集中を提唱した本がマーケティング関係者に大いに売れ、当時すでにランチェスターの理論がほとんど忘れられていた欧米に、黛などによって逆に紹介されたことがあったと言う。

専門分野が異なるので、筆者はこのランチェスター戦略のことは全く知らなかったが、当時かなり有名になったもののようであるから、その後同じ理論が誰かによって第2次大戦の戦闘の分析にも用いられたという可能性は、充分考えられると思う。

アメリカのプリンストン大学における研究で、数学のノーベル賞と言われるフィールズ賞を受け、1976年に東京大学の教授となり、のちには理学部長をも務め

た数学者小平邦彦博士は、無口な学者肌であった。あるとき先生は、「学者は専門バカだ、専門バカだ、と言うが、学者に専門が無かったら、ただのバカだ」という発言をなさって、評判になったことがある。その意味するところについては「学者は自分の専門においては、バカと言われることのないよう、的確な判断をしなければならない」ということのほかに、「学者は専門のことしか知らない専門バカでもかまわない」のだという解釈もあった。

しかし筆者の理解としては、先生の発言の主旨は、とかく日本では何かことがあると、ジャーナリズムやメディアが、それとはあまり関係のない分野で功績のあった学者を引っぱり出してきて、あれこれと解説させたり意見を求めたりする風潮があり、それを心良く思われなかった教授が、「学者はとかく視野が狭いから、自分の専門分野での発言にはことさら注意する必要がある。もし他の分野のことが自分の分野での判断に少なからぬ影響があるというときに、それを忘れ、自分の分野だけを孤立させ、その中だけでものごとを判断した結果が正しくなかったりすると、それこそ物笑いの種になるよ」という、いましめのことばとしてのものではなかったかと思っている。

本稿を準備するにあたって筆者は、この小平教授のことばを思い出しては、絶えず何か間違ったことを書いていないかに気を配ったつもりである。それでもここには、筆者のいままでの研究分野から踏み出した考察が少なからずあるので、思わぬ錯誤を犯した可能性が考えられる。諸賢がこれをお読みになられるにあたっては、書かれてあることをそのまま鵜呑みにされず、いつも自分でお考えになって、内容を判断なさっていただきたいと思う。そして、もし何かおかしいと思われることが書いてあったら、率直なお便りをいただければ幸いである。

**謝辞**

1944年から45年にかけてのフィリピン戦線において、数か月にわたる死闘に加わった日本陸軍のある将校のかた、および1941年からのシンガポール攻略作戦をはじめ、中国を含むアジアの広範囲の地域において戦闘に参加された、日本陸軍航空隊のある将校のかたは、本稿の草稿をお読み下さり、いろいろと貴重なご意見をお寄せ下さった。また、もと日本海軍の、ある将校のかたのご意見を伺うこともできた。さらに本稿を査読された山口東京理科大学の清水忠雄教授からも、いろいろと建設的なご批評をいただいた。それらご意見の多くはここに取り入れさせていただいたつもりであるが、しかし本稿中にいかなる不備や不適切な判断があったとしても、それはすべて筆者の責任である。

その後、黛治利氏は詳しいお考えを述べた長文を寄せられた。その中で氏は、第2次大戦以降の戦闘においては、通信と情報処理とが戦闘の結果を大幅に左右したこと、したがって量よりも質が重要になったことを、多くの具体例を挙げて述べられておられる。そのお考えには大いに傾聴すべきものがあると筆者も考えるが、主としてそれは今後の問題についてであり、紙面のつごうで、詳細は省略させていただくことにした。

なおほかには、本稿における戦闘力の分析と表記法の分析とがかなりかけ離れたテーマなので、あるいは別べつにまとめたほうがよいのではないかというご意見を寄せられたかたがある。しかし、黛氏からのご指摘にも示唆されていたことであるが、戦闘における情報の伝達の重要さにおいて、この二つは実はかなり密接に関わっているのだという事実が、いまに至るまで一般にはよく理解されているとは思えないので、筆者はこの二つをあえていっしょに論じることは重要であると考える次第である。

最後に、本稿を書くにあたって、たびたび浄書タイプをして下さったのは木稲宏実さんである。

これらのかたがたに対して、ここに厚い感謝の意を表わしておきたい。

**参考文献**

[1] 阿川弘之，山本五十六，新潮社，1965年.[1969年に新版がでたが，この旧版のほうが，より写実的と思う]

[2] 阿川弘之，井上成美，新潮社，1986年.

[3] NHK取材班，エレクトロニクスが戦いを制す，ドキュメント太平洋戦争，角川書店，1993年.

[4] NHK，長篠の戦い——鉄砲伝説の真実，「堂々日本史」シリーズ，1997年11月25日放映.

[5] エンソー，R.C.K.(内山正熊・訳)，第2次世界大戦史，岩波新書・青232，1956年.
[Robert Charles Kirkwood Ensor, A Miniature History of the War, Down to the End of the War in Europe, 1945]

[6] 生出寿(おいで・ひさし)，凡将山本五十六，現代史出版会，1983年.[徳間文庫版，1986年]

[7] 生出寿，砲術艦長・黛治夫，光人社，1988年.

[8] 岡田益吉，日本陸軍英傑伝，光人社，1972年．
[9] 奥宮正武，大艦巨砲主義の盛衰，朝日ソノラマ，1989年．
[10] 小倉磬夫，古老の思いこみ(2)，朝日カメラ，1997年1月号，pp. 140-141．
[11] 神原周(しゅう)，新しき科学技術のために，河出書房，1946年．
[12] 日下公人(きみんど)，名誉ある孤立の研究，PHP研究所，1993年，p. 39．
[13] クロパトキン，アレクセイ・ニコラエヴィッチ(Aleksei Nikolaevich Kuropatkin)，ロシア軍隊と日露戦争，1902．
[14] 児島襄(のぼる)，誤算の論理——戦史に学ぶ失敗の構造，文芸春秋社，1987年．
[15] 杉森久英(ひさひで)，辻政信，文芸春秋新社，1963年．[河出文庫版，参謀・辻政信，1982年]
[16] 高木俊朗(としろう)，抗命，インパールII，文春文庫151-2，1976年．
[17] 高橋昭男，技術系の文書作法，共立出版，1995年．
[18] 田中光子，文字は認知処理を支配するか一漢字とアルファベット一，関西学院大学文学研究科心理学専攻修士論文，1997年3月．
[19] 千早正隆，日本海軍の戦略発想，プレジデント社，1982年．
[20] 手塚晃，言語・思考の枠組としての文字システムの評価，日本語学，第6巻，第8号，pp. 88-105，1987年．
[21] 戸部良一，寺本義也，鎌田伸一，杉之尾孝生，村井友秀，野中郁次郎，失敗の本質一日本軍の組織論的研究，ダイヤモンド社，1984年．
[22] 内藤初穂，海軍技術戦記，図書出版社，1976年．
[23] 中谷宇吉郎，千里眼其の他，附記，春草雑記，pp. 30-39，生活社，1947年．
[24] 中谷宇吉郎，針葉油雑感，科学世界，1946年2月号．[上掲書，pp. 215-236]
[25] 新見政一・他，日本海軍の良識一提督新見政一，自伝と追想，原書房，1995年．
[26] 藤村作(つくる)，国語問題と英語科問題，白水社，1940年．
[27] 福井静夫，日本の軍艦——わが造船技術の発達と艦艇の変遷，出版協同社，1956年．
[28] 保科孝一，国語問題五十年，三養書房，1949年．
[29] Horodeck, Richard A., The Role of Sound in Reading and Writing Kanji, Ph.D. dissertation, Cornell University, 1987.
[30] 黛治夫，海軍砲戦史談，原書房，1972年．
[31] 黛治夫，艦砲射撃の歴史，原書房，1977年．
[32] 黛治利(はるとし)，私信，1998年1月25日．
[33] 宮野澄(とおる)，最後の海軍大将・井上成美，文芸春秋社，1982年．[文春文庫版，1985年]
[34] 山田尚勇(ひさお)，横書きの歴史・現状と評価，文学(岩波)，第55巻，第6号，pp. 25-44，1987年．
[35] 山田尚勇，文字論の科学的検討，学術情報センター紀要，第4号，pp. 261-318，1991年．
[36] 山田尚勇，創造性の発露について(論文集)，学術情報センター，1994年，vi+143 pp.
[37] 山田尚勇，情報化社会の国際化と日本語，学術情報センター紀要，第9号，pp. 33-71，1997年．

## 付録：兵力の逐次投入の効果を示す一般式

ここでは中学教育レベルの算術的演算のみを用いて、本文の第11節で取り扱った、兵力の逐次投入の問題をもっと一般的な形で考える。初めに敵と味方の戦闘単位数(以後員数と呼ぶ)をそれぞれ、$e_1$および$f_1=e_1/s$とする。$s$は味方と比べた敵の員数的優勢度である。また技術力などを含めて、味方の対敵練度を$r$とする。

### (1) 味方の員数を$n$分の1ずつに分けて逐次投入することによる損失

ここで2回目以降$(n-1)$回目までは、残存員数が投入員数の$k$倍$(0\leqq k\leqq 1)$になったときに、次回の投入を行なうとする。[$k=1$のときは、明らかに初戦における全員数の一括投入を意味し、$k=0$のときは投入員数が全滅してから次の投入を行なうことを意味する。]また最終投入後は全滅するまで戦うものとする。

いま味方の第$n$回目の投入時の員数を$g_n$で表わすとすると、初戦における敵・味方の員数はそれぞれ$e_1$および$g_1$であり

$$g_1=f_1/n=e_1/sn \tag{a}$$

である。第1回戦の結果における敵の残存員数 $e_2$、すなわち第2回戦初頭における敵の員数は、本文の第3節で得た式(2)を用いて、

$$e_1^2-e_2^2=r(g_1^2-k^2\cdot g^2)=(r/s^2n^2)\cdot e_1^2\cdot(1-k^2)$$

から $e_2^2$が求まる。すなわち

$$e_2^2=e_1^2-(r\cdot e_1^2/s^2n^2)(1-k^2) \tag{b}$$

つぎに第2回戦初頭における味方の員数は

$$g_2=k\cdot g_1+f_1/n=(ke_1/sn)+(e_1/sn)=(e_1/sn)(1+k) \tag{c}$$

したがって、再び式(2)を用い

$$e_2^2-e_3^2=r(g_2^2-k^2\cdot g_2^2)=(r\cdot e_1^2/s^2n^2)(1-k^2)(1+k)^2$$

が得られる。ゆえに

$$e_3^2=e_1^2-(r\cdot e_1^2/s^2n^2)(1-k^2)[1+(1+k)^2] \tag{d}$$

同様にして第3回戦では

$$g_3=k\cdot g_2+f_1/n=(e_1/sn)(1+k)k+(e_1/sn)=(e_1/sn)(1+k+k^2) \tag{e}$$

$$e_3^2-e_4^2=r(g_3^2-k^2\cdot g_3^2)=(r\cdot e_1^2/s^2n^2)(1+k+k^2)^2(1-k^2)$$

が求まる。ゆえに

$$e_4^2=e_1^2-(r\cdot e_1^2/s^2n^2)(1-k^2)[1+(1+k)^2+(1+k+k^2)^2] \tag{f}$$

以上の結果からの見通しを踏まえ、ここで数学的帰納法を用いて、一般式を出すことにする。すなわち、上の式(a)、(c)、(e)の形から、$g_m$および $e_m$をそれぞれ

$$g_m=(e_1/sn)\cdot\sum_{i=0}^{m-1}k^i,\quad(m\geqq1) \tag{g}$$

および

$$e_m^2=e_1^2-(r\cdot e_1^2/s^2n^2)(1-k^2)\cdot\sum_{j=0}^{m-2}\left(\sum_{i=0}^{j}k^i\right)^2,\quad(m\geqq2) \tag{h}$$

と仮定してみると、$g_{m+1}$の定義により

$$g_{m+1}=k\cdot g_m+(e_1/sn)=(ke_1/sn)\cdot\sum_{i=0}^{m-1}k^i+(e_1/sn)$$

$$=(e_1/sn)\cdot\left(\sum_{i=1}^{m}k^i+1\right)=(e_1/sn)\cdot\sum_{i=0}^{m}k^i \tag{i}$$

また式(2)から

$$e_m^2-e_{m+1}^2=r(g_m^2-k^2\cdot g_m^2)=r\cdot g_m^2(1-k^2)$$

$$=(r\cdot e_1^2/s^2n^2)\cdot\left(\sum_{i=0}^{m-1}k^i\right)^2\cdot(1-k^2) \tag{j}$$

すなわち

$$e_{m+1}^2=e_m^2-r\cdot g_m^2\cdot(1-k^2)$$

$$=e_1^2-(r\cdot e_1^2/s^2n^2)(1-k^2)\left[\sum_{j=0}^{m-2}\left(\sum_{i=0}^{j}k^i\right)^2+\left(\sum_{i=0}^{m-1}k^i\right)^2\right]$$

$$=e_1^2-(r\cdot e_1^2/s^2n^2)(1-k^2)\cdot\sum_{j=0}^{m-1}\left(\sum_{i=0}^{j}k^i\right)^2 \tag{k}$$

したがって式(i)、(k)はそれぞれ式(g)、(h)で添字の値を1だけ進めたかたちであり、仮定した式(g)および(h)が一般の場合に正しいことがわかる。

さて、最後に $m=n$ となったときには、残った全員数を投入して最後まで戦うのだから、

$$g_n=(e_1/sn)\cdot\sum_{i=0}^{n-1}k^i \tag{l}$$

を用いて、最終的に敵の残存員数 $e_{n+1}$ としては、式(2)から

$$e_n^2-e_{n+1}^2=r\cdot g_n^2 \qquad (m)$$

すなわち

$$e_{n-1}^2=e_n^2-r\cdot g_n^2$$

式(g)と式(h)とにおいて、それぞれ $m=n$ とし、この右辺に代入すると

$$e_{n-1}^2=e_1^2-(r\cdot e_1^2/s^2n^2)\left[(1-k^2)\cdot\sum_{j=0}^{n-2}\left(\sum_{i=0}^{j}k^i\right)^2+\left(\sum_{i=0}^{n-1}k^i\right)^2\right] \qquad (o)$$

この式から、味方が全滅したときの敵の残存兵力数 $e_{n+1}$ が求められる。そして、初頭での数 $e_1$ に対するそのときの敵の残存率 $e_{n+1}/e_1$ の一般式は、自乗の形として表わすと

$$(e_{n+1}/e_1)^2=1-(r/s^2n^2)\left[(1-k^2)\cdot\sum_{j=0}^{n-2}\left(\sum_{i=0}^{j}k^i\right)^2+\left(\sum_{i=0}^{n-1}k^i\right)^2\right] \qquad (p)$$

となる。（この右辺は等比級数の和を含むから、それを与える、下の(II)でみられるような、よく知られた公式を用いれば、もう少し簡単にできるが、当面の問題には、このままでも充分であろう。）

この式の右辺には戦闘初頭の敵方の総数 $e_1$ も味方の総数 $f_1$ も出ていないから、明らかにこれは $e_1$ および $f_1$ には関係なく、味方に対する敵の員数的優勢度 $s=e_1/f_1$ と $k$、$n$、$r$ だけによって決まる数値となる。

いまこの式(p)を、本文の第11節において計算した具体例に対して用い、兵力の逐次投入の効果を検証してみると、$r=s=1$、$k=1/2$、$n=3$ であったから、

$$\begin{aligned}(e_4/e_1)^2&=1-(1/3^2)\left[(1-(1/2)^2)\cdot\sum_{j=0}^{1}\left(\sum_{i=0}^{j}(1/2)^i\right)^2+\left(\sum_{i=0}^{2}(1/2)^i\right)^2\right]\\&=1-(1/9)\left[(3/4)(1+(1+(1/2))^2+(1+(1/2)+(1/4))^2\right]\\&=0.3889\end{aligned}$$

したがって残存率は

$$e_{n-1}/e_1=0.6236$$

となり、当然のことながら、本文で直接得た残存敵兵力の計算値、62パーセント強と一致する。

なお、断わるまでもないことであろうが、上の基本式(g)、(h)、(p)においては、$k$、$n$、$r$、$s$ などの値の選びかたによっては、第 $n$ 回の会戦に至るまでのある段階 $m$ で敵が全滅することも起こるから、そのときはそれ以後の $m$ の指定は無意味となる。

### (II)　逐次投入による損失を補うに必要な兵力の増強量

いまある戦局において、すでに展開しているある員数の敵の兵力と戦う戦局において、味方に必要とされる兵器の生産あるいは兵員の補給が間に合わないために、兵力が一気に整えられないということが起こり、何回かに分散して逐次に投入しなければならないというときには、敵に対して互格に戦うために必要となる味方の兵力の総和は、結果として敵の兵力よりも多くなってしまう。したがって以下では、そうした兵力の逐次投入を余儀なくされたときに、味方に必要となってくる、投入兵力の増加量を求めてみることにする。

いま敵方の展開している兵力数を $e_1$ とする。それに対し、味方は兵力が足りず、とりあえず $m(m<e_1)$ しか兵力を揃えられないとする。また味方の対敵総合練度を、いままでの場合いと同じようにように $r(r>0)$ とする。

計算が簡単になるので、ここでも(I)の例と同じように、戦闘中の味方の兵力が投入段階時の総数の $k$ 倍($0\leqq k\leqq 1$)にまで失われたときに、新しく $m$ を追加投入する場合いを考えるものとする。いっぽう敵方には、追加投入はいっさいないものとする。

すると味方が $n$ 回目($n\geqq 1$)に兵力を投入した時点では、味方の兵力員数 $f_n$ は、

$$f_1=m,\quad\left(=m\cdot\sum_{i=0}^{0}k^i\right) \qquad (q)$$

$$f_2=mk+m=m(1+k),\quad\left(=m\cdot\sum_{i=0}^{1}k^i\right) \qquad (r)$$

軍の戦闘能力を左右したのは何か

$$f_3=m(1+k)\cdot k+m=m(1+k+k^2)=m\cdot\sum_{i=0}^{2}k^i \tag{s}$$

となり、明らかに

$$f_n=m\cdot\sum_{i=0}^{n-1}k^i \tag{t}$$

あとの便宜のために、ここで

$$S_{n-1}=1+k+k^2+\cdots\cdots+k^{n-1}=\sum_{i=0}^{n-1}k^i \tag{u}$$

の値を求めておくと、中学の数学でよく知られているようにして

$$kS_{n-1}-S_{n-1}=(k-1)S_{n-1}=\sum_{i=1}^{n}k^i-\sum_{i=0}^{n-1}k^i=k^n-1 \tag{v}$$

したがって

$$S_{n-1}=\sum_{i=0}^{n-1}k^i=(1-k^n)/(1-k) \tag{w}$$

と、よく知られている式が求まる。

したがって式(t)から

$$f_n=m(1-k^n)/(1-k) \tag{x}$$

よって $n$ 回目の味方兵力の投入時の敵の員数は、本文で求めた式(2)と以上の式とから、一般に

$$e_1^2-e_2^2=rm^2(1-k^2)\left(\sum_{i=0}^{0}k^i\right)^2=rm^2(1-k^2)(1-k)^2/(1-k)^2 \tag{y}$$

$$e_2^2-e_3^2=rm^2(1-k^2)\left(\sum_{i=0}^{1}k^i\right)^2=rm^2(1-k^2)(1-k^2)^2/(1-k)^2 \tag{z}$$

・
・
・

$$e_{n-1}^2-e_n^2=rm^2(1-k^2)\left(\sum_{i=0}^{n-2}k^i\right)^2=rm^2(1-k^2)(1-k^{n-1})^2/(1-k)^2 \tag{aa}$$

いまもしこのようにした $n$ 回の逐次兵力投入によって、味方は初めに $e_1$ の員数だった敵とはじめて互格に戦い得るのだということは、その後の戦闘で、敵・味方ともに全滅することを意味するから、そのときは

$$e_{n+1}=0 \tag{ab}$$

となり、かつ式(t)と(w)から

$$f_{n+1}=m\cdot\sum_{i=0}^{n}k^i=S_n=0 \tag{ac}$$

だから、式(2)、(w)から

$$e_n^2=r\cdot f_n^2=rm^2(1-k^n)^2/(1-k)^2 \tag{ad}$$

したがって、この式(ad)から始めて、式(aa)、…、(z)、(y)と逆にたどることにより、

$$e_{n-1}^2=rm^2[(1-k^2)(1-k^{n-1})^2+(1-k^n)^2]/(1-k)^2 \tag{ae}$$

$$e_{n-2}^2=rm^2[(1-k^2)(1-k^{n-2})^2+(1-k^2)(1-k^{n-1})^2+(1-k^n)^2]/(1-k)^2 \tag{af}$$

式(ad)、(ae)、(af)の形をみることにより、

$$e_{n-p}^2=rm^2[(1-k^2)\cdot\sum_{i=n-p}^{n-1}(1-k^i)^2+(1-k^n)^2]/(1-k)^2 \tag{ag}$$

と置いてみると、式(2)により

$$e_{n-p-1}^2=rm^2[(1-k^2)\cdot\sum_{i=n-p-1}^{n-1}(1-k^i)^2+(1-k^n)^2]/(1-k)^2 \tag{ah}$$

となることが分かるから、数学的帰納法により、式(ag)は$0\leqq p\leqq n-1$について成立することが分かる。ここで、$i$

を指標とする何か一般の式 $F(i)$ について、$\sum_{i=p}^{q} F(i)$、$q<p$、のときには集める項が無いことを意味するから、

$$\sum_{i=p}^{q} F(i)=0, \qquad q<p \tag{ai}$$

である。したがって式(ad)、(ae)、(af)、(ag)などは、まとめて

$$e_{r-p}{}^2=rm^2[(1-k^2)\cdot\sum_{i=n-p}^{n-1}(1-k^i)^2+(1-k^n)^2]/(1-k)^2,\ (0\leqq p\leqq n-1) \tag{aj}$$

と書けることが分かる。それで

$$e_1{}^2=rm^2[(1-k^2)\cdot\sum_{i=1}^{n-1}(1-k^i)^2+(1-k^n)^2]/(1-k)^2 \tag{ak}$$

ここで敵は初頭に $e=e_1$ を投入するだけであるが、味方は員数 $m$ ずつを $n$ 回投入しているから、$f=nm$ である。したがって $n$ 回に分けて分散逐次投入される味方が、員数 $e$ の敵と互格に戦えるのに必要となる兵力 $f$ と、敵兵力 $e$ との比は、自乗の形で示すと

$$f^2/e^2=(1-k)^2n^2/r[(1-k^2)\cdot\sum_{i=1}^{n-1}(1-k^i)^2+(1-k^n)^2] \tag{al}$$

また、もし分散逐次投入によるこの戦力の減少分を、技術力を含めた、味方の対敵練度 $r$ で補おうとすれば、$f^2/e^2=1$から

$$r=(1-k)^2n^2/[(1-k^2)\cdot\sum_{i=1}^{n-1}(1-k^i)^2+(1-k^n)^2] \tag{am}$$

の $r$ が必要となる。

いま具体例として、$r=1$、$k=1/2$、$n=3$ の場合いを考えると、式(al)を用い、

$$f^2/e^2=0.25\times9/[0.75(0.5^2+0.75^2)+0.875^2]$$
$$=2.25/1.375=1.636$$

ゆえに

$$f/e=1.279$$

すなわち、味方は約28パーセントも余分の総兵力数の投入を必要とすることになる。

もし兵力の増強投入を行なわず、その分を対敵練度で補うとするならば、式(3)から、味方の練度としては $r=1.636$ であることが要求される。当然ながら、式(am)は本文第3節で求めた式(4)の与えている結果と整合している。

大東亜戦争において日本は、兵器の生産能力が不充分であり、しかも兵站能力でも劣っていたので、作戦にあたって兵員や兵器の緊急の配備が間に合わないことが多く、ここに述べたような分散逐次投入となるのが常であった。にもかかわらず、最終的には総員数において敵を圧倒するだけの供給ができず、また、技術力を含めた対敵練度のほうもかんばしくなく、結局は敗北に敗北が重なることになった。

以上で述べた二つの場合いの検討では、計算を初歩的なものに留めるために、兵力の新規投入の時期を、前の投入後における全兵力が半減した時としたが、実戦ではむしろ投入から投入への時間の経過が等間隔となることのほうが実情に近いかと思われる。しかし、その場合いの計算は、時間が媒介変数として入った、もっと複雑なものとなるので、ここではこれ以上踏み込まないことにしよう。それでも兵力の分散逐次投入は、そうした場合いにおいても決して好ましいことにならないことは、これら二つの場合いの計算結果から推して、もう直観的にお分かりいただけることと思う。

理工系の背景をお持ちのかたがたで、本稿における数理的な取り扱いが初歩に過ぎると思われた向きには、練習問題として、この等間隔時間ごとの兵力の投入の場合いの計算を試みてみられるのも一興ではないであろうか。

学術情報センター紀要　第10号(1998)

研究論文

# OCR認識誤りを含む書誌情報の認識

# Analysis of Bibliography including OCR Misrecognition

東京大学工学系研究科　早川　公泉
Kimimoto HAYAKAWA
Graduate School of Engineering, University of Tokyo
学術情報センター　高須　淳宏
Atshuhiro TAKASU
National Center for Science Information Systems
学術情報センター　安達　淳
Jun ADACHI
National Center for Science Information Systems

**要旨**

学術論文誌を扱う電子図書館では、書誌情報にリンクを張り、関連する文献を相互に結びつけるのが望ましい。しかし文献情報を画像データから抽出する場合、OCRの認識誤りによって書誌項目の抽出が難しい。そこで本稿では、確率的パターン解析によって、OCRを通して得られたテキストデータから、参考文献の情報を認識するための手法を提案する。

提案する手法は、認識済みの書誌方法から項目の属性を推定し、確率文脈自由文法を用いてその結果を修正するものである。本稿では提案される手法が書誌情報の認識に有用であることを示す。

**ABSTRACT**

At digital libraries providing academic journals, documents which have relation each other should be linked. But,it is dificult to analyze bibliography owing to effect of OCR misrecognition. This paper presents a method for analysis of bibliography which obtained through OCR.

The objectives of this paper are to presume bibliographich attributes using yet decided bibliography and to correct this result by probablistic context-free grammars. The paper shows the efectiveness of the presented method for analysis of bibliography.

[キーワード]　書誌情報、OCR認識誤り、確率文脈自由文法、電子図書館
[Keywords]　Bibliography, OCR Misrecognition, Probablistic context-free grammars, Digital Library

## 1 はじめに

電子図書館において、既存の紙に印刷された出版物は、あらかじめ電子化を想定されて製作された文書とは異なる扱いを必要とする。既存の出版物においても、印刷前の情報は電子化されている場合は多いが、一般にこの電子情報を扱いやすい形で入手するのは困難である。そのためスキャナを用いて、印刷物から電子化を行ない、文書画像として扱うことになる。

電子図書館ならではの特徴として、検索や即時参照が挙げられるが、このためにはテキストデータが必要となる。文書画像からテキストを得る手段としてOCRがあるが、必ず何らかの認識誤りが含まれる。そのためOCR認識誤りに対応したテキスト処理が必要となる。

電子図書館における利便性の向上の一つとして、文書画像間に自動的にハイパーリンクを設けることが挙げられる。具体的なリンク例としては、学術論文の目次と記事、参考文献と該当文献などの間などがある。

本研究では、論文から参考文献へのリンクの自動生成を行なうために、OCRによって得られたテキストデータから、参考文献の書誌的事項の認識を目指している。

## 2　研究の背景

### 2.1　既存の文書の電子化

そもそも、文書の基本はテキストデータである。現在では文書の作成にコンピュータは欠かせないといってよく、きちんとした文書を仕上げる場合はまずテキストデータを入力することになる。あらかじめ電子化を想定しているならば、このテキストデータを活用すべきである。

しかし既存の出版物においてはこのテキストデータを得られるとは限らない。例え得られたとしても、印刷物と同じように画面上に表示させるのは困難である。無論AdobeのPDFのように、画面上と印刷時の形態を一致させるものもあるが、印刷時の形態を画面上に再現できるのではない。テキストデータを画面に表示させるためには、文書の著者が作成時に表示形態を定める必要があるのである。

その点、文書画像を用いた場合は、印刷物の形態をそのまま画面に再現することが可能である。スキャナを用いて、十分実用に耐え得る精度の画像を容易に得ることができる。そのため、電子図書館で既存の出版物を扱う場合、文書画像として扱うのが妥当である。画像データとして扱う場合、テキストデータに比べてデータ量が格段に多く、全文検索などの電子化ならではの特徴を生かすことが出来ない。検索などのサービスを提供するためにはOCRを用いて文書画像からテキストデータを生成して対応することになる。

文書の電子化に要する手間は掛からなければ掛からないほど望ましい。その分の手間をより多くの文書の電子化に割けるためである。OCRによって得られたテキストデータに含まれる誤りを人手によって修正するのは大変手間が掛かる。そのため、多少の誤りが含まれているテキストでも、そのまま処理が行なえる方が良い。このテキストデータは電子図書館の利用者の目に触れない裏側の処理で用いれば良いため、誤りを含んだままでも構わない。

### 2.2　参考文献項目の認識

論文における参考文献の各項目から、対象となる文献へのリンクを設けるためには、リンク先の情報を持つデータベースとの照合が必要となる。照合を行なうためには、OCRによって得られたテキストデータから、タイトル等の属性を認識しなければならない。

誤りを含むマッチングについては研究が行なわれている[1]が、参考文献項目の属性を認識する手法については本格的な研究が行なわれていない。

そこで、本研究では、OCRによって得られた参考文献の各項目から、タイトル等の属性の認識について追求していく。

## 3　参考文献項目の構造

### 3.1　参考文献項目の例

参考文献項目の記述は論文誌ごとに異なる。情報系の学術論文誌における典型的な例を例1に挙げる。

データベースとの照合を行なう場合、通常タイトルが用いられるが、上の例から判るように、必ずしもタイトル部分が特定出来るようなデリミタで区切られているとは限らない。実際の参考文献項目では、このような形の整ったものは少ない。タイトル部にデリミタとして用いられている記号が含まれている場合も多々あり、どの部分がタイトルを表しているか判別が難しい場合の方が多い。

特に判別が困難であるのが複数のタイトルを持つ場合である。論文集の中に含まれる論文や、翻訳本などではタイトルや著者の記述形式が一定ではない。例として例2に示す。このように、人によって異なる記述の揺れを吸収しなければ
ならない。

認識する必要があるのはタイトルだけではない。物理の分野においては例3のようにタイトルの記述がないものも見受けられる。

これはいささか極端な例ではあるが、参考文献の特定にタイトル以外の要素も必要とすることはよくある。例えば、著者によっては同じタイトルで複数の論文を発表することも少なくない。このような場合には、タイトル以外に、論文の発表された文献名や年月が必要となる。

### 3.2　参考文献項目の品詞

参考文献項目の属性を以下のように定義し、以下この属性を品詞と呼ぶことにする。

#### 3.2.1　タイトル

論文などの主題を表しており、論文誌によっては論

天野　要：代用電荷法に基づく2重連結領域等角写像の数値計算法，情報処理学会論文誌,Vol.29,No.10,pp.914-924(1988).

小坂哲夫,鷹見淳一,嵯峨山茂樹."話者混合逐次状態分割法による不特定話者音声認識と話者適応",信学論(A),J77-A,2,pp.103-111(1994-02).

國藤:発想支援システムの研究開発動向とその課題,人工知能学会誌,Vol.8,No.5,pp.552-559(1993).

例 1　情報系の参考文献項目の例

Cohen,P.R.,Feigenbaum,E.A.(編),田中幸吉,淵　一博(監訳):人工知能ハンドブック第3巻,pp.504-522,共立出版(1984).

Ravden,S.and Johnson,G.:Evaluating Usability of Human-Computer Interfaces, EllisHorwood Limited (1989). (東(監訳):ユーザ・インタフェイスの実践的評価法-チェックリストアプローチによる使いやすさの向上-,海文堂(1993).)

例 2　翻訳本の参考文献項目の例

Anderson P W 1958 Phys. Rev. 109 1492-505

例 3　物理分野の参考文献項目の例

文のタイトルと本のタイトルが別々に存在するものもある。データベースとの照合を行なう際に最も重要な品詞である。

### 3.2.2　サブタイトル

本のタイトル以外に、その本が属するシリーズ名や第何版であるかなど、データベースとの照合を行なう際に必要ではないが、タイトルに準ずるものをサブタイトルとする。

### 3.2.3　著者名、姓、名

通常は文献を記述した人物名であるが、時には学会などの団体の場合もある。著者名が日本語表記の場合、姓のみを記すか姓名共に示すか決まった規則はない。

著者名が英語表記の場合は、姓と名に分けて扱ったほうが良い。

例4の例では姓名の間にも著者名間と同じデリミタが用いられている。従って、デリミタを用いて機械的に各要素に分割しようとする場合には姓と名を別々の品詞として扱う。

### 3.2.4　雑誌名、研究報告名

論文の掲載されている書誌が雑誌の場合では、その雑誌名も参考文献項目に示される。研究報告名も同様であるが、雑誌よりも一般的でない書誌は全て研究報告として扱うことにする。これらの品詞でデータベースとのマッチングをおこなわなければならない際に、両者の区別が必要となる。

### 3.2.5　出版元

書籍では必ず、雑誌などでもその書誌の出版元が示されることがある。

### 3.2.6　地名

その文献の出版元や研究会が行なわれた場所などである。ここでは国名等も地名として扱うことにする。

### 3.2.7　巻、号、頁

巻、号は雑誌などに用いられ、頁はほとんどの書誌で用いられている。学会の全国大会などにおける文献では、巻と頁が一括りにされて示されることが多く、

Cameron, E. J., Griffetch, N., Lin, Y., Nilson, M. E., Schnure, W. K. and Velthuijsen, H. : A Feature-Interaction Benchmark for IN and Beyond, IEEE Communication Magazine, Vol.31, No.3, pp.64-69(1993).

例 4　著者名が英語表記の参考文献項目の例

この場合の品詞は便宜上巻として扱うことにする。

#### 3.2.8　発行年月

よほどの理由がない限り、参考文献項目には必ず発行年月が含まれる。翻訳本のような場合では、複数の発行年月が記述されることもある。発行年は通常西暦で記述されることが多いが、発行月は著者によって記述が異なり、英語表記やただの数字であったり、記述は一定でない。発行月は省略されることも多い。

### 3.3　OCR による認識誤り

実際に処理を行なわなければならないのは、OCR 処理によって得られたテキストデータに対してである。そのため OCR の起こす誤りについて把握しておく必要がある。以下は情報処理学会論文誌1995年1月号の参考文献項目に対して、学術情報センター研究開発部で用いられている東芝製 OCR で処理して得られた結果から分析したものである。処理した文字全体の認識率は99%程であったが、文字の種類によって誤りを起こしやすいものとそうでないものとがあった。

#### 3.3.1　OCR の認識できない文字

OCR が認識できない文字は全く別の文字に認識されてしまう。例としては、ö → 6 などが挙げられる。

#### 3.3.2　良く似ている文字

人間でも注意しないと間違えそうな文字は OCR でも間違える。二 → ニ や p → P などである。OCR のならではの間違え方もあり、10 → lo や o → () といったものが見受けられる。比較的よく起こる。

#### 3.3.3　スペースの挿入脱落

スペースは形がないため OCR ではよく間違える。特に本来スペースのあるべき単語間のスペースが脱落してしまう誤りは頻繁に起こる。

#### 3.3.4　それ以外の誤り

あまり似ていなくても、元の画像の品質が悪いと誤りは生じる。例としては、s → 白 などが挙げられる。よく起こる誤りではなく、あらかじめ予想するのも難しいケースが多い。

## 4　提案する手法

本研究で提案する手法では、同種の論文誌における既に品詞の付与されている項目を利用するという手法を取る。最初に人手で品詞を付与する手間は掛かるが、品詞毎の特徴を解析する必要がないため、対象となる論文誌を広げるのが容易である。また、データベースに依存しないため、未知の文献に対しても既知の文献と同様に扱うことができる。

全体の処理は、OCR によって得られたテキストデータをデリミタによって各要素に切りだし、切り出された各要素の品詞を特定するという流れで行なわれる。

切り出された各要素はまず文字列処理により品詞を推定し、さらにそれを文法を用いて補正する。品詞を特定には、あらかじめ品詞が付与されている既知の参考文献項目から抽出した品詞群及び文法を用いる。

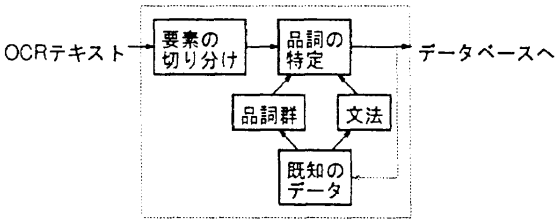



図 1　処理の全体像

## 5　要素の切りだし

まずデリミタを用いて、参考文献項目の各要素を分割する。論文誌によって要素間のデリミタは異なり、全ての論文誌に対応するのは困難である。幸い情報系の論文誌では参考文献項目の記述が似ており、まとめて扱うことも可能である。以後、情報処理学会論文誌を中心に、情報系の論文誌を対象として話を進めていく

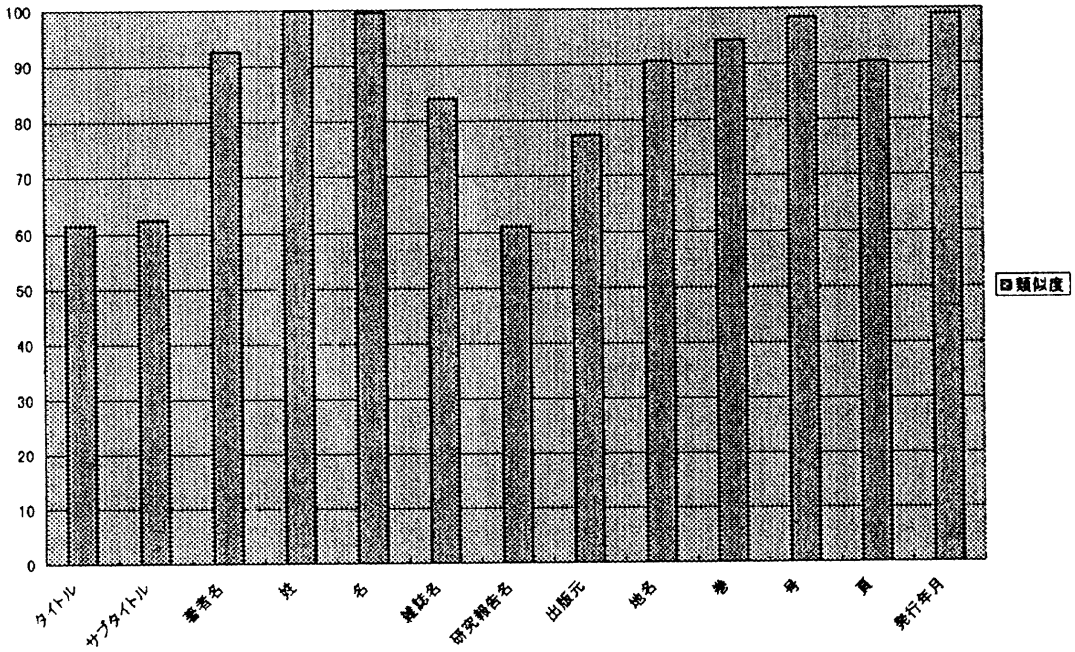



図 2　正解率

ことにする。

情報系の論文誌においては、',' が主に用いられ、部分的に ':' 、'(' や ')' が用いられている。幸いこれらの文字のOCR認識誤りは0.1%以下であり、さほどOCRの影響は受けない。むしろタイトルなどの品詞中にデリミタと同種の文字が含まれている場合のほうが問題となる。デリミタかそうでないかの区別は困難であるため、機械的にデリミタと定義した文字で区切ることにし、切り過ぎた要素は品詞の判定のさいに判別することにする。

また著者名が英語表記の場合、表4のように著者名間の最後の区切りが ',' ではなく 'and' になる。ここでは 'and' はデリミタとはみなさないことにする。最後の著者の名とその前の著者の姓がくっついてしまうが、この問題に関しては、その要素が著者名であることが認識できれば、その時点で'and'で区切ることによって解決できる。

## 6　文字列処理による品詞の判定

字句解析によって切り出された各要素の品詞は、まずその文字の並びから推定する。

参考文献項目の各要素は品詞ごとにある程度共通の特徴を持つ。例えば巻であれば 'Vol.' で始まったり、頁であれば 'pp.' で始まったりする。そこで、あらかじめ品詞の付与されている既知の要素との類似度を求め、その類似度によって要素の品詞を定める。

### 6.1　文字列間の類似度の定義

文字列間の類似度を定めるのに、最大一致文字列(Longest Common Subsequences)[2]を用いることにする。これは文字列間において、重複しない複数の部分文字列の総和を最大になるようにとったものである。単純な部分文字列の一致をとる場合と異なり、文字列にOCR認識誤りが含まれていてもそこを除くことが可能である。文字列間の最大一致文字列長をその文字列間の類似度と定義する。文字列長による正規化を行なわないのは、文字列長が長い場合に類似度が大きくなる影響よりも、正規化をおこなった場合に生じる、文字列長が短い場合に類似度が大きくなってしまう影響のほうが害が大きいためである。

### 6.2　品詞の推定

参考文献項目から切り出された要素は、既に正しい品詞が付与されている既知の品詞群の各要素とそれぞれ類似度を求め、最も類似度が大きかった要素の品詞を、求める要素の品詞とみなすことにする。最大の類似度を持つ要素が複数ある場合は、最も文字列長が近いものを優先する

### 6.3 情報処理学会論文誌1995年1月号の例

情報処理学会論文誌1995年1月号の参考文献項目の各要素に対しあらかじめ正しい品詞の付与をおこない、各要素毎に他の要素との類似度を求めて品詞の推定を行なった。ここで品詞毎の正解率とは、

$$\frac{\text{正しく認識された要素数}}{\text{その品詞と判定された要素数}} \quad (1)$$

と定義する。

当然ながら、品詞間の要素の相関性が高い巻号、頁、発行年月などの正解率は良いが、タイトルや研究報告名のように相関性の低い品詞では良い結果が得られていない。

## 7 文法による品詞の補正

### 7.1 文法の導入

最大一致文字列長を用いた品詞の判定だけではタイトル等を正確に判定することは困難である。そこで、参考文献項目にある程度の文法規則が見られることを利用して品詞の推定の補正を行なうことにする。

### 7.2 参考文献項目の文法

参考文献項目における文法規則は、実際の論文誌から抽出することになる。そのためには幾つかの品詞が集まって構成される句を定義する必要がある。句は以下のように定義する。

#### 7.2.1 タイトル部

通常タイトルとサブタイトルは並べて記述される。これらをまとめてタイトル部と呼ぶことにする。

#### 7.2.2 出版元情報

出版元は地名と結び付くことが多いため、出版元と地名を含むものを出版元情報と定める。

#### 7.2.3 雑誌情報

雑誌では大抵の場合巻や号が併記される。これらを含むものを雑誌情報とする。

#### 7.2.4 研究報告情報

研究報告の場合、雑誌と同様に巻や号が併記されるだけでなく、研究会を主宰した組織や開かれた地名も併記されることある。そこで、研究報告、巻、号、出版元情報をまとめて研究報告情報とする。

#### 7.2.5 書誌情報

出版元情報、雑誌情報、研究報告情報に加えて頁を加えたものを書誌情報とする。タイトルや著者名に関する情報がなくても、書誌情報のみで参考文献を特定することができる。

#### 7.2.6 参考文献項目

全ての品詞や句がまとまったものが参考文献項目となる。

### 7.3 確率文脈自由文法

参考文献の項目に対して、確率文脈自由文法を適用して各要素の品詞の推定を行う。確率文脈自由文法は文脈自由文法に対して確率を付与し、複数生じる構文木の順序付けを行うものである。

確率文脈自由文法では、各生成規則$A\rightarrow\alpha$に対して、

$$\Sigma_\alpha P(A\rightarrow\alpha)=1 \quad (2)$$

となる確率$P(A\rightarrow\alpha)$を付与する。ある特定の構文木$t$を得る文$w$の生成確率は

$$P(w,t)=\Pi_{A\rightarrow\alpha}P(A\rightarrow\alpha) \quad (3)$$

と、各生成規則の確率の和で表され、文$w$全体の生成確率は

$$P(w)=\Sigma_t P(w,t) \quad (4)$$

と表せる。確率文脈自由文法では、全ての文の生成確率の和に対して

$$\Sigma_w P(w)=1 \quad (5)$$

が成り立っている。

### 7.4 確率文脈自由文法導入の問題点

参考文献の項目に確率文脈自由文法を適用する際の障害は、参考文献の項目の各要素に対して、品詞からの生成確率を求めることが不可能で

ある点である。そこで、確率文脈自由文法を適用する範囲を前終端記号までとし、品詞と要素の間にはあらたな確率を付与することにする。

### 7.5 品詞確率

ある要素がどれくらいの確率でどの品詞を表すかを、前章の類似度を用いた文字列処理による判定から求めることができる。

正解率を求めた場合と同様の操作を行ない、品詞Aと判定されたときに実際は品詞Bである確率を

$$\frac{\text{そのうち実際の品詞が}B\text{である要素数}}{\text{品詞}A\text{と判定された要素数}} \quad (6)$$

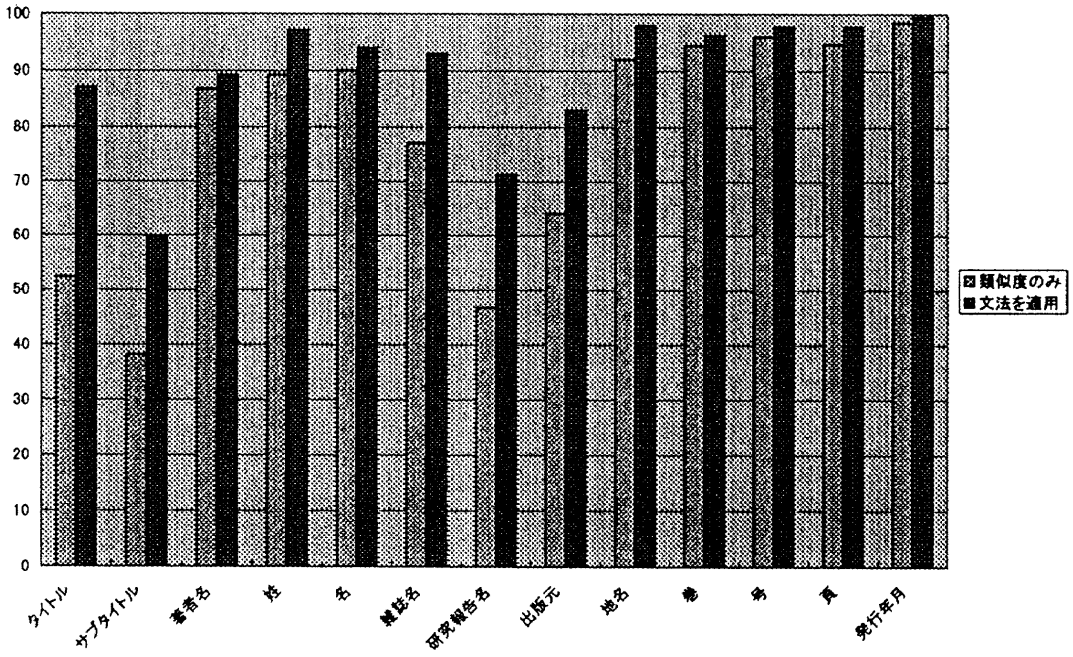



図 3　正解率

で定義する。

この確率を品詞確率と呼ぶことにする。

### 7.6　確率文脈自由文法の拡張

確率文脈自由文法において、品詞と要素の間の生成確率の代わりに品詞確率を用いる。このことにより、生成確率と品詞確率の全ての積をとっても構文木の生成確率とは異なるものになるが、この値をその構文木の得点としてそのまま用い、構文木の候補の中で最も得点の高いものをその参考文献項目の構文木とみなす。この構文木によって各要素の品詞をあらたに付与しなおす。

品詞確率が0となる場合があることから、まれに構文木が求められないことがある。その場合については文法による補正が行なえないため、文字列処理による判定をそのまま用いる。

## 8　評価

### 8.1　対象データ

情報処理学会論文誌1995年1月号の参考文献項目301件についてあらかじめ正しい品詞の付与をおこない、それを元に2月号～6月号の項目1676件の品詞の判定をおこなった。

### 8.2　プログラム

プログラムはjperl5でおこなった。確率文脈自由文法の性格上、文法規則を変形させるとその生成確率が変化してしまうために、構文木を2分木にすることができず、パーザ部分も含めて自前でコーディングをおこなった。

構文解析の方法は多々あるが、大抵は2分木を対象としているため、ここでは単純に再帰的にトップダウンに構文解析をおこなうことにした。動的計画法を用いて部分木ごとの得点を算出し、その結果をハッシュテーブルに保存しておくことにより、処理時間の短縮を図っている。

### 8.3　実験結果

1676件中17件がデリミタによる切り出しに失敗し、残りの1659件について各品詞の属性の正解率は以下のようになった。比較用に文法による補正を加えていないものも共に示してある。

サブタイトルの正解率が悪いが、これはサブタイトルをもつ文献項目が非常に少ないためである。研究報告も若干悪いが、これは文法的には雑誌と区別がつきにくいためであり、間違いのほとんどは雑誌である。

### 8.4　具体例

例5は品詞の推定がうまくいった書式例である。著者

Yasumoto, K., Higashino, T., Matsuura, T.and Taniguchi, K.: PROSPEX : A Graphical LOTOS Simulator for Protocol Specifications with NNodes, IEICE Trans. on Comm., Vol. E 75-B, No.10, pp.1015-1023 (1992).

例 5　品詞の判定がうまくいった例

Nehari, Z.: Conformal MaPPing, pp.333-392, McGraw-Hill, New York (1952) ; Dover, New York (1975).

例 6　品詞の判定がうまくいかなかった例

が他数おり、タイトルが複数に分割されてしまう場合、著者部分とタイトル部分の判定が困難であるが、デリミタの種類によらずともうまく認識される。

一方品詞の推定がうまくいかなかった書式例としては例6のものが挙げられる。

この例では一冊の本に対して複数の出版元がなどが示されている。この例のような文法は1月号には含まれておらず、正確な文法を認識できなかった。さらに出版元の間のデリミタに独自の記号を用いているため、要素の分割がうまくいっていない。

全般的な傾向として、典型的な文法であれば、各要素が細分化されていても比較的正確に認識できている。逆に未知の文法が用いられている場合は、各要素が細分化されていなくても品詞の正確な推定は困難である。

## 9　今後の課題と検討

本研究では、あらかじめ品詞が判っている既知のデータから、参考文献項目の品詞の推定が行なえることを立証した。今後の課題としては、以下のものを挙げられる。

### 9.1　デリミタの認識誤りへの対応

デリミタとして用いられる記号の認識誤りは、文字全体に対して低いため、現在では考慮されていない。デリミタの認識誤りを含む考文献項目は全体の2%以下程度であるが、品詞の認識の精度を向上させるためには無視出来なくなってくる。これらの誤りの大部分は、記号が脱落してしまうものや隣接した文字と予期せぬ結合を起こすものなど、対応が困難であるものだが、何らかの対策を検討する必要がある。

### 9.2　未知の文法への対応

参考文献項目の品詞の認識に失敗した原因の中に、未知の文法に対応できなかった為によるものの占める割合は多い。参考文献項目の記述は、著者による記述の揺れが大きい。そのためあらかじめ抽出した文法では対応しきれない例が見受けられた。さらなる精度の向上を図るためには、未知の文法に対応できる手法が求められる。

### 9.3　おわりに

本研究では、既存の出版物を扱う電子図書館において、その利便性の向上の一環として論文間にハイパーリンクを設けるために必要な技術として、OCR の認識誤りを含む参考文献の項目からタイトル等の情報を認識する方法を提案し、その有効性を検証した。

本手法では対象となる参考文献の既知未知を問わずに、参考文献項目の情報を認識することが可能であり、またデリミタに依存しておらず、同種の論文誌の項目に対してあらかじめ品詞を付与する方法を採用しているため、書式の異なる論文誌に対しても移植の負担が少ない。

今後の課題として、OCR 認識誤りへの対応の強化や未知の文法への対応を図ることによって、さらなる認識精度の向上が期待される。

## 参考文献

[1] 高須　淳宏, 文書画像データからの書誌情報の抽出とマッチング, 情報学基礎, 45-6, pp.33-38, 1997.

[2] Charniak,E., Statical Language Learning, The MIT Press, 1993.

学術情報センター紀要　第10号(1998)

研究論文

# HTTPメッセージのコンテンツ変換を行う共通フィルタサーバの設計と試作

# Design and Implementation of Common Filter Server for HTTP Message Contents Conversion

学術情報センター　相澤　彰子
Akiko AIZAWA
National Center for Science Information Systems
電子技術総合研究所　佐藤　豊
Yutaka SATO
Electrotechnical Laboratory

**要旨**

アプリケーションレベルでHTTPメッセージを中継するプロキシサーバは、広域ネットワーク上のHTTPを効率的に伝達するために重要な役割を果たす。また、プロキシサーバ上でのHTTPメッセージのコンテンツ変換は、HTTPによる情報流通の高度化に向けての多様な可能性を持つ。このことから本研究では、メッセージ中継時に既存のプロキシサーバと連携してコンテンツ変換を行う「共通フィルタサーバ」(Common Filter Server, CF Server)の実現に向けて、その設計および試作を行った。CFサーバの設計では、汎用のアプリケーションプロトコル中継サーバであるDeleGateの利用を前提として、DeleGateが持つCFI(Common Filter Interface)と呼ばれるコンテンツ変換機能を、HTTPメッセージヘッダのテンプレートを登録したデータベースを用いて拡張した。また試作はスクリプト言語であるPerlを用いて行い、一般的なクライアントとの間で動作確認を行った。今後の課題としてCFサーバとDeleGateとの連携、CFサーバの高速化、応用面での検討があげられる。

**ABSTRACT**

HTTP proxy servers, the application level gateways which transfer HTTP messages between servers and clients, play an important role in efficient transmission of HTTP messages over wide area networks. These proxy servers also provide a number of possibilities for the advanced use of the current HTTP-based information systems. With these points as background, this research focuses on the design and implementation of 'Common Filter Server' (CF server) which performs contents conversion of transit messages in cooperation with existing HTTP proxy servers. In our design of CF server, we assume DeleGate, a general-purpose application protocol gateway with unique contents conversion mechanism called CFI (Common Filter Interface), as cooperating proxy server and extend CFI function of DeleGate using a database of HTTP message header templates. Our CF server is implemented using script language Perl to confirm the interaction with popular HTTP clients. Future issues include cooperative operation between CF server and DeleGate, the speed up of the CF server, and consideration of application examples.

[キーワード]　HTTP、プロキシサーバ、情報変換フィルタ、応用層プロトコル中継、インターネット、DeleGate、CFI、共通フィルタサーバ

[Keywords]　HTTP, proxy server, information conversion filter, application protocol gateway, Internet, DeleGate, Common Filter Interface, Common Filter Server

## 1 はじめに

WWW(World Wide Web)の通信プロトコルであるHTTP(Hyper Text Transfer Protocol)は、現在インターネット上でもっとも通信量が多いアプリケーションプロトコルであり、情報サービスの高度化や通信回線の効率的利用の観点から、その運用技術が注目されている。

広域ネットワーク上でHTTPを効率的に伝達するためには、アプリケーションレベルでHTTPメッセージを中継して情報の流れを制御する「プロキシサーバ」が重要な役割を果たす。プロキシサーバの上では、異なるネットワークや計算機環境にあるクライアントとサーバがテキストや画像など種々の形式のコンテンツをやりとりする。ここで中継時に、プロキシサーバ上でコンテンツを加工する変換サービスを提供すれば、マルチメディアを含むこれらのコンテンツを、クライアントやサーバ各自の環境に適合させて効率的に受け渡すことができると考えられる。

このことから本研究では、プロキシサーバによるコンテンツ変換サービスに注目し、「プロファイル」と呼ぶHTTPメッセージヘッダのテンプレートを用いて、変換プログラムを自動的に起動したり、変換プログラムの名前をプロキシサーバに通知したりする「共通フィルタサーバ」(Common Filter Server, 以下CFサーバ)の実現を検討した。このように中継用のサーバとは独立した専用のサーバにより変換処理を実現することは、中継点での負荷の分散が期待できること、複数のプロキシサーバ間で変換機能や制御情報を共有できるなどの利点を持つ。

以下2.でCFサーバの背景および目的を述べ、3.でその動作概要と構成を示す。4.ではプロファイルの記述形式を定め、CFサーバ自身がコンテンツ変換を行いながらメッセージを中継する場合、およびCFサーバが受け取ったメッセージヘッダとプロファイルとの照合結果をプロキシサーバに通知する場合の2者について手順を定める。最後に5.で今後の課題やCFサーバの応用について考察する。

## 2 CFサーバの背景と目的

### 2.1 HTTPメッセージ

HTTPメッセージはRFC822[1]で定められたmessage記述形式を用いて定義されており、伝達されるメッセージのメタ情報(属性や制御情報)を含む「ヘッダ」と伝達するメッセージの内容そのものである「ボディ」すなわち「コンテンツ」から構成される。図1にHTTP/1.1[2]による仕様定義の一部を示す。

```
generic-message   =   start-line
                      *message-header
                      CRLF
                      [ message-body ]
start-line        =   Request-Line | Status-Line
message-header    =   field-name ":" [ field-value ] CRLF
field-name        =   token
field-value       =   *( field-content | LWS )
field-content     =   < the OCTETs making up the
                      field-value and consisting of either
                      *TEXT or combinations of token,
                      tspecials, and quoted-string >
...
```

図1 HTTP/1.1による仕様定義の一部

図の定義において、generic-messageは全メッセージに共通の記述形式であり、<CRLF>(空行)によりヘッダ部とコンテンツ部に区切られている。generic-messageの1行目であるstart-lineは要求または応答のメッセージ種別やHTTPのプロトコルバージョンなど基本的な情報を持つ。2行目から始まるmessage-headerは、各行が<属性名：属性値>の組で表されるテキスト情報であり、種々の制御情報を含んでいる。図中には示していないが仕様中ではさらに、図中のmessage-headerを、(1)HTTPのプロトコルバージョンに固有のメッセージ属性を示すgeneral-header、(2)要求または応答ごとに共通な属性を示すrequest/response-header、(3)メッセージのコンテンツに関する種々の属性を示すentity-header、の3つの構成要素で定義している。このうち最後のentity-headerでは、アプリケーション側で新たな属性を追加することを許しており、メッセージのコンテンツに応じた柔軟な記述が可能になっている。

### 2.2 HTTPプロキシサーバ

プロキシサーバとは一般に、クライアントから受け付けた要求メッセージをサーバに中継し、サーバから受け取った応答メッセージをクライアントに転送するアプリケーションレベルのプログラムである。

図2に示すように、通常アプリケーションはIPが提供するクライアントーサーバ間の経路を論理的な1本の通信路とみて通信を行い、メッセージ中継や経路制御はすべてIP層において行われる(図2(a))。これに対してプロキシサーバを介してメッセージの中継を行う場合には、アプリケーション層がメッセージ中継に

かかわることになる(図2(b))。

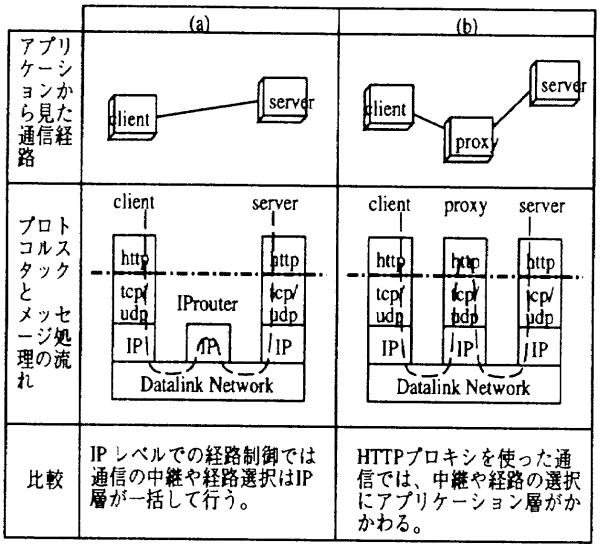



図2　HTTP プロキシによるアプリケーションレベルでのメッセージ中継

プロキシサーバでは中継時に、いったんメッセージをアプリケーションプログラムが読み込んで解釈した上で転送を行うため、専用ルータによるIPレベルでの中継と比較するとスループット性能は低下する。しかしプロキシサーバは中継メッセージのキャッシュやコネクション保持などの機能を備えており、これらを活用することによって、クライアントが直接サーバとやりとりを行う場合と比較して、効率のよい通信が可能になる。

さらにプロキシサーバどうしを多段に接続することにより、コンテンツを考慮した柔軟な経路制御をアプリケーションレベルで行うことも可能である。たとえばSquid[3]などのソフトウェアでは、ICP(Internet Cache Protocol)と呼ぶ制御用のプロトコルを用いたプロキシサーバの組織的な連携が可能になっている。すなわち、プロキシサーバどうしがキャッシュのエントリー情報を交換し、必要に応じてコンテンツを相互に利用しあう協調キャッシュ機能が提供されていて、メッセージの中継時に最寄りのキャッシュからコンテンツを受け取ることができる。

### 2.3　CF サーバの目的

HTTP ではメッセージ自体が、中継やコンテンツの表示に必要となる属性情報(たとえばメッセージ種別やブラウザで表示可能なコンテンツのタイプなど)をすべて含むことを前提としている。そこで、このようなヘッダ情報を適切に管理して中継時に参照するようにすれば、表示クライアントではなくプロキシサーバにおいて、ネットワーク負荷や通信相手の計算機環境を配慮したコンテンツの変換処理を行うことが可能になる。

本稿で利用を前提としているDeleGate[4]では、CFI(Common Filter Interface)という仕組みによって、すでにこのような機能を実現している。具体的にはプログラム起動時のオプション指定によって、中継時のローカルURLの書き換えや漢字フィルタの適用を自動的に行えるようになっており、サイト固有のネットワーク環境を損なわずに外部ネットワークと通信が行えるよう工夫がされている。しかし中継時の変換処理は上記のように単純なものだけではなく、たとえば機械翻訳やメディア変換など、メッセージの属性に応じた多様な形態が考えられる。また高速なHTTPメッセージ中継が要求される状況では、中継時の負荷の問題から、DeleGate上ですべての変換処理を実現することには限界がある。

そこで本研究では、DeleGateのCFIと同様の変換機能を、CFサーバと呼ぶ独立の専用サーバを用いて実現することを考え、その設計およびプロトタイプの実装を行った。すでに述べたように、このように専用のサーバで変換処理を実現することで、中継点での負荷分散や、複数のプロキシサーバ間での情報共有などの利点が期待できる。次節ではこのようなサーバの概要について述べる。

## 3　CF サーバの動作概要と構成

### 3.1　CF サーバの動作概要

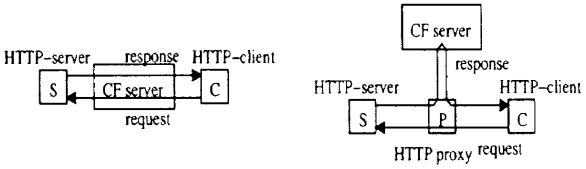
まず、CFサーバは、プロファイルと呼ぶメッセージヘッダのテンプレートを管理していて、動作時に実際のメッセージヘッダと照合をとることにより、コンテンツに対して適用すべき変換プログラムの名前を知る。以下、このような変換プログラムをDeleGateにならってCFIスクリプトと呼ぶ。CFIスクリプトは一般的にコンテンツに対する変換フィルタを指している。したがって、標準入力からバイト列入力を受け取り、何らかの処理によって加工し、その結果を標準出力に出すプログラムはすべてCFIスクリプトとして用いることができる。漢字コード変換フィルタは広く使われている例であるが、画像の符号化方式変換や自動翻

訳、コンテンツに対するキーワード付与や点数付けサービスなどもCFIスクリプトとして実現可能である[5]。

次にこのようにして中継メッセージに対応したCFIスクリプトが得られると、CFサーバは受け取ったメッセージに応じて以下の2種類のモードで動作する。

- **動作モード(1)**
  それ自体が単独でHTTPプロキシサーバとしてクライアントとサーバ間の中継を行い、中継するメッセージに対してCFIスクリプトを適用する(図3(a))。
- **動作モード(2)**
  実際に中継を行う他のプロキシサーバからの要求に応じてCFIスクリプトを適用したりCFIスクリプトの名前を通知したりする(図3(b))。

具体的には、Content-Type: image/mpegである応答メッセージに対して、Pフレームを除いて送出するメディアスケーリングを行ったり(動作モード(1))、DeleGateや他のCFサーバからの問い合わせに対して適用すべきCFIスクリプトの名前を返したりする(動作モード(2))。それぞれの動作モードにおける具体的な手順については4.2および4.3で述べる。




(a)単独のHTTPプロキシサーバとして動作する場合　　(b)他のプロキシサーバからの要求に対応して動作する場合

図 3　CFサーバの2種類の動作モード

なお、CFIスクリプトはHTTPの要求と応答のいずれのメッセージについても適用可能であるが、簡単のため本稿では、サーバからクライアントに向かう応答メッセージへの適用を前提として手順を述べる。また現在は、メッセージヘッダを書き換えることは許しておらず、変換処理はメッセージのコンテンツ部分にのみ適用される。

### 3.2　CFサーバの構成

以上の動作を実現するために、CFサーバは「プロファイルデータベース」、「中継／変換サーバ」、「汎用プロキシサーバ」の3つの要素から構成されている。

(1) **プロファイルデータベース**
HTTPヘッダのテンプレートを管理するデータベース。HTTPメッセージの特定の属性のパターンを条件として、中継時に起動するべきCFIスクリプトを登録する形になっている。

(2) **中継／変換サーバ**
ホストマシン上の指定したポートにおいて外部のクライアントやサーバやプロキシからHTTPメッセージを受け取り、プロファイルデータベースを参照するプログラム。必要に応じてCFIスクリプトを起動してメッセージのコンテンツを変換したりCFIスクリプト名を相手に通知したりする。

(3) **汎用プロキシサーバ**
通常に配布されているプロキシサーバ。中継／変換サーバからHTTP要求メッセージを受け取ると、ローカルディスクあるいは遠隔サーバからデータを取り寄せて応答として返す。

各動作モードにおけるそれぞれの構成要素の役割については、次節の図4および図5に示してある。

汎用プロキシサーバをCFサーバの中に組み込んでいるのは、外部に対してHTTPの仕様を満たすプロキシとしての動作を保証するためである。中継／変換サーバ自体は処理の簡単化のため、動作モード(2)の場合を除いては受け取ったHTTPメッセージの解釈は行わない。中継メッセージのヘッダを一連の文字列とみてデータベースのエントリーと照合をとるだけである。受け取ったメッセージのプロトコル上の構文チェック等の機能は汎用プロキシサーバに一任する形になっている。

## 4　CFサーバの実装

### 4.1　プロファイル定義ファイルの記述形式

現在の実装では、cfi.confとaliases.dbという2つの定義ファイルをプログラム起動時に読み込むことで、プロファイルデータベースの初期化を行っている。いずれの定義ファイルについてもPerlの正規表現が使用可能である。実際、本研究で試作したCFサーバはPerlで実装しており、動作時には定義ファイルの各行をPerl検索コマンドのパターンとみなして照合を行っている。

cfi.confでは、<CRLF>をプロファイルどうしの区切り記号として、各プロファイルを以下の形式で記

述する。

```
--
field-name1: perl-regular-expression1
field-name2: perl-regular-expression2
...
+
Filter: <CFI-script-name>
--
```

先頭から'+'符号までがプロファイル、それ以降 Filter で始まる行は、指定されたプロファイルに適合するメッセージに適応すべきフィルタの名前(あるいは仕様)である。このようにプロファイルは、field-name: field-value の形式のテンプレートの集まり(以下、属性テンプレートと呼ぶ)として定義されている。プロファイルデータベースの参照では、登録されたすべての属性テンプレートに対してメッセージヘッダに照合する行が存在する場合にのみ、両者の間で照合関係が成立したとみなす。複数のプロファイルに対して照合が成立する場合には、最初に照合した1つだけを結果として返す。

cfi.conf の中で@で始まる文字列はマクロ変数に対応している。これらのマクロ変数は aliases.db において、UNIX の /etc/aliases と同様の形式で以下のように記述する。

```
aliase-name: member1, member2, ...
```

なお、以上の cfi.conf の記述形式は、DeleGate の CFI 起動定義ファイルである CFI スクリプトの形式に準じている。DeleGate からの拡張は、マクロ変数の使用を許したこと、Perl の正規表現を可能にしたことの2点である。

最後に DeleGate で配布される CFI 起動定義ファイル中のサンプル記述を用いた簡単な例を示す。この例では、「EUC コード表示のクライアントに対してテキストを中継する場合には漢字変換フィルタでコード変換を行うこと」を指示している。DeleGate からの変更点として、euc-viewers をマクロ変数で定義し、aliases.db の中での記述に Perl の正規表現を用いている。

(1) cfi.conf の記述例

```
--
#
# character code conversion for specific
# User-Agents
#
Content-Type: text/
User-Agent:\@euc-viewers
Output-Charset: X-EUC
+
Filter: nkf -e
--
```

(2) aliases.db の記述例

```
#
# definition of euc-viewers
#
euc-viewers: (Lynx | NCSA Mosaic)
```

### 4.2 単独の中継／変換サーバとしての動作

CF サーバが通常の HTTP メッセージを中継する場合には、CF サーバ上でコンテンツ変換処理を行う。この場合に、通信相手となるクライアントやサーバが通常のプロキシサーバとの違いを意識せずに CF サーバと通信が行えることが必要条件である。具体的には図4に示す手順で通信を行う。

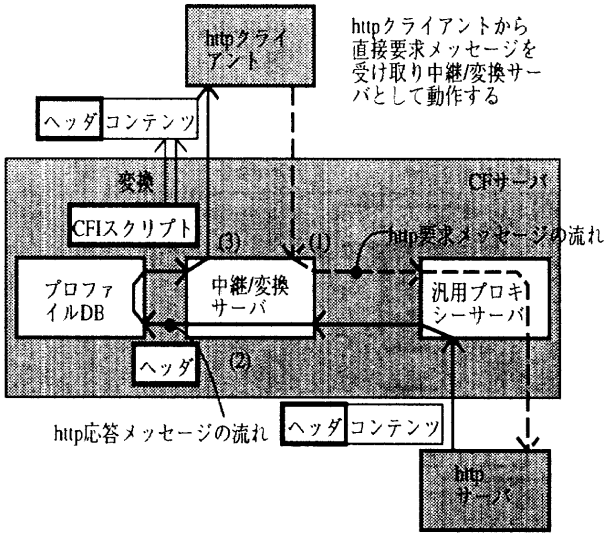



図 4　単独の中継／変換サーバとしての動作手順

- **手順(1)**：CF サーバがクライアントから受け取った要求メッセージを汎用プロキシサーバに転送する。
- **手順(2)**：CF サーバは汎用プロキシサーバから応答メッセージを受け取ると、そのヘッダ情報に基づきプロファイルデータベースを参照して、照合する CFI スクリプトの有無および適合する場合にはその名前を得る。

・**手順**(3)：CFサーバがクライアントに応答メッセージを転送する。照合するCFIスクリプトが存在する場合には、そのCFIスクリプトを起動してコンテンツ変換を行いながら、クライアントに応答メッセージを転送する。照合するCFIスクリプトがなければ、そのままクライアントに応答メッセージを転送する。

ここでプロファイルデータベースへの参照は、ヘッダとコンテンツの区切り(<CRLF>)を読み込んだ時点で行う。データベースの参照が終了したら、以降はサーバから順次コンテンツを受け取り、必要に応じてCFIスクリプト変換フィルタをかけながらクライアントへの転送を行う。効率のため、転送時に全メッセージをバッファに読み込むことはしない。

### 4.3 プロファイルデータベースの参照によるDeleGateとの連携

この場合にはCFサーバはDeleGateからの要求に応じてデータベースを参照するだけで、実際のコンテンツ変換処理はDeleGate上において行われることになる。具体的にはCFサーバとDeleGateは図5の手順で通信を行う。

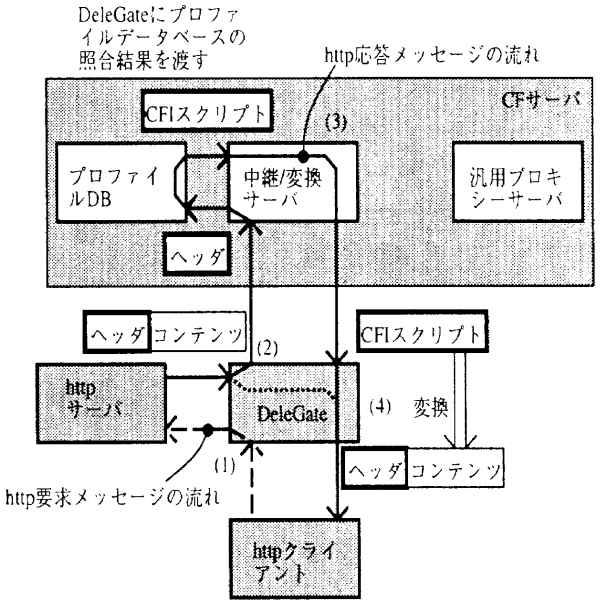



図5　プロファイルデータベースの参照によるDeleGateとの連携手順

・**手順**(1)：DeleGateがクライアントから受け取った要求メッセージをサーバに転送する。

・**手順**(2)：DeleGateはサーバから応答メッセージを受け取ると、ヘッダ情報をCFサーバに送る。

・**手順**(3)：CFサーバはプロファイルデータベースを参照して、照合するCFIスクリプトの有無および照合する場合にはその名前をDeleGateに通知する。

・**手順**(4)：DeleGateはCFサーバから受け取った情報をもとにクライアントに応答メッセージを転送する。照合するCFIスクリプトが存在する場合には、そのCFIスクリプトを起動してコンテンツ変換を行いながら、クライアントに応答メッセージを転送する。照合するCFIスクリプトがなければ、そのままクライアントに応答メッセージを転送する。

CFサーバとDeleGateの間でやりとりするメッセージの形式として、たとえば以下が考えられる。まず、DeleGateからCFサーバへの要求において、DeleGateはHTTP/1.0またはHTTP/1.1のPOSTメソッドを用いてCFサーバに対してデータベースの参照を依頼する。HTTP/1.0を用いる場合には、このPOSTメッセージのヘッダを以下のようにする。

```
POST / HTTP/1.0
Content-Type:
        application/x-http-response-header
X-[additional-header-line] *
```

POSTメッセージのコンテンツ部にはDeleGateがサーバから受け取った応答メッセージのヘッダ部の内容がそのまま渡される。またPOSTメッセージヘッダ部のX-で始まる属性行は、本来の応答メッセージには含まれないが、中継／変換処理において重要な情報、たとえばクライアントホストの名前などをDeleGateからCFサーバに対して通知するために用いる。一方、HTTP/1.1を用いる場合には、POSTメッセージのヘッダは以下のようにして、Connection: keep-aliveのオプションにより、一度設定したDeleGateとCFサーバの間のコネクションを保持したまま連続してプロファイル照合を行えるようにする。

```
POST / HTTP/1.1
Content-Type:
        application/x-http-response-header
X-[additional-header-line] *
```

```
Connection: keep-alive
Transfer-Encoding: chunked
```

ここで連続するメッセージの区切りを明示的に示すために、コンテンツを任意長のブロックに分割してブロックごとに長さを指定するChunked Encodingを用いている。POSTメッセージのコンテンツ部にはHTTP/1.0の場合と同様にDeleGateがサーバから受け取った応答メッセージのヘッダ部の内容がそのまま渡される。

CFサーバからDeleGateへの応答においては、上記に対応して、CFサーバはHTTP/1.0またはHTTP/1.1を用いてDeleGateに対してデータベースの参照結果を応答する。HTTP/1.0を用いる場合には、応答メッセージのヘッダは以下のようになる。

```
HTTP/1.0 200 OK
Content-Type: application/x-cfi-script
```

同様にHTTP/1.1を用いる場合には以下のようになる。

```
HTTP/1.1 200 OK
Content-Type: application/x-cfi-script
Connection: keep-alive
Transfer-Encoding: chunked
```

いずれの場合についても、メッセージのコンテンツ部には、プロファイルデータベースに登録されていたCFIスクリプトの名前が書き込まれる。

## 5　今後の課題

これまでにCFサーバのプロトタイプをPerlで実装し、4.2に述べた方法によってクライアントからCFサーバを直接呼び出して中継／変換動作を実現した。4.3の方法によるDeleGateとの連携については、今後さらに検討する予定である。

CFサーバの動作確認は試験的に登録した高々数個のプロファイルを用いて行ったため、CFサーバの処理速度がボトルネックになることはなかった。しかし、実際に多数のプロファイルを登録してサービスを行う場合には、中継、文字列照合、CFIスクリプト起動処理などの高速化が課題となると考えられる。

またCFIスクリプトについては多様な使い方が想定されるため、標準的なライブラリを整備しておくことも重要である。CFIスクリプトライブラリとしてまず考えられるのは、表現形式を自動変換する漢字コードや画像符号化方式変換フィルタや言語間の変換を行う自動翻訳プログラムなどである。それ以外にも、利用者管理データベースなど大規模なデータベースを用いてアクセス制御を行うもの、コンテンツの要約など意味的な変換処理を伴うもの、キーワード抽出などの処理によりコンテンツの選別を行うものなど、多様な情報フィルタが考えられる。

さらに、複数の中継サーバから共通に参照可能であることを利用すると、プロキシサーバ間の並列動作をスケジュールしキャッシュの分散を管理するための集中制御サーバとしても応用できる可能性がある。プロファイルごとに利用統計情報を収集して、従来の方法よりも柔軟なログ情報の収集が可能な統計サーバを実現することも考えられる。

## 参考文献


[1] Crocker,D,H., "Standard for the Format of ARPA Internet Text Messages," STD11, RFC 822, UDEL (1982).

[2] Fielding,R.; Gettys,J.; Mogul,J.C.; Frystyk,H.; Berners-Lee,T., "Hypertext Transfer Protocol - HTTP /1.1," IETF Internet-Draft <draft-ietf-http-v11-spec-06> (1996).

[3] http://www.nlanr.net/Squid/(Squid Internet Object Cache).

[4] http://wall.etl.go.jp/delegate/(ETL DeleGate Home Page).

[5] http://www.dna.affrc.go.jp/ugawa/w3conf-japan/(DeleGate Proxy を使ったWeb情報の英和逐語訳サーバ -EtoJ_Proxy-).

学術情報センター紀要　第10号(1998)

研究論文

# An Application-oriented Approach for HyTime Structured Document Management

# HyTime構造を持つ文書管理のための応用指向アプローチ

Frederic ANDRES
National Center for Science Information Systems
学術情報センター　フレデリック　アンドレス
John F. BUFORD
University of Massachussetts Lowell/Interactive Media Group
マサチューセッツ大学ローウェル
インタラクティブメディアグループ　ジョン　F.ビュフォード
Kinji ONO
National Center for Science Information Systems
学術情報センター　小野　欽司

**ABSTRACT**

In this article, we point out the important functionality needed by emerging multimedia applications such hypermedia presentations or digital library retrieval systems which require next generation database systems. Uniform management of hypermedia data is required to be suitable to various kinds of applications with different characteristics (data types, data model, data format, i/o devices). DBMSs provide efficient data storage facilities but still lack of customizability according to target applications. Moreover, content-based and structure-based retrieval managements are required by modern information retrieval systems. In order to combine the requirements of information retrieval systems and opened DBMS, we have implemented information retrieval functions inside the Application-Oriented DBMS Phasme. The document representation is either SGML or HyTime. SGML or HyTime documents are stored inside Phasme and are accessed using full text retrieval functionality. Such functionality are implemented as Phasme plugins and are stored inside Phasme. The storage management of the documents is independent from the way the user application will retrieve them.

The developments achieved so far inside the AHYDS project (Active HYpermedia Delivery System) currently under process at NACSIS illustrate the chosen architecture design of the retrieval system. The performance of the current prototype is evaluated on a 40 Gbs document Benchmark showing that our approach yields excellent results.

**要旨**

本文では、ハイパーメディア表現のようなマルチメディアのアプリケーションや次世代データベースシステムに要求されるデジタル検索システムに必要な機能であるマルチメディア配送バスについて考察している。種々のアプリケーションが異なった特性(データタイプ、データモデル、データフォーマット、I/Oデバイス)と適合する為に、ハイパーメディアデータの管理が要求される。DBMSはそのアプリケーションに左右されない効率のよいデータ記憶機能を供給する必要がある。さらにコンテンツと構造の基盤となる検索管理は最新情報検索システムよって要求される。情報検索システムとDBMSのオープン化の必要条件を満たす為に、本論文で、応用指向DBMS Phasme

の内部検索機能の実装をｆしている。文書表現はSGMLまたはHyTimeである。SGMLまたはHyTimeの文書はPhasme内部に記憶され、最大限のテキスト検索機能を使用しアクセスされる。このような機能は、Phasmeのプラグインとして使用され、Phasme内部に記憶される。文書の記憶管理は、ユーザーアプリケーションの検索技法から独立している。

NACSISにおいてAHYDS(アクティブ　ハイパーメディア配送システム)プロジェクトの検索システムのアーキテクチュアーデザイン過程にあるプロトタイプの性能は、40GBドキュメントに対するベンチマークテストにおいて良い結果を得ている。

[Keywords]　Active DBMS, Hypermedia, QoS, Document retrieval, Query Processing
[キーワード]　アクティブDBMS、ハイパーメディア、QoS、文書検索、問い合わせ処理

## 1 Introduction

During the last years, multimedia information systems based on client/server architecture have integrated various kinds of user tools such as authoring tools, browsers, presentation tools, or animation tools. Information themselves are also various (e.g. various types, various formats). Meanwhile, hypermedia information relate multimedia data by linking them together and enable navigation through links.

Existing systems do not yet provide all the functionality required by the hypermedia information systems. As an example, it is necessary to support content-based and structure-based retrieval [Buford97] as well as database query mechanisms for hypermedia document management. Moreover, information retrieval issues are a key input for document management as it is mentioned in [Kow98]. Also database services (concurrency control, recovery, versioning) are very important for efficient multimedia management.

The contribution of this paper is twofold: first we present the implementation of an hypermedia retrieval system using Phasme DBMS, and second we evaluate the performance of this implementation.

### Combining an Application-oriented DBMS with an Information Retrieval

This paper shows how the hypermedia document description in HyTime[HyT97] and retrieval functions can be mapped to the application-oriented system Phasme, a state of the art, parallel database micro-kernel. Phasme stores data in a uniform way independently of the various kinds of user applications (different data formats, data types and semantics). Moreover, it provides traditional DBMS services such as persistency, concurrency control, recovery and versioning. The vertical customisability (from the data definition to the execution model) enables the Application-Oriented DBMS to be tailored according to the requirements of the application (e.g. index types, word matching algorithm).

### High Performance Document Retrieval

This paper presents a system that execute information retrieval with high performance. This system translates the hypermedia document description (in HyTime) into the Phasme EBG data structure. Phasme is used as a back-end for the information retrieval execution. The demonstration of the performance has been done using the Binary Document Management[AA95] benchmark and a 5Gbs set of hypermedia documents. The data set, and the queries of this benchmark illustrate the performance of our system.

### Scope and outline of this paper

Object-relational databases typically bring together relational table storage and query processing, and the object-orientation of object-oriented DBMSs.

Object-Relational systems, with a capability of vertical customisability (adaptability from the data definition to the execution model) and with a uni-

form data storage, will be able to adapt themselves dynamically according to the application requirements and to provide high performance.

The end-goal of the AHYDS project is to demonstrate a system with all these features inside an hypermedia application. The work described in this paper is limited to document retrieval management.

The organization of the remainder of the paper is as follows. Section 2 describes the architecture of the AHYDS platform. In Section 3, technical aspects of the hypermedia delivery system are sketched, while Section 4 analyses the performance experiments and results. Some related works are presented in Section 5. Finally, Section 6 summaries and concludes the paper.

## 2 The Active HYpermedia Delivery System (AHYDS)

This section reminds the major requirement of hypermedia retrieval systems. Then, the management of SGML/HyTime Documents within an AODBMS is discussed. Finally the architecture of the AHYDS platform is described.

### 2.1 Requirements of hypermedia document retrieval applications

Hypermedia document management in many library applications requires the following features from the information server:

(1) *Support for structured document management*.

Such documents can include only text or text and media information such as audio, images, and moving pictures. Several document languages such as SGML, HyTime, or ODA have been developed in order to satisfy this requirement.

(2) *Support of DBMS services*

Transaction control, data integrity and data sharing, and version control are required for handling a large amount of multimedia data under dynamic multi-user environments.

(3) *Uniform data management independently of the application data model*.

Uniform storage structure is appropriate to deliver information to different kinds of multimedia applications (presentational type, conversational type) based on different kinds of data models (relational, object oriented).

(4) *Durability against the standardization*.

Several standard generalized markup languages have been proposed. SQL/MM is now emerging to support full text retrieval functions. The interface of the multimedia server must be flexible and customisable enough to support new languages.

(5) *Efficiency and quality of service for the data delivery*.

Combining functionality with efficiency and quality of services is mandatory in order to support both time-dependent and resolution -dependent media data.

### 2.2 Handling SGML/HyTime Documents within an AODBMS

We have chosen to focus on the integration of the information retrieval functionality inside the AODBMS for the following reasons: (1) it allows a greater flexibility and extensibility with regards to the information retrieval system's requirements in terms of standards (e.g. information manipulation languages, information representation, application data models, data exchange); and (2) it provides high performance and customisability mechanisms to assure durable hypermedia systems. The advantages of such a coupling are a) process structure -based and content-based query processing closed to the storage data, b) high degree of concurrency, and c) update management inside the DBMS.

Such an integration reduces the number of layers between the application and the document storage manager.

The advantages are: a) a full AODBMS functionality, e.g. vertical customisability of the kernel to support efficiently information retrieval functionality, uniform storage management independently of the application's semantics;

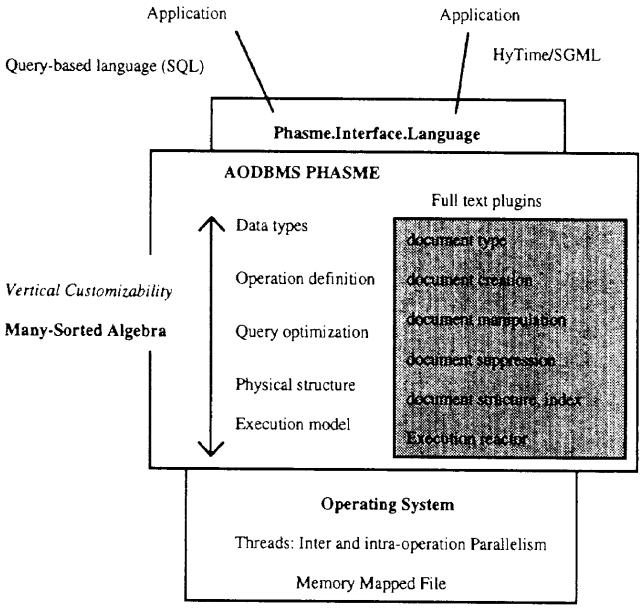



Figure 1 System architecture

```
Plug-ins HyDocument;
  ITEM HyDocument
  ADD     hydocADD;
  DEL     hydocDEL;
  GET     hydocGET;
  SIZE    hydocSIZE;
  COMP    hydocCOMP;
  INIT    hydocNULL;
  INDEX   hydocINDEX;
  MEM     hydocMEM;
  TOA     hydoc2a;
  ATO     a2hydoc;
END HyDocument;
```

Figure 2 Hypermedia Document Plugins

b) no overhead due to the traditional mapping between application data and database storage and c) the durability of the information system against the evolution of standards, and the hardware trends.

The overall architecture is shown in figure 1. Information about the SGML/HyTime support can be found in [Techno96]. It defined the framework to store and to retrieve SGML and HyTime documents. A short overview is given in the next section.

One of the objectives of the AHYDS (Active Hypermedia Delivery System project) is to use an application-oriented database application framework for structured and hypermedia document storage. The data representation inside the database is based on the Extended Binary Graph structure (EBG) of Phasme[AO97]. Each component of the document corresponds to an EBG. The EBGs corresponding to a document's elements make up a hierarchy. Leaf EBGs contain text or media information.

### 2.3 Architecture of the AHYDS platform

We applied the application-oriented approach to the SGML/HyTime framework described in the previous section.

The integration of the IR-functionality with the AODBMS requires the design and the development of plugins in order to customise vertically the Phasme DBMS. On the one hand, hypermedia document retrieval processing is done inside the DBMS kernel closed to the data storage manager. On the other hand, the plugins mechanism enables to adapt the DBMS according to the real application requirements.

The Hypermedia Delivery System allows a dynamic creation of new components of hypermedia documents. Combined structure- and content-oriented queries are done within the user query language or user application requirements.

## 3 Technical Aspects of the Hypermedia document management

### 3.1 The HyTime plugins

SGML/HyTime support is provided by the HyMinder Library[Techno96].

SGML/HyTime documents are stored in the Phasme DBMS. Each SGML/HyTime element is stored as a SGML/HyTime item of Phasme. It enables various granularity creation and manipulation of hypermedia documents. The Hypermedia docu-

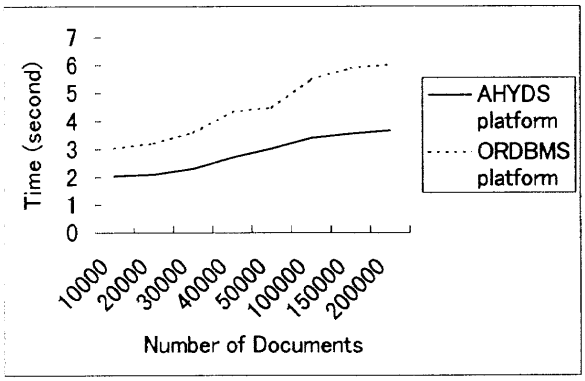



Figure 3　Document retrieval performance using AHYDS platform and using an Object-Relational DBMS (ORDBMS).

ment plugins is given in Figure 2.

The plugins includes both hypermedia document manipulations, conversion from EBG format to ascii format, and index management. For example, hydocADD and hydocDEL respectively inserts into Phasme data structure and deletes a document from Phasme data structure. The operation hydocGET retrieves a document or some parts of it from the database. A sample of the Data Type Definition (DTD) is given in appendix A.

### 3.2 Document Retrieval Queries

Document retrieval queries enable both structural and content-based access. To illustrate our approach, the document retrieval query part is formulated as arguments of the plugins operation hydocGET. The examples of queries are based on the many-sorted algebra of the Phasme EBG data structure.

**"Select all the paragaphs of documents having an image and an audio summary about 'Kamakura temples' ":**

hydocGET(hydoc.paragraph, hydoc.image.isabout("Kamakura temples")
and hydoc.audio.isabout("Kamakura temples"));

**"Select the title and the abstract of documents created in 1996 containing a paragraph about 'Soccer'":**

hydocGET((hydoc.title,hydoc.abstract),hydoc.datecreation(1996) and hydoc.paragraph.isabout("soccer"));

## 4 Performance Experiments

This section describes the results of experiments that measure the efficiency of document manipulation inside the AHYDS platform.

### 4.1 Configuration

Our server platform was a SPARC 1000 with 128 Mb of main memory and 8 SPARC processors. It has a RAID disk (level 0) with 60GB as capacity. The document set includes tags such as title, author, abstract, keywords, and at least 3 chapters with images and videos. The test-bed server of the Phasme DBMS version 2.02 has been used as the data storage manager of AHYDS. The client test software just retrieves the document from the server.

### 4.2 Performance results

The first set of measurements evaluates the performance of the access method of the hypermedia document in the context of the text retrieval issues. Figure 3 shows the average elapsed time of document retrieval queries varying the number of documents from 10 000 to 200 000. The hypermedia document identifier is randomly chosen for one couple of measures from the Hypermedia application running on the ADHYDS platform and on the ORDBMS platform. The average size of a document is 0.2 MB. The maximum size of documents managed inside both the AHYDS platform and the ORDBMS is 40 Gbs of documents.

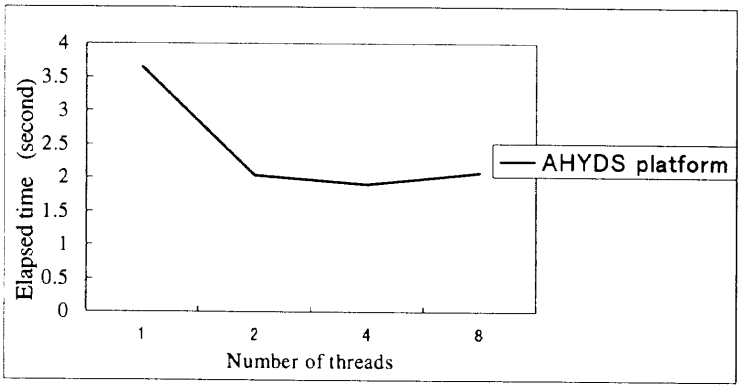



Figure 4 Hypermedia retrieval and intra-operation parallelism.

#### 4.2.1 Inter-operation parallelism

Figure 4 shows the influence of the intra-operation parallelism of Phase DBMS on a document retrieval access. Database is set to 200 000 documents. This experience was run on a SUN ULTRA 2 workstation. We varied the number of threads from 1 to 8 threads. We have horizontally clustered the set of hypermedia documents. In this way, the set of threads can be linked to the set of hypermedia document buckets. The intra-operation parallelism mechanism of the execution model considerably improves the document retrieval performance.

The multithreading management over 4 threads introduces an overhead due to the small number of processors. There are overheads due to some thread swappings over the processors. Some improvements will be introduced using pattern programming approach.

#### 4.2.2 User scalability and Elapsed Time

An important issue is the scalability and the influence on the elapsed time of the increase of the number of users in order to receive a complete document. Figure 5 shows the variation of the performance (elapsed time) of the AHYDS platform increasing the number of users. We verified that the AHYDS system supports the increase of users. The ORDBMS is becoming better when the number of users increase due to caching effects. Then, the elapsed time is increasing as the caching effect is no more improving the elapsed time according to the number of users.

#### 4.3.3 User scalability and number of documents delivered per second

An important issue is the behavior of a delivery system on the number of delivered documents according to the number of users querying the system. Figure 6 shows the variation of the number of documents delivered per second of both the AHYDS platform and the ORDBMS-based platform increasing the number of users.

The ORDBMS platform does not support the scalability when we increase the number of users.

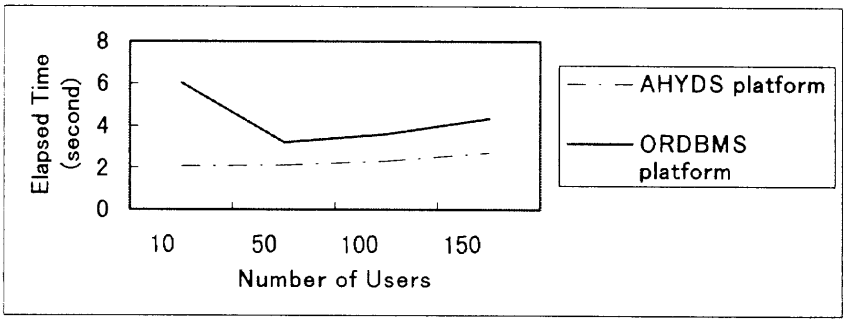



Figure 5 Influence of the number of users on the elapsed time

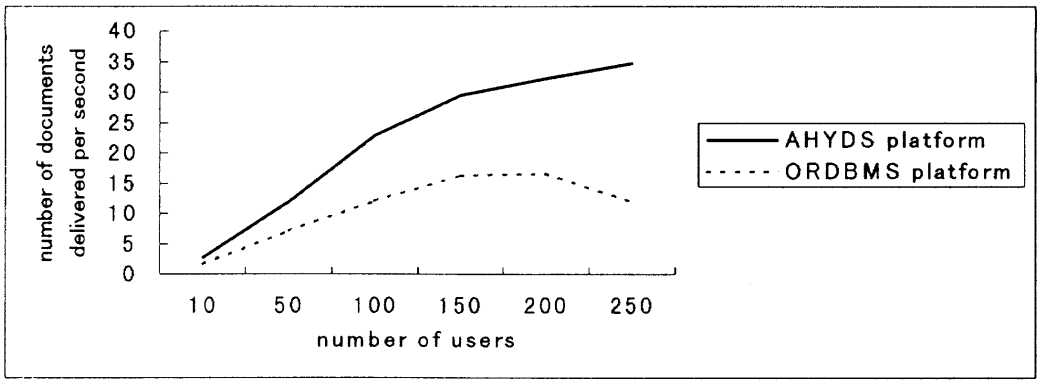



Figure 6　Influence of the number of users on the number of documents delivered per second

Our AHYDS supports the increase of users even if the number of users goes over 200 users.

## 5 Related Works

Quite a number of research prototypes of hypermedia database management systems[AKM94, CACS94, Hara96, Kim96, Volz96, Ozsu97] have been developed over the last years, mostly on OODBMS platforms.

As an example of commercial development the Oracle Universal Server[ORA96] is extensible for the development of so-called "cartridges" which are manageable objects. The Oracle Full text retrieval (TextServer) has been extended to manage new types of data inside a document including images and videos. It will be compared in the future in order to stress the hypermedia document management.

## 6 Conclusion

In this paper, we have described our approach to build an information engine system providing the integrated functionality of information retrieval system based on non query languages (SGML, HyTime) inside the Phasme Application-oriented DBMS. We have pointed out the problems that have arisen when traditional DBMSs are used. Further, we believe that our approach is rather flexible for the following reasons: (1) Uniform data management independently of the language or text representation standards. (2) Customisability of the information server according to different users' tools. (3) the plugins-based hypermedia document management.

The performance evaluation demonstrates convincingly that a DBMS kernel based on an application-oriented capabilities can be scaled to manage documents and to support information retrieval.

The following issues remain for future investigation: integration of HyQ (OQL-based query language) for HyTime information retrieval with the AHYDS platform, relevant feedback from very large data set (40 Gbs of data). Very large database experimentation will be done inside the Text REtrieval Conference (TREC) evaluation for the text retrieval processing. Beyond that, we are interested in finding dynamic optimization rules to provide high performance. Finally, the management of emerging textual representation languages such as XML, Pdf and full text SQL languages has to be investigated. In this field, the Phasme SQL compiler is currently extended.

## References

[AKM94] Andrews,K.; Kappe,F.; Maurer,H., "The Hyper-G network Information System", *Information Processing and Management, Special Issue: selected proceedings of the Workshop on Distributed Multimedia Systems*, Graz, Austria, Nov. 1994.

[AA95] Andres,F.; Asada,K., "The BDM Benchmark: an example of use on G-BASE",

*IEICE'95*, Fukuoka, Japan, March 1995.

[AO97] Andres,F.; Ono,K., "Phasme: A High Performance Parallel Application-oriented DBMS" in *Informatica Journal, Special issue on Parallel and Distributed Database Systems*, 1997.

[Buford97] Buford,J.F.; Rutledge,L., "Third generation Distributed Hypermedia Systems" in *Multimedia Information Management Handbook* (ed. W. Grozky), Prentice Hall, 1997.

[CACS94] Christophides,V.; Abiteboul,S.; Cluet, S.; Scholl,M., "From Structured Documents to Novel Query Facilities", in *Proceedings on ACM SIGMOD International Conference on Management of Data*, pp.313-324, 1994.

[DeFazio95] DeFasio,S.; Daoud,A.; Smith,L.A.; Srinivasan,J., "Integrating IR and RDBMS using cooperative indexing" in *Proc. of the 18th Annual Int. SIGIR Conf. on Research and Development in Information Retrieval*, pp.84-92, 1995.

[Hara96] Hara,Y.; Hirata,K.; Takano,H.; Kawasaki,S., "Hypermedia Database" Himotoki "and Its Application" in *Proc. Int' l Conference on ICDE*, 1996, pp 372-379.

[HyT97] HyTime ISO/IEC 10744:1997.

[Kim96] Kim,H.; Zhoo,Z.; Shin H., and Chang J., "An Object-oriented Hypermedia System for Structured Documents" in *Proc. Pacific DBMS'96*, Hong Kong, 1996, pp.286-295.

[Kow98] Kowalski,Gerald., "Information Retrieval Systems Theory and Implementation", *Kluwer Academic Publishers*, second printing, 1998.

[ORA96] Network Computing Archictecture. An Oracle White Paper, September 1996.

[Ozsu97] Ozsu,M.T.; Iglinski,Paul.; Szafron, Duane.; El-Medani,Sherine.; Junghanns, Manuela., "An Object-oriented SGML/ Hytime Compliant Multimedia Database Management System", in *Proc. 5th International Multimedia Conference (ACM Multimedia 97)*, Seattle, WA, November 1997, pp.239-249.

[Volz96] Volz,M.; Aberer,K.; Bohm,K., "An OODBMS-IRS Coupling for Structured Documents" in *Proc. IEEE ICDE*, pp.10-19, 1996.

[Techno96] Technoteacher "Hyminder User Guide", June 1996.

## Appendix A:

```
DOCTYPE document SYSTEM "document.dtd"
<!- HyTime Modules Used →
<?HyTime support base>
<?HyTime support measure>
<?HyTime support sched manyaxes =3>
<?HyTime support hyperlink>

<!ATTLIST  document
  id ID          # REQUIRED
  HyTime NAME  # FIXED HyDoc>
<!ATTLIST quote
   source  CDATA  # IMPLIED>
<!ATTLIST author
         designation  CDATA # IMPLIED>

<!ELEMENT X  -- EMPTY>
<!ATTLIST X
 HyTime NAME # FIXED "axis"
 axismeas CDATA # FIXED "virspace"
 axisdim  CDATA # FIXED "1024">
<!ELEMENT Y  -- EMPTY>
<!ATTLIST Y
 HyTime NAME # FIXED "axis"
 axismeas CDATA # FIXED "virspace"
 axisdim  CDATA # FIXED "900">
<!ELEMENT  (audio | video | text)  -
     EMPTY!>
<!ATTLIST (audio | video | text)
  ebg CDATA # REQUIRED
  --- Hytime Attributed ---
  Hytime NAME  # FIXED "event"
 exspec IDREFS  # REQUIRED>
```

学術情報センター紀要　第10号(1998)

研究論文

# Striping and Transfer Alternation of VOD Data on Tape-Based Tertiary Storage Libraries

# テープベース3次記憶ライブラリー上のVODデータの異なるディスクへの分配と転送方策の検討

Jihad BOULOS
National Center for Science Information Systems
学術情報センター　ジハド　ブロス
Kinji ONO
National Center for Science Information Systems
学術情報センター　小野　欽司

**ABSTRACT**

Video-on-Demand (VOD) servers are becoming feasible. These servers have voluminous data to store and manage. If only disk-based secondary storage systems are used to store and manage this huge amount of data the system cost would be extensively high. A tape-based tertiary storage system seems to be a reasonable solution to lowering the cost of storage and management of this continuous data. However, the usage of a tertiary storage system to store large continuous data introduces several issues. These are mainly the replacement policy on disks, the decomposition and the placement of continuous data chunks on tapes, and the scheduling of multiple requests for materializing objects from tapes to disks. In this paper we address these issues and we propose solutions based on some heuristics we experimented in a simulator. We first extend a replacement policy that has been proposed for a single user environment to a multi-user one with several servicing streams. We then study different policies for continuous object decomposition and chunks placement on tapes under different characteristics of the tertiary storage drives. Finally, we propose a scheduling algorithm for object materialization; this algorithm guarantees the materialization on disks of all chunks of an object at their service deadlines in a pipelined service. We present the results of some simulations we made to measure the impacts of our proposed algorithms on the average latency time of the system.

**要旨**

ビデオーオンーディマンド(VOD)サーバーは、現実のものとなりつつある。これらのサーバーは、記憶及び管理すべき大量のデータを持っている。もし、ハードディスクベースの2次記憶システムだけでこの巨大な量のデータを記憶し管理するなら、システムコストは非常に高価なものとなろう。テープベース3次記憶システムは、この連続的データの記憶と管理のコストを下げる合理的な解決策になると思われる。しかし、大量の連続的データを3次記憶システムで記憶するには、いくつかの問題点がある。これらは、主にディスク上のリプレイスメントポリシー(再配置方策)で、テープ上にある連続データのかたまりの分解と配置、そしてテープからディスクへオブジェクトを実現する為の多様な要求のスケジューリングである。本論文ではこれらの問題点を扱い、シュミレーターで実験したいくつかのヒューリスティックスに基づく解決策を提案する。最初にシングルユーザー環境の為に提案されているリプレイスメントポリシーを一連のサービスストリームがあるマルチユーザー環境に拡大した。それから連続オブジェクトの為に3次記憶ドライブの異なった特質下にあるテープ上の交換とデータのかたまりの配置について異なったポリシーを検討する。最後にオブジェクトの具体化の為のスケジューリングアルゴリズムを提案する。こ

のアルゴリズムは、パイプラインサービスでサービス限界にあるオブジェクトの全データのかたまりのディスク上での具体化を保証する。

本論文では、我々の提案するアルゴリズムによるシステムの平均遅延時間の影響を測定したいくつかのシミレーション結果についても考察する。

[Keywords]　Continuous Media Servers, Replacement Policy, Scheduling, Hierarchical Storage Systems.

[キーワード]　連続メディアサーバ、再配置方策、スケジューリング、階層型記憶システム

## 1 Introduction

Rapid advancements in information technology have led to the development of consumer-targeted multimedia applications and services having voluminous-sized data requiring a large storage capacity within their underlying storage subsystem. Most multimedia data is currently stored on magnetic hard disks, yet the proliferation of such applications is generating massive amounts of data such that a storage subsystem based solely on hard disks will be too expensive to support and manage. This limitation is an underlying reason for developing hierarchical storage subsystems in databases [Ston91]. The video-on-demand (VOD) server is one particular multimedia application requiring huge storage capacity.

While current issues associated with VOD servers are focusing on the physical layout on disks of continuous objects (movies) such that the number of serviced streams can be maximized [BMGJ94, GVK+95, WYS95, ÖRS96, GÖS97, PER97], the number of available on-line objects will soon become an important issue as well. In other words, viewers must have the ability to request an object from a large collection and then be able to view it within a reasonable latency time. The problem arising is that some objects will only be requested infrequently, while others-especially new ones-may be simultaneously requested by any number of viewers. Because the cost per gigabyte using magnetic hard disks or tertiary storage systems is different by more than an order of magnitude, it is more economical to store all objects on a tertiary storage system and to replicate only the most requested ones at a certain time on disks. This is also beneficial for fault-tolerance -as a copy of any object exists on the tertiary storage- as well as for lowering the cost and physical size of the storage system.

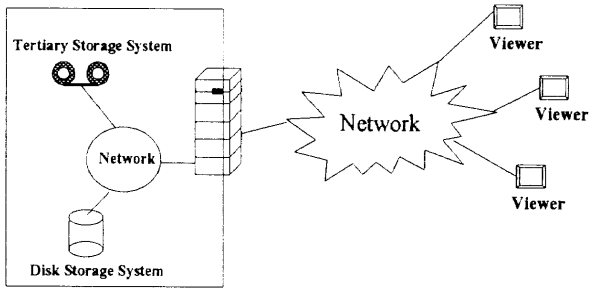



Figure 1　The rectangle contains the storage systems concerned by this study.

In the present study, we consider the issues of placing continuous objects on a tape-based tertiary storage system and also the ensuing replacement and scheduling policies (Figure 1). These issues have been studied previously, yet not from the standpoint considered here. That is, we focus on a VOD server which can **service multiple streams with a multitude of videos using a large number of hard disks and a tape-based tertiary storage system with several drives**. The size of the continuous objects is assumed to be several times larger than the secondary storage capacity, which is considered to be a realistic assumption. Video streaming companies and production studios are expected to manage a huge number of films, and in some situations, *e.g.*, production studios, a film cannot be compressed at production time during editing/assembly because it must process high quality

frames and lossy compression techniques are not convenient. Under such conditions and in light of the potentially large capacity of currently available tapes and the high transfer bandwidth of their drives[1], a tape library is considered to be advantageous in comparison to other kinds of tertiary storage systems.

With continuous objects stored on tapes it is envisioned that when a request arrives for a partially or completely non-disk resident object, the system starts transferring non-disk resident portions from the tertiary storage system to disks, and after some described conditions are satisfied the service commences while the transfer continues. Under these "pipelined" transfers, *i.e.*, between tapes and disks on one side and disks and memory on the other side, the service can commence prior to the end of the object transfer with minimum delays incurred. The envisioned system requires that several problems be addressed, however. Some examples are given to point out three major problems.

As an example of the first problem, if a request for a non-disk resident object arrives, space equal to the object size must be freed up on disks before the transfer from tape(s) can start. One solution proposed by [GS94] involves using a disk manager to compute a certain "heat" for each disk-resident object and remove parts or all of the least heated object(s) to free up the needed space. This replacement policy, however, was designed for a personal computer with a single user and not for a multi-user/multi-stream system in which a less heated object on disks may be undergoing servicing at the time a request arrives while other more heated objects might not be servicing. In an extreme case all the objects on disks could be undergoing servicing when a request for a non-disk resident object arrives, and therefore freeing up space under these conditions would require an extension to the replacement policy proposed by [GS94].

As an example of the second problem, if a tape library with a 3.0 MBps bandwidth drive is employed and two 1.0 GB requests with a display rate of 1.0 MBps each arrive within a short time, the second request must wait until the first finishes its materialization, *i.e.*, a latency time greater than 333 seconds. This gives an average latency time of more than 150 seconds and a large standard deviation in response time. If, however, transfer alternation is allowed, the first request can be interrupted after a certain time period and the transfer for the second can then commence. Since the drive transfer bandwidth is larger than the sum of the display rate of both requests, even with an overhead for seeking time and possibly tape exchanges, the average latency time for both requests might be in the range of a few tens of seconds with a reasonable standard deviation. The idea of alternating request transfers seems very simple and cost effective. If on the other hand, one or more other requests arrive in the meantime and again interrupt the transfer, this could create an interruption chain that prevents previously interrupted services from meeting their deadlines, *i.e.*, being materialized at the proper service time. This example indicates the need for a real-time-like scheduling of continuous object materialization from tertiary storage systems to disks.

As an example of the third problem, if a tape library has multiple drives, each of which has a transfer rate of 1.5 MBps, and a request arrives for a 1.0 GB continuous object with a display rate of 2.0 MBps. If the object is contiguously placed on one tape in equal-sized blocks (termed here as "chunks") and a pipelining technique is employed, the object display cannot start until some specified object size has been materialized on the disks. The latency time for this object is the time to load the tape, seek the location of the first object chunk, and materialize at least the first 250 MB [1000 - (1.5/2.0 × 1000) = 250 MB], *i.e.*, 166 seconds (250/1.5). If, on the other hand, the object chunks are alternately distributed on two tapes, two drives may be used to transfer them. The latency time in

[1]Tapes of more than 150 GB storage capacity and drives of 15 MBps transfer rate exist at the time of writing.

this case is the time to load the tapes, seek the desired location within the tapes, and materialize the first fraction of the first chunk of the object. Obviously then, distributing object chunks over several tapes may be beneficial.

Regarding the size of the chunks, the transfer of an object from tapes to disks must not be interrupted at a random position during alternating transfers between objects. For different tape technologies, physical block sizes are configured at the time the tape is formatted and transfers occur at the granularity of a physical block. This, however, would no doubt produce a large number of alternations between transfers; and accordingly, we chose to transfer chunks laying on a set of contiguous physical blocks. Needless to say, distributing object chunks on one or more tapes requires the system administrator to specify the chunk size, with this assessment greatly affecting the average latency time and the overall tertiary storage performance. That is, a large chunk size would yield a large latency time when alternating the transfers between two or more requested objects, whereas a small chunk size would probably yield a large number of alternations and hence consume much time in seeking chunk locations for the different requested objects and potential tape exchanges.

To address the above problems, we consider the following issues: 1) a replacement policy for freeing disk space to store non-disk resident chunks of a requested object, 2) a placement strategy for object chunks on one or more tapes, and 3) a real-time-like scheduling algorithm for materializing continuous objects from tapes.

In Section 2 related works are discussed, while in Section 3 we describe the system architecture and the framework under which the issues of this paper are relevant. Section 4 presents an extension to a multi-user environment of a replacement policy to free up disk space for continuous objects, and Section 5 considers the issue of chunk size and placement on tapes. An adapted real-time-like scheduling algorithm for transferring object chunks from tapes to disks is presented in Section 6. The results of simulations designed to clarify our choices are discussed in Section 7, with concluding remarks being given in Section 8.

## 2 Related Works

Several studies have addressed hierarchical storage systems for continuous or conventional data, although none have considered the general framework for continuous data management on a hierarchical storage system.

### Replacement Policy

As mentioned, [GS94] considered a PC-based hierarchical storage system for continuous media, proposing an algorithm for object replacement between the tertiary and the disk storage subsystems of PCs with one disk and one CD-ROM drive. In this framework, whenever the bandwidth of the tertiary drive is larger than the continuous data consumption rate, the data is transferred directly from the tertiary storage to the viewer. To address the issues introduced by a larger environment with parallel requests, we modify and extend this replacement policy.

[BR96] describes the hierarchical storage management system in Berkeley's distributed VOD system, where a continuous object must be requested hours before its service time and the object is completely removed from disks when space is needed. However, neither object decomposition nor service alternation are considered.

### Materialization Scheduling

[LLW95] describes a hierarchical storage system for a VOD server, proposing a scheduling algorithm for alternating data transfers in order to improve the response time and minimize required disk space. However, tertiary storage drives having a lower bandwidth than the service rate are not considered. The tape library is based on a single drive mainly, and transfer alternations between different requests are made according to a "time slice" which is an amount of time allocated for the transfer of a fraction of an object, though object decomposition and chunk placement are not considered. In addition, the analytical formulae do not include seeks,

which we will show account for a significant proportion of transfer costs especially when alternation is extensive.

[HS96] investigates the effectiveness of several scheduling algorithms for a batch of random I/Os on a serpentine DLT4000 tape in which scheduling algorithms with time constraints are considered. This work confirms the feasibility of accurate analytical modeling of tape drive behavior.

### Chunk Placement

The issue of data placement on tertiary storage systems was recently addressed by [CTZ97], where data blocks are placed according to an *a-priori* access pattern, *i.e.*, the most requested objects are placed together at the middle of a tertiary storage media, also referred to as a platter, in order to reduce exchange and seek times. While this method is plausible, it seems likely that access patterns for various kinds of data will change over time. This would definitely be the case for a VOD server where new objects are frequently added and are the most requested ones. The issue of multimedia data placement was also addressed by [TP97], where blocks of a continuous object are placed on a platter according to their access time within that object; the aim being to save on the secondary storage bandwidth by arranging blocks on platters in a manner that allows particular blocks to be transferred to memory at their service time. This optimization technique is not addressed here since we do not discuss secondary storage bandwidth.

### Chunk Size

[Sar95] and [YD96] have investigated how the performance of tertiary storage systems is affected by the size of chunks. [Sar95] performed an analytical analysis on the effect of "fragment" size on performance, where an upper bound of the fragment size is specified under a relational execution model incorporating scans and joins. Fragment size is related to the parameters of request size distribution and the kind/degree of sharing between queries -in addition to the underlying tertiary storage system parameters. These two parameters, however, are irrelevant under our framework as will be shown later. Instead, the consumption rate of continuous data and the ratio between the secondary storage capacity and the data size are the most important parameters. In the work by [YD96], an optimal range for the "tile" size of satellite images is analyzed and found to be between 32 and 512 KB for a Quantum DLT4000 tape drive. Their results, however, do not apply to continuous data in which chunk size must be in the order of tens or even hundreds of MB, *i.e.*, a continuous object consists of several GB of data compared to a typical satellite image of about a 100 MB.

### Striping

[DK93] and [GMW95] considered robotic storage libraries, analyzing the performance and trade-off of striping large files on several tapes. Both studies deal with non-continuous large data files in which no transfer alternation and synchronization is necessary under their framework, finding that striping is beneficial when tertiary drives are lightly loaded, though it has drawbacks when the drives are highly loaded. While our results on chunk distribution are consistent with theirs in some cases, due to synchronization aspects, chunk distribution in our framework is sometimes beneficial even for a highly loaded system.

## 3 System Architecture

Figure 2 shows the architecture of the envisioned system. A tape-based tertiary storage system and a multiple-disk secondary storage system are connected to a large server through a high-speed network. Transfer of an object from tapes to disks must pass through the server memory if the hardware and software connection interfaces (*e.g.*, SCSI) of different storage components do not permit direct transfer from tape drives to disk drives. Other connection types (*e.g.*, HiPPI) allow direct transfer from tape to disk drives. In both cases though, the considered tasks are managed by a server-based database management system.

Different types of tertiary storage systems exist. We previously studied a CD-ROM jukebox tertiary

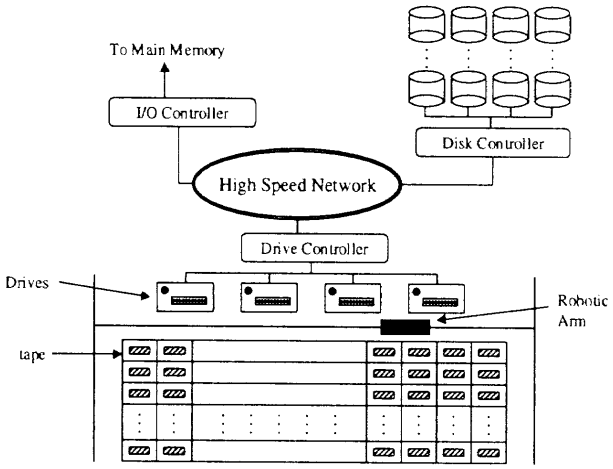



Figure 2 System Architecture.

storage system for a VOD server [BO97]. This system has a major drawback in that one platter has a limited storage capacity, and as such, since one continuous object is normally contained on several platters, a large platter exchange overhead is incurred. A tape-based tertiary storage system is therefore considered more convenient and economical. Two different technologies exist for currently available tapes, with their differences having been addressed by [HS96]; thus, we do not consider this issue here. In our simulations, we assumed that the seek time between two locations on a tape is proportional to their positions on that tape; an assumption which does not hold for serpentine tapes and therefore this is considered as the worst case seek time between two positions in the simulation. As shown in Figure 2, a robotic arm moves a requested tape to/from a drive from/to its cell in the library. We assume there is no contention for this arm, and considering the current state of robotic technology, this is a realistic assumption. Table 1 summarizes major parameters of the system architecture with their corresponding meanings.

The secondary storage system is based on an array of disks, and although several factors must be considered to determine the proportion of disk space versus data size, the most important factor is latency time versus cost. However, to justify the need of a tertiary storage system, the data size must be several times greater than the disk capacity. We define the ratio between the secondary storage capacity and the data size in the tertiary storage system as $STR$, *i.e.*, $STR=\frac{n_d \cdot c_d}{\sum_{i=1}^{O} L_i}$. Materializing object chunks is defined as replicating the chunks from the tertiary tapes to disks. It is surmised that disk partitions on which the continuous objects are stored are managed directly by the OS or a special purpose continuous media file system. This approach is necessary because the physical layout of continuous objects on disks is different from that of conventional data and continuous service must be guaranteed.

Table 1 Typical parameters used in our study.

| **Tertiary Storage Parameters** | |
|---|---|
| Number of Drives | $k$ |
| Tape Exchange Time | $t_{ex}$ |
| Seek Rate | $t_{tseek}$ |
| Read Rate per Drive | $r_t$ |
| Number of Tapes | $n_t$ |
| Tape Capacity | $c_t$ |
| Chunk Size | $B_t$ |

| **Disks Parameters** | |
|---|---|
| Read Transfer Rate | $r_d$ |
| Seek Time (worst case) | $t_{dseek}$ |
| Rotate Time (worst case) | $t_{lat}$ |
| Disk Capacity | $c_d$ |
| Block Size | $B_d$ |
| Number of Disks | $n_d$ |

| **Database Parameters** | |
|---|---|
| Initial Available Objects | $O$ |
| Regularly Added Objects | $A$ |
| Objects Length | $L_i$ |
| Display Rate | $D$ |
| Memory Block Size | $B_m$ |

The number of continuous objects $O$ stored on tapes in the tape library is large. We assume that all the objects have the same service rate, and that

at a certain time a subset of popular objects must be materialized on disks because of the larger transfer rate of disks and the possibility of multiple streams of the same object. When an object is requested, if it is completely materialized on disks, then the service may start immediately if sufficient disk bandwidth is available. If, on the other hand, some fraction of the requested object is not on disks, then the non-disk resident fraction is requested from the tape library. Of interest here are such cases, *i.e.*, requested objects are either partially or fully non-disk resident.

Whenever a request arrives, sufficient disk space must be freed up to materialize the non-disk resident fraction of the requested object, and once the free space is available, the transfer from tapes to disks may start. In parallel with the transfer, the service may start in a pipelined manner although some fraction of the object must be materialized on disks before starting the service. Adopting the definition from [GS94], the ratio between the transfer rate per tape drive and the consumption rate of an object is termed as the Production Consumption Ratio ($PCR=\frac{r_t}{D}$). If $PCR \geq 1$, then under their framework, the request may be serviced directly from the tertiary memory. Under ours, however, this criterion poses a big problem because a popular object could be requested many times within a short time; and if so, it is likely that the tertiary storage drive could not service all requests from multiple service points. Moreover, if $PCR > 1$, servicing a stream directly from the tertiary storage drive would force the drive to stop and resume its transfer several times such that drive bandwidth is wasted.

When a new request arrives for an object having non-disk resident chunks, the system cannot simultaneously (in parallel) send a request for freeing disk space and transferring the non-disk resident chunks since sufficient disk space must be guaranteed before starting the transfer of non-disk resident chunks. A parallel request may lead to a situation where the transfer from tapes is guaranteed at a certain time while the disk space is not guaranteed yet.

### 3.1 Object Decomposition

The continuous objects have different sizes. An object $O_i$ is split into $n_i$ chunks $O_{i,1}$, $O_{i,2}$, $\cdots$, $O_{i,n_i}$ which may be stored on just one tape in a contiguous manner, inter-lapped with chunks from other objects, or be distributed on more than one tape. A chunk is the unit of transfer from tapes to disks and of space freed up on disks.

Upon system installation, some number $O$ of continuous objects will be placed on a particular number of tapes. At either constant or variable time intervals (*e.g.*, one week), a number of new objects $A$ are added to the system and $O \leftarrow O+A$. These objects are initially added on tapes, and upon being requested, a new object is materialized on disks. It is expected that the request frequency of new objects will reach a peak level a short time after being added to the system.

### 3.2 Latency Time

The ultimate objective of the server is to minimize the latency time while guaranteeing an uninterrupted service stream. Latency time is defined as the elapsed time between request arrival and the beginning of service. For a new request of an object with non-disk resident chunks, the following conditions must be satisfied before service commencement.

- Sufficient disk space must be freed up to store non-disk resident chunks of the object.
- The scheduling algorithm of the tape library drives must guarantee the materialization on disks of any non-disk resident chunk before its service time.
- Some fraction of the object is disk resident.
- Sufficient disk bandwidth is available for parallel servicing the requested stream and materializing non-disk resident chunks.

Regarding the last condition, this is not an issue studied here, *i.e.*, we assume that if the disk bandwidth cannot satisfy a new request, the request is rejected and thus the service request is not han-

dled. Hence, the latency time $T_i$ for a request of object $O_i$ having non-disk resident chunks is calculated as

$$T_i = L_{FreeSpace} + L_{Schedule} + M_i \quad (1)$$

where the variables in (1) are described next.

$L_{FreeSpace}$ is the time needed to free up disk space equivalent to the non-disk resident fraction of $O_i$. If all disk-resident objects are servicing at the time a request for $O_i$ arrives, freeing disk space equivalent to the required space may not be possible and the request must wait until sufficient disk space is available. In this case $L_{FreeSpace}$ cannot be estimated in advance. If, however, the needed space can be immediately freed up, then $L_{FreeSpace}$ is assumed to be equal to zero, *i.e.*, the time for disk space management is not factored in.

$L_{Schedule}$ is the time needed by the tertiary storage scheduling algorithm to place the request for $O_i$ in a list of requests and the elapsed time between the end of the scheduling algorithm and the arrival of the list element carrying request $O_i$ to the drive handling it. If the scheduling algorithm is FCFS or has a low complexity, then the execution time of the scheduling algorithm is insignificant, *i.e.*, considered to be zero. If the request list is empty upon arrival of request $O_i$, then it is placed at the top of the list and its elapsed time for a drive is zero. In this case, the scheduling cost is also insignificant, as there is only one request to schedule.

$M_i$ is the time to materialize the $FirstSlice(O_i)$, which is the fraction of object $O_i$ that must be materialized on disks before service commencement, being related to the value of $PCR$, *i.e.*,

$$FirstSlice(O_i) = L_i - (\min(PCR, 1) \times L_i) + B_m \quad (2)$$

where $\min(PCR, 1)$ expresses the fact that when $PCR \geq 1$, the size of any $FirstSlice(O_i)$ is equal to the size of $B_m$, independent from $L_i$. Accordingly, $M_i$ can be computed using

$$M_i = P(t_{ex}) + StartPosition(O_{j,l}) / t_{tseek} + FirstSlice(O_i) / r_t \quad (3)$$

*where* $P(t_{ex})$ is a function that returns the cost in seconds of a tape exchange if needed. Also included in this parameter is (i) the total time for rewinding the currently loaded tape in the target drive to a position where eject is allowed and (ii) the time for the robot arm to shelve the tape and get/load the requested tape.

For any object $O_i$ on the tertiary storage system, if a portion $O_{i,p_1}$ of that object is to be conserved on disks to eliminate $M_i$ this portion must be greater or equal to $DiskSlice(O_i)$ defined as:

$$DiskSlice(O_i) = (t_{ex} + StartPosition(O_{i,p_2}) / t_{tseek}) \times r_t + FirstSlice(O_i) \quad (4)$$

where $t_{ex}$ is (i) the time to rewind a tape such that it can be ejected (worst case), unloaded, and placed back in its cell, and (ii) the time to get a new tape and load it in a drive. $StartPosition(O_{i,p_2})$ is the start position on the tape of the second portion of $O_i$.

### 3.3 Multiple Drives

The tape library is comprised of $k$ drives ($k > 1$) in which all are assumed to have the same characteristics. The transfer bandwidth of the tape library becomes $k.r_t$. The following three cases apply regarding pipelining and chunks placement on tapes:

- $k.PCR < 1$.
- $PCR < 1$ and $k.PCR \geq 1$.
- $PCR \geq 1$.

These three cases are the initial points from which the design of a tape-based tertiary storage system for continuous data must start, the cost of which is primarily determined by the number of drives/tapes comprising it. One important question that must be addressed is: "What is the optimal number of drives in a tertiary storage subsystem such that it is nearly always available for servicing a transfer request yet at the same time economical?" The answer is heavily dependent on $PCR$, $STR$, and the object access patterns -as well as drive characteristics. A highly skewed access pattern requires that a small number

of objects be materialized on disks in a certain length of time, while a more uniform access pattern requires a higher transfer rate to satisfy requests in the same amount of time.

We determine the optimal number of drives, $k$, as follows. Let $\lambda$ be the average number of requested objects to be transferred from the tertiary storage subsystem to the disks during some $\Delta t$ (*e.g.*, 30 min); $n_{A_i}$ be the number of tapes on which the ith requested object is stored ($n_{A_i} \geq 1$); and $L_i$ be the length in MB of object $O_i$. To reach equilibrium of the tertiary storage, the following formula must hold:

$$\frac{1}{k} \cdot \sum_{i=1}^{\lambda} \left\lceil \frac{L_i}{r_t} + n_{A_i} \cdot \sum_{j=1}^{O_{i,n_i}} \left( P(t_{ex_j}) + Abs(StartPosition(O_{i,n_i}) - EndPosition(O_{i,n_{i-1}}))/t_{tseek} \right) \right\rceil \leq \Delta t \qquad (5)$$

that gives $$k = \left\lceil \frac{1}{\Delta t} \cdot \sum_{i=1}^{\lambda} \left\lceil \frac{L_i}{r_t} + n_{A_i} \cdot \sum_{j=1}^{O_{i,n_i}} \left( P(t_{ex_j}) + Abs(StartPosition(O_{i,n_i}) - EndPosition(O_{i,n_{i-1}}))/t_{tseek} \right) \right\rceil \right\rceil$$

where $P(t_{ex_j})$ is the function that returns the tape exchange time when chunk $O_{i,n_j}$ is the next to be transferred, with the other variables being self explanatory. If no transfer alternation is made between the transfers of two contiguous chunks of the same object, then $P(t_{ex_j})$ and the seek time are both equal to zero. In other words, suppose 10 continuous objects of 1.125 GB each (the MPEG-I 1.5 Mbps bit stream for 100 min) are to be transferred from the tertiary storage subsystem to hard disks in 30 minutes. Suppose also that each object is stored on 1 tape and decomposed into 6 contiguous chunks that are placed at the middle of a 20 GB capacity tape. The transfer rate of one drive is 6.0 MBps, its load time is 50 seconds, and its seek speed is 200 MBps. Under these conditions, we obtain $k=2$, whereas with everything the same and 3.0 GB continuous objects (the NTSC 4.0 Mbps bit stream for 100 minutes), $k$ must be no less than 4.

## 4 Replacement Policy

Replacement policy is concerned with freeing space on disks for materializing an object that has been requested and is partial or fully non-disk resident. An unlimited number of drives in the tape library and $PCR \geq 1$ are assumed here. In other words, whenever a request for an object is sent to the tape library there is always sufficient free tape drives and for all objects $O_i$ in $O$ $FirstSlice(O_i) = B_m$ holds. Under these assumptions, the locations of the chunks of any object and chunk size are not an issue; hence, $L_{Schedule}$ is equal to zero and $M_i$ is bounded by a constant.

[GS94] has proposed a replacement policy (PIRATE) that is suitable for a single stream environment and is needed only when $PCR < 1$. We extend this replacement policy to a multiple-stream environment. In our framework this replacement policy must be applied independent from the value of $PCR$ (*i.e.*, for both $PCR < 1$ and $PCR \geq 1$.)

The system computes a heat $Heat(O_i)$ for each continuous object $O_i$ for the last P period (*e.g.*, one week), being calculated as the ratio of the number of requests that $O_i$ has received during $P$ to that of all the requests received by the system also during $P$, *i.e.*, $\sum_{i=1}^{0} Heat(O_i) = 1$.

Figure 3 shows the extended replacement policy PIRATE', An extended version of the algorithm Ext-PIRATE was optimized to balance the latency time with its standard deviation by deleting a certain computed amount of the $FirstSlice(O_i)$ of an object $O_i$. This optimization was not considered in our study because we are trying here to devise a way that avoids the blocking of arriving requests. The replacement policy algorithm commences by deleting chunks of objects which have the lowest heat values and are not undergoing servicing, deleting chunks backwards from the end to the beginning of an object. For a non-servicing victim, $O_i$, $O_i$ chunks coming after the $DiskSlice(O_i)$ are deleted first. If more space is still required, deletion similarly proceeds to chunks falling within the $DiskSlice(O_i)$ until the first chunk of that victim is reached. Deleted last are those chunks from servicing objects

```
Procedure PIRATE' (RequestedObject) {
    NeededSpace = RequestedObject.Length - RequestedObject.OnDisks
    Repeat
        victim = object Oi on disks such that
            1) Oi is not servicing
            2) Oi has the lowest heat
            3) OnDisks(Oi) > DiskSlice(Oi)
        If (victim is NOT null) then
            FreedSpace = OnDisks(Oi) - DiskSlice(Oi)        // in the granularity of a chunk
        Else
            victim = object Oi on disks such that
                1) Oi is not servicing
                2) Oi has the lowest heat
            If (victim is NOT null) then
                FreedSpace = OnDisks(Oi)
            Else
                victim = object Oi on disks such that
                    1) Oi has the lowest heat
                If (victim is NOT null) then
                    FreedSpace = ServicePoint(Oi) - DiskSlice(Oi)
        If (FreedSpace > NeededSpace) then
            OnDisks(Oi) = OnDisks(Oi) - NeededSpace         // in the granularity of a chunk
            NeededSpace = 0
        Else
            OnDisks(Oi) = OnDisks(Oi) - FreedSpace
            NeededSpace = NeededSpace - FreedSpace
    Until (NeededSpace = 0 or one loop is made without freeing any chunk)
    If (NeededSpace > 0) then
        send a wait signal
}
```

Figure 3 Replacement Policy.

and falling after $DiskSlice(O_i)$ and before the chunk in which the service is currently proceeding. It is this last deleted portion which signifies the algorithm's major modification. If at that point sufficient space is not freed up, then a signal is transmitted to the request telling it to wait for some unspecified amount of time until a stream finishes its service such that waiting requests can now be handled.

The extended algorithm is compared to the original version in Section 7. In the next two sections we consider a number of disks $n_d$ always sufficiently large in order to avoid blocking any request and hence preventing $L_{FreeSpace}$ from influencing the latency time of the other algorithms used for comparison.

## 5 Chunks Size and Distribution

As distribution of continuous data chunks on multiple tapes can be beneficial, we evaluated several different placement strategies in simulation experiments, comparing only the most effective strategies regarding selection of chunk size/distribution in Section 7. From the results, we discuss the potential gains in latency time when using striping and/or overlapping.

At system installation, a number of objects $O$ will be placed on tapes, and due to tape capacity being typically several times larger than object size, several objects will normally be placed on one tape. In an iterative manner, a small number of objects $A$ is chosen from the initial number $O$ to be

placed on $n_A$ tapes, where $n_A$ is termed the stripe size. Parameters $A$ and $n_A$ are determined according to $PCR$ and $\sum_{i=1}^{A} L_i \leq n_A . c_t$, although a good rule-of-thumb is that $n_A \leq k$.

We define the *Utilization Factor UF* of a drive to be the fraction of drive time spent in transferring data, *i.e.*, one minus the fraction of time it is idle, exchanging tapes, or seeking locations in tapes. The utilization factor of the tertiary system is $k.UF$, and the objective of any data placement strategy and scheduling algorithm is to maximize $k.UF$, which is equivalent to minimizing the average latency time for all requests. This last point is considered obvious and therefore we present no formal proof.

Although striping an object on $k$ tapes yields a good $k.UF$ since all drives are transferring data when a request is made, this also consumes time in exchanging tapes and seeking locations, *i.e.*, $k.UF$ is lowered. Therefore, these two contradicting outcomes of striping must be balanced. One successful heuristic method is to distribute only a portion of an object's chunks. That is, a *Distribution Factor DF* gives the number of chunks from each object in $A$ to be distributed and overlapped with other objects chunks on $n_A$, with the remaining chunks of an object being stored in a contiguous manner on a single tape. Parameters $DF$ and $n_A$ are accordingly based on $PCR$. We denote the size of a chunk as $B_t$, whose optimal value, *i.e.*, yielding the best latency time, also varies according to $PCR$. It is important to note, however, that optimal $B_t$ under a light workload may have a poorer latency time than other chunk sizes under a heavy workload; hence, $B_t$ is also dependent on system workload.

### 5.1 k.PCR<1

When $k.PCR<1$, distributing chunks on $n_A=k$ tapes showed a significant improvement in latency time when the tertiary system is lightly to moderately loaded, being due to the fact that *FirstSlice* $(O_i)$ is very large in this case. Decomposing *FirstSlice* $(O_i)$ and distributing its chunks on multiple tapes allows parallel transfer of the chunks and hence lowers $M_i$ in Eq. (3) since $r_t$ is multiplied by $k$. An upper bound on the chunk size in this situation is $\lceil B_t = FirstSlice(O_i)/k \rceil$. Theoretically, $M_i$ is divided by $k$, with $B_t = FirstSlice(O_i)/2.k$ showing the lowest average $M_i$ since grouping and overlapping the first chunks from different objects on a single tape provides a better average seek time for the first chunks of all objects whose first chunks are placed near the head of a tape. This also provides a balance against the cost of initiating a read, *i.e.*, lowering the number transfer requests. No alternation of transfers is possible in this case. Figure 4 shows an example in which three objects are distributed on two tapes in which their first chunks are overlapped in a round-robin manner.

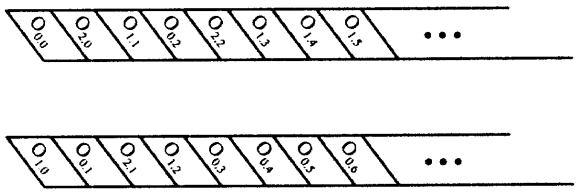



Figure 4　Distribution of three continuous objects on two tapes.

The optimal $DF$ under $k.PCR<1$ was determined as the average size of *FirstSlice* $(O_i)$ divided by the size of a chunk, *i.e.*,.

$DF = \dfrac{\sum_{i=1}^{0} FirstSlice(O_i)}{O \cdot B_t}$. To guarantee a distribution of an object's $DF$ chunks on all $n_A$, the number of added objects $A$ and the number of tapes $n_A$ used to store them must have a GCD (Grand Common Divider) equal to 1; and if $A$ is a prime number the condition is satisfied.

### 5.2 PCR<1 and k.PCR>1

If $PCR<1$ and $k.PCR \geq 1$, this situation is not much different from $k.PCR<1$. The main difference being that $n_A$ might be less than $k$, and in such a case, the lowest average latency occurs when 1) $n_A$ is minimal and 2) $n_A.PCR \geq 1$. In other words, a minimum number of tape exchanges and seeks is

balanced against reducing the size of $FirstSlice(O_i)$ of all objects in $O$. Accordingly, $DiskSlice(O_i)$ of each object is no longer dependent on the particular size $L_i$ of that object, rather it depends on the value of $B_t$; hence, $FirstSlice(O_i)$ becomes $FirstSlice(O_i) = B_t - (B_t \times PCR) + B_m$ which is independent from $L_i$ and is the same for all objects. An upper bound on the size of a chunk in this case is $\lceil B_t = FirstSlice(O_i)/n_A \rceil$. In our measurements $B_t = FirstSlice(O_i)/n_A$ provided the lowest average $M_i$. Once again no alternation of transfers on the same drive can be performed here. The optimal $DF$ was its minimum value, *i.e.*, $DF = n_A$, which provides a lower number of tape exchanges and seeks and prevents a request from utilizing a drive more than it needs; thereby preventing any delay in satisfying subsequent requests.

### 5.3 PCR>1

When $PCR > 1$, striping was only advantageous when the system was lightly loaded, though it yielded catastrophic consequences when the system was moderately to highly loaded. This outcome is due to the fact that the read rate is relatively high and the ensuing tape exchange and seek times account for a significant cost relative to the read time. Hence $n_A$ is best when no striping is permitted, *i.e.*, $n_A = 1$. In this case $FirstSlice(O_i) = B_m$ is independent of the size of any object.

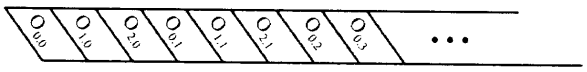



Figure 5 Distribution of three continuous objects on one tape.

An alternative to striping large objects on multiple tapes when $PCR > 1$ is to decompose an object into several chunks and overlap different chunks from different objects on the same tape (Figure 5). The first chunks of all objects on one tape are placed near the head of the tape, which lowers the average seek time for the first $B_t$ of all objects.

$B_t$ must have a greater consumption time than the time for rewinding and switching a tape, seeking and transferring another $B_t'$, and again rewinding and switching a tape and seeking the start position of a third $B_t''$. This case can be formulated as follows:

$$B_t \geq (t_{ex} + StartPosition(B_t')/t_{tseek} + B_t'/r_t + t_{ex} + StartPosition(B_t'')/t_{tseek}) \times D \quad (6)$$

where the switch and seek times must be taken as the worst case. $B_t$ may be equal to $L_i$, and if so, every object $O_i$ consists of one chunk and no alternation is subsequently allowed, although as mentioned alternation can be beneficial. In this case, $DF$ of an object becomes the number of chunks of an object, *i.e.*, an object is decomposed into $DF$ chunks independent of its size while satisfying Eq. (6). Simulations showed that $DF = 4$-$6$ gives nearly optimal average latency. Accordingly, each $O_i$ has its own $B_t$ such that $B_t(O_i) = L_i/DF$; although this complicates the management of space on tapes. From simulations we found one solution is to set $B_t$ of all $O_i$ to the same value, being the average $L_i$ divided by $DF$, *i.e.*, $B_t = \frac{\sum_{i=1}^{0} L_i}{O \cdot DF}$

## 6 Transfer Scheduling

Scheduling the transfers of continuous objects from a tertiary storage system with a pipelined service requires a different strategy from scheduling random I/Os. While this issue has several similarities with real-time scheduling problems, it cannot be stated as such. In real-time scheduling algorithms context switching cost is assumed to be negligible or constant [LL73, GP96]. Under our framework, however, tape exchange and seek costs are important parameters as they significantly affect the utilization factor and system performance accordingly; and in fact, extensive switching between different transfers may have disastrous consequences on performance. Moreover, real-time systems have tasks with periodic or aperiodic activities that cannot be executed before being requested, which does not apply to our framework where the transfer of a chunk may commence long

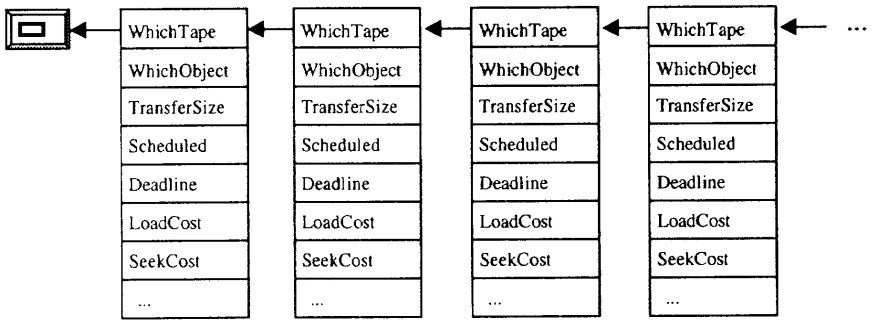



Figure 6　Some parameters of the data structure of each task in the drive task list.

before other preceding chunks from the same object are consumed.

The two cases do though have an obvious similarity in that they both impose a firm-time constraint on completing their jobs before some deadline. Direct adaptation of real-time scheduling algorithms [LL73] would, however, be too restrictive for admission control of a new transfer request because exchanges and seeks must be computed as worst cases.

The problem of scheduling transfer requests from tapes to disks has already been shown to be complex [Sar95, HS96], and this could be even more complex for transfer requests with deadlines, although such complexity can be alleviated by the use of heuristics. To improve the latency time of the system under different conditions, we experimentally investigated several different heuristics; finding that some scheduling choices perform well under a particular chunk distribution/system workload but demonstrate catastrophic performance under other circumstances. We observed that *PCR* and *STR* are in most cases the decisive parameters when considering resultant quality of the implemented heuristic algorithm.

For admission control and transfer scheduling, the system manager processes requests as follows. At the arrival of a request, the system makes a list of the non-disk resident chunks and transmits it to the scheduler. This list consists of elements, termed as tasks hereafter, that specify each requested chunk, on which tape it is stored, its position on that tape, its position within the object, and several other parameters. Figure 6 depicts a list of tasks waiting to be serviced at a drive, while Figure 7 presents the algorithms that the scheduler uses to insert in the lists of different drives new tasks without compromising the deadline of any task already in the drive lists. With this policy, the scheduler reports immediately $L_{Scheduat}$ for that object, and administrators or users may cancel a request if they are not willing to wait for the reported latency.

The two following heuristics showed valuable applicability:

- Balancing workload between drives (obvious optimization practice).
- Never migrates a tape from one drive to another. This implies that when a task requests a chunk from tape $t_i$ and this tape is already scheduled on drive $d_i$, even if there is another less loaded drive $d_j$, then $t_i$ is forced on $d_i$ (higher priority than the first point; reduces tape exchanges).

The algorithm is self-explanatory. Its implementation counts for one-third of the simulator and it took the largest amount of work to be enhanced and verified. We present its results in the following section.

## 7　Simulation Results

The majority of the discussions and results given thus far have been based on the results of a simulated system. That is, to better understand tertiary storage system behavior under different workloads, we conducted extensive simulations on a model of

```
Procedure TransferSchedule(List of Requested Chunks for Object O_i) {
    Map the list of requested chunks to a list of tasks
    For the first task Task_i,1
        Find the least active drive
        PlaceFirstTask(Task_i,1) on the list of tasks of the chosen drive
    Foreach other task Task_i,j from the rest of the list
        Task_i,j.Deadline = Task_i,1.Deadline + Chunk_i,j.PositionInObjcet_i / D
    Foreach other task Task_i,j from the rest of the list
        Choose a drive
        PlaceOtherTasks(Task_i,j) on the list of tasks of the chosen drive
        If Task_i,j is placed after its deadline by s seconds then
            Delay the deadline of all the tasks from this request by s
    GuaranteedService = task_i,1.Deadline
}

Procedure PlaceFirstTask(Task_i,1) {
    If drive task list is empty then
        Place Task_i,1 on head of list
        Task_i,1.Service = CurrentTime()
        Task_i,1.Deadline = CurrentTime()
    Else
        Place Task_i,1 as soon as possible while guaranteeing that for each task Task_di in the drive task
          list coming after Task_i,1 Task_di.Deadline - Task_di.Service is greater or equal to the needed
          time to do all the necessary tape exchanges, seeks, and Task_i,1.ChunkSize reading.
        Adjust the service time of all that drive tasks Task_di coming after Task_i,1.
        Task_i,1.Service = the sum of the service time of all drive tasks list preceding Task_i,1
                + all the exchanges and seeks costs
                - already serviced time for the first task in this list
        Task_i,1.Deadline = Task_i,1.Service
}

Procedure PlaceOtherTasks(Task_i,j) {
    If drive task list is empty then
        Place Task_i,j on head of list
        Task_i,j.Service = CurrentTime()
    Else
        Place Task_i,j in the list of its chosen drive with the Earliest Deadline First policy while guaranteeing that
          for each task Task_di in the drive tasks list coming after Task_i,j Task_di.Deadline - Task_di.Serviced is greater
          than the needed time to do all the necessary tape exchanges, seeks, and Task_i,j.ChunkSize reading.
        Adjust the service time of all the drive tasks Task_di coming after Task_i,j.
        Task_i,j.Service = the service time of task Task_dj immediately preceding Task_i,j in the drive list
                + Task_dj service length + all the exchanges and seeks costs
    If (Task_i,j.Service > Task_i,j.DeadLine)
        Signal a delay of s = Task_i,j.Service - Task_i,j.DeadLine
}
```

Figure 7 Scheduling Algorithms showing suitability for task placement.

the proposed system architecture implemented in a combination of C and C++ (about 2,000 lines) as part of a simulation program integrated into the CSIM18 [CSIM94] simulation package. As full

Table 2　Fixed values used in system performance measurements/simulations.

| Disks Parameters | |
|---|---|
| $r_d$ | 6.0 MBps |
| $t_{dseek}$ | 15 ms |
| $t_{tat}$ | 8ms |
| $c_r$ | 4.55 GB |
| $B_d$ | 0.256 MB |
| $B_m$ | 0.256 KB |

Table 3　Summary in parameter values used in four simulated scenarios.

| | First Scenario | Second Scenario | Third Scenario | Fourth Scenario |
|---|---|---|---|---|
| $t_{ex}$ | 40 sec. | 100 sec. | 70 sec. | 30 sec. |
| $t_{tseek}$ | 300 MBps | 50 MBps | 150 MBps | 500 MBps |
| $r_t$ | 6.0 MBps | 0.5 MBps | 2.0 MBps | 15.0 MBps |
| $c_t$ | 50 GB | 5 GB | 50 GB | 165 GB |
| $B_t$ | 500 MB | 200 MB | 1.2 GB | 7.5 GB |
| $k$ | ∞(40) | 4 | 4 | 4 |
| $n_d$ | 50 | 20 | 100/200/400 | 500 |
| $D$ | 10.0 Mbps | 24.0 Mbps | 24.0 Mbps | 48.0 Mbps |
| $A$ | 7 | 7 | 9 | 19 |
| $O$ | 504 | 504 | 504 | 513 |
| $L_i/D$ | 30～120 min. | 6～15 min. | 30～60 min. | 68～84 min |

control on the lists and their member entries is required, and CSIM18 does not permit such control, we implemented our own queuing system for the tape drive lists. The simulator is comprised of three components: the request generator, tertiary storage manager, and disk manager.

All the simulations were executed on a bi-processor Sun-Ultra2 workstation running Solaris 2.5.1 with 256 MB of RAM. Simulation results are given for 5,000 requests, although no statistics were taken until 1,000 requests were processed in order to prevent the statistics from being biased by the initial state of the system and such that only statistics for a heated system were reported. Simulations lasted on average from 1 to 15 minutes, which allows evaluation of the complexity of the scheduling algorithm. Data structures and lists consumed between 3 and 50 MB of RAM, with large memory consumption only occurring when objects were decomposed into a large number of chunks and the system workload was pushed to its limit; a situation requiring large data structures at the beginning of a simulation and additional memory allocation for the lists consisting of several hundred elements. One lesson learned is never to divide an object into more than ten chunks.

The request generator is an open system with the inter-arrival time between requests being distributed according to an exponential time distribution with a varied mean. We made simulations using two access patterns to the objects -uniform and Zipf -like (*i.e.*, skewed)- being fairly representative distributions of real workloads such that system behavior could be analyzed under two extreme access patterns. The skewed access pattern had its moving peak at the new added objects to the sys-

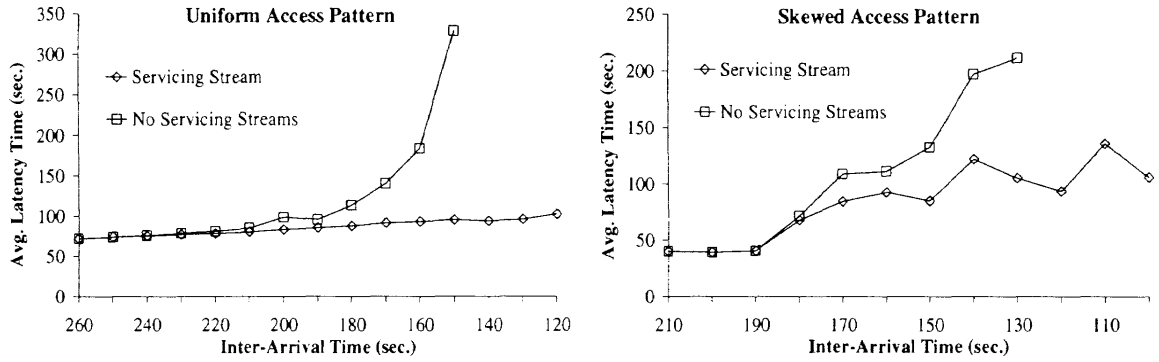



Figure 8 Scenario 1 results: average latency (waiting) times for disk space when chunks from the servicing streams could be removed (service stream) versus not removed (no service stream) for a (a) uniform and (b) skewed access pattern.

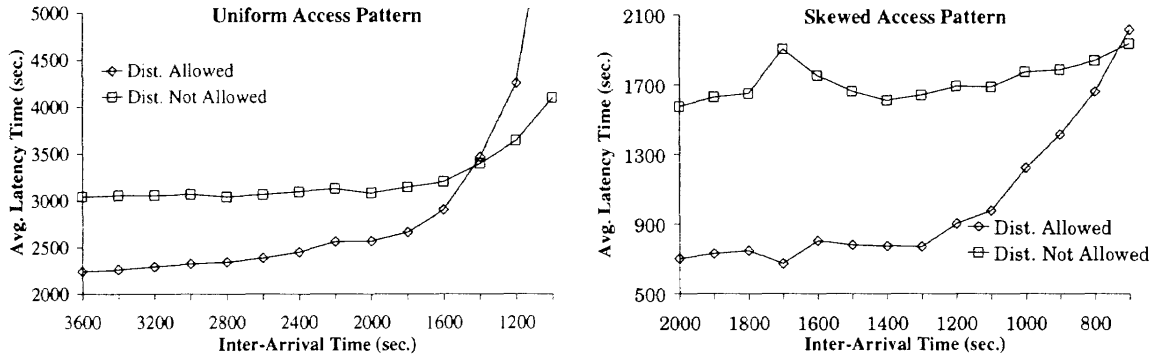



Figure 9 Scenario 2 results: average latency times for disk space when striping was allowed (dist. allowed) versus single tape placement (dist. not allowed) with $k$. $PCR<1$ for a (a) uniform and (b) skewed access pattern.

tem.

The following subsections present simulation results and analyze system behavior under various conditions. All simulations were repeated and parameters were varied several times in order to ensure reproducible results and to exclude the possibility of special cases being a factor in our observations. Tables 2 and 3 respectively summarize the values of fixed parameters and those varied in four simulated scenarios. Parameter $t_{ex}$ accounts for the time to unload a tape, place it back in its cell, get a new tape, and load it in a drive, while the time period for different objects is uniformly distributed between the sizes given in line $L_t/D$. Although simulations were carried out using commercially available video tape libraries, we only report the results for hypothetical tape libraries elucidating the behavior of the proposed algorithms under extreme conditions.

### 7.1 Replacement Policy

Scenario 1 (Table 3) was run to analyze differences in waiting time for disk space when chunks from the servicing streams are allowed to be removed or not, *i.e.*, those chunks falling between $DiskSlice(O_i)$ of a servicing object and the chunk in which the service is currently proceeding. Figure 8 shows results for this replacement policy, where the uniform and skewed access patterns both indicate better performance using our modified algorithm allowing chunks from servicing streams to be removed. End points are inter-arrival times where requests could no longer free up space. The skewed access pattern indicates variations in the average

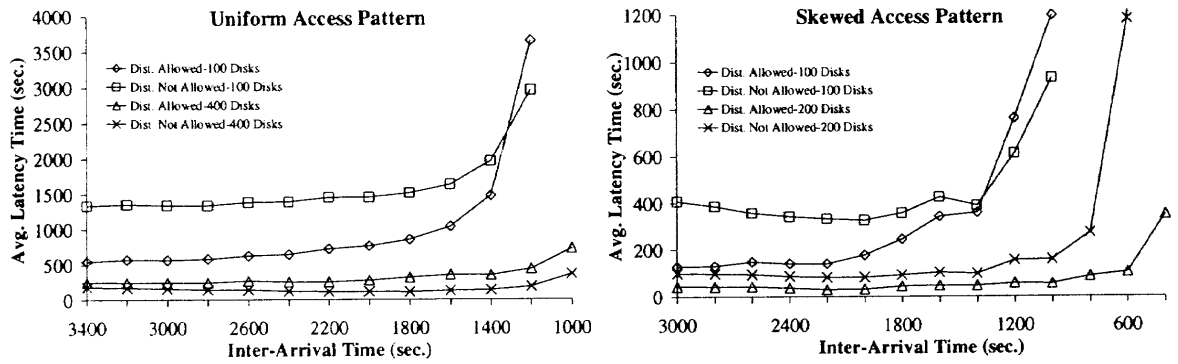



Figure 10　Average latency times under different conditions when $PCR<1$ and $k$. $PCR>1$.

latency time at times when new objects were added to the tape library, *i.e.*, the new objects affected the needed disk space immediately after being added to the system due to their high popularity and the large number of requests for them. Such behavior does not occur with the uniform access pattern because new objects are requested at the same frequency as old ones.

### 7.2 k.PCR<1

Scenario 2 considered $k.PCR<1$ and $STR\approx 0.1$ in which an object has an average size of 1.89 GB and can be transferred by one drive in 3780 s. $B_t$ was varied, and a value of 200 MB was found to be most effective. Figure 9 shows the results, where for both access patterns the distribution of chunks on 4 tapes ($n_A=k=4$) with $DF=9$ yielded lower average latency times when the system was lightly

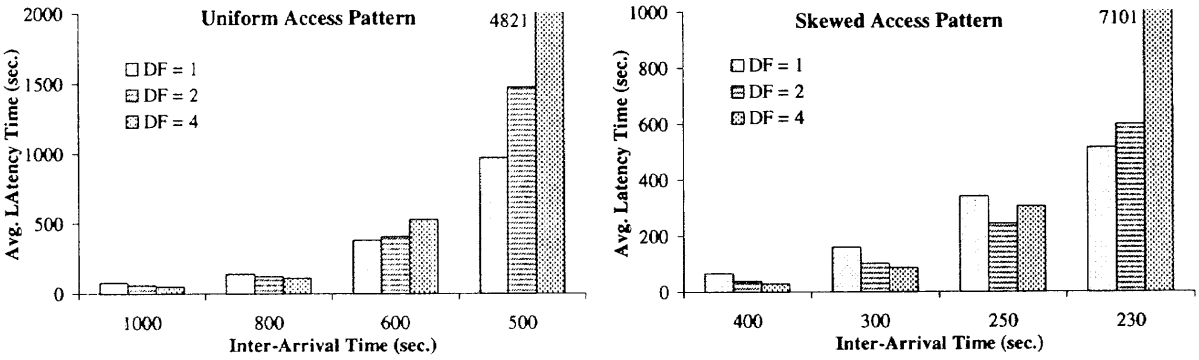



Figure 11　Distribution within the same tape when $PCR \geq 1$.

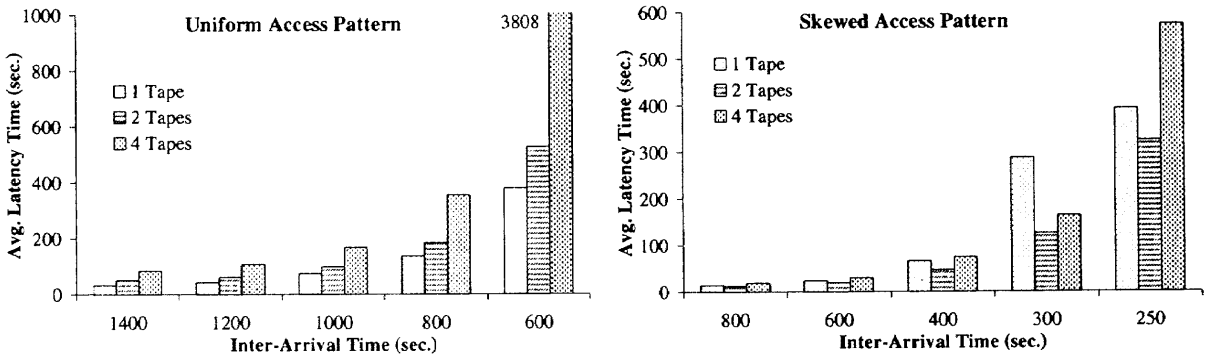



Figure 12　Distribution on different tapes when $PCR \geq 1$.

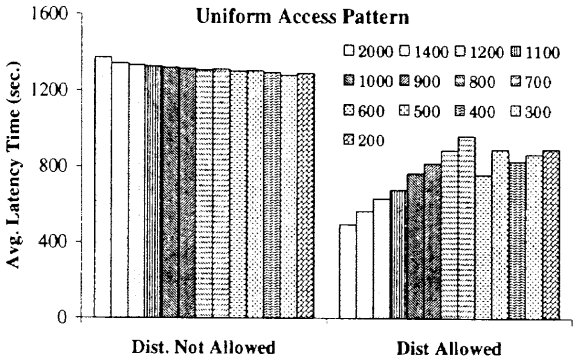



Figure 13a Varied chunk size on one and two tapes for the Third Scenario.

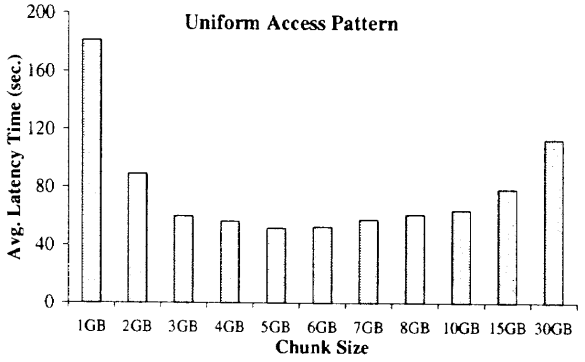



Figure 13b Varied chunk size on one tape for the Fourth Scenario.

loaded. For a highly loaded system, however, increases in tape exchange/seek times overcame the gains from parallel transfers on multiple drives.

### 7.3 PCR<1 and k.PCR≥1

Scenario 3 considered the case of $PCR<1$ and $k.PCR\geq 1$, with results being shown in Figure 10. This scenario is not as simple to analyze as that of 1 and 2. Regarding both access patterns with $STR \approx 0.1$ (*i.e.*, 100 disks), note that disk capacity was smaller than the sum of all the $DiskSlise(O_i)$, $n_d \cdot c_d < \sum_{i=1}^{0} DiskSlice(O_i)$ such that striping on two tapes (*i.e.*, $n_A=2$) improved the average latency time when the system was lightly loaded, although this gain was reversed by increasing exchange/seek times for a highly loaded system. However, for the uniform access pattern with $STR \approx 0.5$ (*i.e.*, 400 disks), disk capacity became greater than the sum of all the $DiskSlice(O_i)$, $n_d \cdot c_d > \sum_{i=1}^{0} DiskSlice(O_i)$, such that striping yielded a higher average latency time. For the skewed access pattern, when the number of disks was increased to 200, object striping still had an advantage over non-striping. With 400 disks, the results were similar but latency times were too small to be apparent. Such behavior is due to the fact that drive load balancing could not be achieved when a tape had a high probability of being requested several times in parallel (new films added on the same tape). With the no striping policy, this put a heavy load on one drive while other drives were idle.

### 7.4 PCR≥1

When $PCR\geq 1$, chunk distribution only shows an advantage in rare cases, typically having a counter-productive effect on performance. Scenario 4 considers this case in which $PCR\geq 1$ and $STR \approx 0.16$. Figures 11 and 12 respectively show the results when chunks were distributed on the same tape and on several tapes.

In Figure 11, average latency time is shown for $n_A=1$ and $DF=1$, 2, and 4. Note that for both access patterns object decomposition and overlapping only marginally improve latency time when the system is lightly loaded, and completely overloading it at moderate loads due to tape exchange/seek times. Figure 12 shows corresponding results when an object is distributed on 1, 2, and 4 tapes, *i.e.*, $n_A=1 \mid 2 \mid 4$. In this case, the uniform access distribution on multiple tapes shows higher latency times than no distribution; while the skewed access with distribution on 2 tapes shows a small advantage due to load balancing effects on drive utilization. In any case, however, distribution on multiple tapes with $PCR\geq 1$ clearly yields a negative effect.

### 7.5 Chunk Size

Figures 13.a and 13.b show the variations in average latency time when chunk size was varied for Scenarios 3 and 4, respectively. In Scenario 3, chunk size ranged from 200 MB to 2 GB with and

without distribution at a fixed inter-arrival time of 3000 seconds. Note that chunk size has almost no effect on the average latency time without distribution, whereas with distribution on 2 tapes ($n_A=2$), larger chunk size yields lower average latency. In Scenario 4, one object was stored on one tape and only the $DF$ was varied at a fixed inter-arrival time of 900 seconds and average object size of 27 GB. Note that a chunk size of 5 or 6 GB yielded the lowest average latency; hence a $DF$ of 5 or 6 is the best value.

With regard to the performance of our scheduling algorithm, it was found to be moderately sensitive to variations in object size and highly sensitive to variations in inter-arrival time. When the object size and the inter-arrival time were fixed, analytical formulae could be written and verified. However, values of the average latency time for a fixed object size $T_{avg}$ and fixed inter-arrival time $t$ increased by a factor of up to ten when object size was varied and the inter-arrival time was an exponential function of $t$. FCFS scheduling with no striping and transfer alternation also showed some sensitivity also, though not as much that exhibited by the proposed algorithm.

## 8 Conclusion

A continuous data management approach using a tape-based hierarchical storage system was investigated. We extended the algorithm of a previously proposed replacement policy for a multi-user environment and subsequently tested it via simulation experiments. Continuous object decomposition into "chunks," chunk size, and their placement were studied under different characteristics of the tertiary storage subsystem, sizes of the secondary storage system, and service rates of continuous objects. Finally, based on the results of applying several heuristics and taking statistics using a simulated system, we developed a real-time like scheduling algorithm for materializing chunks from tapes to disks.

We conclude that continuous object decomposition is advantageous in a lightly loaded tertiary storage subsystem, one expected to have a high hit ratio (*i.e.*, object already on disks) for most requests arriving at large VOD servers; thereby resulting in a lightly loaded tertiary storage system. A second conclusion is that chunk placement on different tapes is only effective when $PCR<1$ and $STR \ll 1$; and in this case, the number of tapes $n_A$ on which chunks from one object are to be placed must be the minimum that give $n_A$. $PCR \geq 1$ and $n_A \leq k$, *i.e.*, being bounded by the number of drives in the tertiary storage subsystem. A third conclusion is that when $PCR \geq 1$, chunks from the same continuous object must be placed on the same tape, and alternating transfers for $r$ requests on the same drive is only advantageous when $r$ is less than $PCR$ by a factor $f$ ($f<1$) that compensates the increase (cost) in tape exchange/seek times.

Future work will be directed at investigating the point at which the bottleneck of the system switches from disk space freeing to chunk materialization congestion. Determining the factor $f$ beyond which transfer alternation with $PCR \geq 1$ gives a negative effect is also planned. Finally, as a combination of continuous and non-continuous data placement and accesses on a tertiary storage system will produce different characteristics from those demonstrated here, we plan to investigate this as a means to elucidate other placement and scheduling policies that can reduce system latency time.

## References

[BMGJ94] Berson,S.; Muntz,R.; Ghandeharizadeh,S.; Ju,X., "Staggered Striping in Multimedia Information Systems," in *Proceedings of the 1994 ACM-SIGMOD Int. Conference on Management of Data*, Minneapolis, Minnesota, May 1994.

[BO97] Boulos,J.; Ono,K., "Replacement Policy of a Multilevel Store for Continuous Media Servers," in *IPSJ Int. Symposium on Next-Generation Information Systems and Technologies*, Fukuoka, Japan, September 1997.

[BR96] Brubeck,D.; Rowe,L., "Hierarchical Storage Management in a Distributed VOD System," *IEEE Multimedia*, Vol. 3, No. 3, Fall 1996.

[CSIM94] CSIM18 Simulation Engine, Mesquite Software Inc., 3925 West Braker Lane,Austin,Texas78759-5321, 1994.

[CTZ97] Christodoulakis,S.; Triantafillou,P.; Zioga,F., "Principles of Optimally Placing Data in Tertiary Storage Libraries," in *Proceedings of the 23rd Very Large Data Bases Con-ference*, Athens,Greece,August 1997.

[DK93] Drapeau,A.; Katz,R., "Striped Tape Arrays," in *Proceedings of the 12th IEEE Symposium on Mass Storage Systems*, Monterey, CA, April 1993.

[GÖS97] Garofalakis,M.; Ozden,B.; Silberschatz,A., "Resource Scheduling in Enhanced Pay-Per-View Continuous Media Databases," in *Proceedings of the 23rd Very Large Data Bases Conference*, Athens, Greece, August 1997.

[GVK+95] Gemmel, D. J., et al., "Multimedia Storage Servers: A Tutorial," in *IEEE Computer*, Vol. 28, No. 5, May 1995.

[GS94] Ghandeharizadeh,S.; Shahabi,C., "On Multimedia Repositories, Personal Computers, and Hierarchical Storage Systems," in *Proceedings of the 2nd ACM Int. Conference on Multimedia*, San Francisco, CA, October 1994.

[GMW95] Golubchik,L.; Muntz,R.; Watson,R., "Analysis of striping techniques in Robotic Storage Libraries," in *Proceedings of the 14th IEEE Symposium on Mass Storage Systems*, Monterey, CA, September 1995.

[GP96] Gopalakrishnan,R.; Parulkar,G., "Bringing Real-time Scheduling Theory and Practice Closer for Multimedia Computing," in *Proceedings of the ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, Philadelphia, May 1996.

[HS96] Hiller,B.; Silberschatz,A., "Random I/O Scheduling in Online Tertiary Storage Systems," in *Proceedings of the 1996 ACM-SIGMOD Int. Conference on Management of Data*, Montreal, Canada, May 1996.

[LLW95] Lau,S.; Lui,J.; Wong,P., "A Cost-effective Near-Line Storage Server for Multimedia System," in *Proceedings of the 11th Int. Conference on Data Engineering*, Taipei, Taiwan, March 1995.

[LL73] Liu,C.; Layland,J., "Scheduling Algorithms for Multiprogramming in a Hard-Real-Time Environment," in *Journal of the ACM*, Vol. 20, No. 1, January 1973, pp. 46-61.

[ÖRS96] Özden,B.; Rastogi,R.; Silberschatz, A., "Disk Striping in Video Server Environments," in *IEEE Int. Conference on Multimedia Computing and Systems*, June 1996.

[PER97] Performance Evaluation Review, ACM SIGMETRICS, Vol. 25, No. 2, September 1997.

[Sar95] Sarawagi,S., "Query Processing in Tertiary Memory Databases," in *Proceedings of the 21st Very Large Data Bases Conference*, Zurich, Switzerland, September 1995.

[Ston91] Stonebraker,M., "Managing Persistent Objects in a Multi-Level Store," in *Proceedings of the 1991 ACM-SIGMOD Int. Conference on Management of Data*, Colorado, May 1991.

[TP97] Triantafillou,P.; Papadakis,T., "On-Demand Data Elevation in a Hierarchical Multimedia Storage Server," in *Proceedings of the 23rd Very Large Data Bases Conference*, Athens,

Greece, August 1997.

[WYS95] Wolf, J.; Yu, P.; Shachani, H., "DASD Dancing: A Disk Load Balancing Optimization Scheme for Video-on-Demand Computer Systems," in *Proceedings of the ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, Ottawa, Canada, June 1995.

[YD96] Yu, J.; DeWitt, D., "Processing Satellite Images on Tertiary Storage: A Study of the Impact of Tile Size on Performance," in *Fifth NASA Mass Storage Systems and Technologies Conference*, Maryland, Sept. 1996.

学術情報センター紀要　第10号(1998)

研究論文

# Data Exchange Bus for Advanced Media Delivery Systems Design and Architecture

# 高度なメディア配送システム設計のためのデータ交換バスとアーキテクチャ

Tredej TORANAWIGRAI
Graduate School of Engineering, University of Tokyo
東京大学大学院工学系研究科　トリーデージ　トラナウィカライ
Frederic ANDRES
National Center for Science Information Systems
学術情報センター　フレデリック　アンドレス
Kinji ONO
National Center for Science Information Systems
学術情報センター　小野　欽司

**ABSTRACT**

The Common Object Request Broker Architecture (CORBA), defined by the Object Management Group (OMG), provides not only flexible operations in heterogeneous communication environment, but it also supports interoperability, reusability, and durability. However, the current implementation of CORBA in commercial products provides low performance comparing to socket and it lacks of key ORB features, eg. IDL compiler, Portable Object Adapter, and Interface Repository; and so on.

Nowadays, the development of new ORBS based on CORBA (version 2.0) has been done in order to increase the performance and to integrate major key elements of ORB. One of them is TAO (The ACE ORB), developed by Washington University. TAO does not only support real-time processing but also enhances maintenance of ORB middleware, and increases portability and reuse of code by the ACE framework.

In this paper, we integrate TAO to the Phasme DBMS, which is used for multimedia management, under the AHYDS project. By using the object technology supported by CORBA, the remote operation and multimedia information retrieval processing can be achieved efficiently.

**要旨**

CORBAは、異種の通信環境下で、柔軟性かつ、相互運用性、再利用性、耐故障性と云う特長を備えている。しかし、現在の商業上のCORBA実装は、socketに比べると、まだ性能が低く、種々のORBの特徴も備えていない。IDL compilerやPoratable Object AdapterやInterface Repositoryなどが、その例である。

現在、性能向上や重要なエレメントを統合するために、CORBAの改良版(CORBA version2.0)がでている。その中に、Washington大学が開発したTAO(The ACE ORB)がある。TAOはreal-time processingの機能を支え、ORBミドルウェアを容易に維持して、Portabilityを増やし、さらにACE frameworkにより、codeの再利用もできる。

本論文は、我々が取り組んでいるAHYDS projectの下で、TAOとmultimedia管理のPhasmeというDBMSを統合する試みを報告する。このCORBAのオブジェクト技術を使うことで、リモートオペレーションやマルチメ

ディア情報を取ることの効果も期待できる。

[Keywords] CORBA, interoperability, reusability, durability, IDL compiler, Portable Object Adapter, Interface Repository, TAO, Phasme
[キーワード] CORBA、相互運用性、再利用性、耐故障性、IDL、コンパイラ、POA、インターフェースレポジトリー、TAO、Phasme

## 1 Introduction

Nowadays, the network computing environment is becoming more diverse. The supercomputers which perform complex calculation, the powerful servers, the computer at user-end including conventional system such as audio-video components, and etc. are included in the network environment in a distributed fashion. The computers and networks become faster and cheaper, while communication software are still slow and get bugs and expensive. This kind of such an environment creates the following problems:

- Making hardware from different vendors to work together, or working in heterogeneous environment is a difficult task.
- It is difficult to get the software to work together. Integration of different software, or achieving interoperability causes both time and money.
- Software development takes too long and costs too much. We need to build all software on a common foundation, instead of starting from scratch. The reusability and durability of software are necessary.

In addition to the problems of incompatible software infrastructures and the necessity of re-invention of core concepts and components when changing the platform, inherent complexity such as latency, partial failures, network partitioning, etc. is also an important problem that should be considered.

Choosing small components of software that can combine flexibly and dynamically to create tools focused on particular needs is better than using a large monolithic do-everything application. Therefore, the concepts of object technology are applied. Object-orientation splits computing problem into components that model the real world, these components will be easier to deal with and interaction between objects will appear logical and well-founded. Object-oriented concepts also solve the following problems.

- Programs are difficult to change or to extend.
- Programs are difficult to maintain.
- Programs take a lot of time to be written.

The key software technology developed to solve the problems mentioned above is the distributed object computing framework(DOC). DOC frameworks facilitate the collaboration of local and remote applications in heterogeneous distributed environments. They eliminate many tedious and error-prone distributed programming tasks from low-level programming, such as socket, and also enable portable developing and maintaining distributed application by automating common network programming tasks. At the heart of DOC middleware are Object Request Brokers (ORBs) such as the Common Object Request Broker Architecture (CORBA).

However, like many other distributed applications, the current implementation of ORBs is still low efficient over high speed network compared to that of the low-level implementation [1]. And because of statically configured software designs, which is hard to maintain and to optimize, conventional ORBs cannot be extended without modifying their source code, which forces recompilation, relinking, and restarting running ORBs and their associated applications [8].

The general purpose of this study is to introduce efficient CORBA2.0 management inside a DBMS and to optimize the ORB mechanisms for the DBMS management in terms of multimedia and hypermedia supports. First, we review the CORBA state

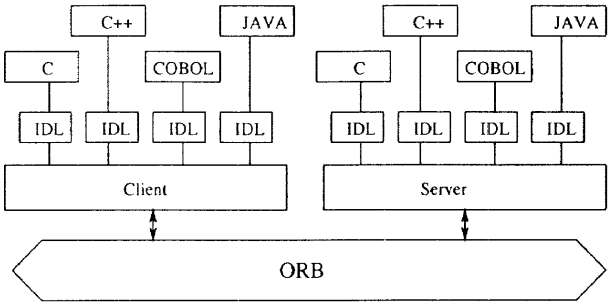



Figure 1　CORBA IDL binding client/server interoperability

of the Art, including starting from the investigation of the ACE (the Adaptive Communication Environment) developed by Washington University. After that, we will integrate CORBA2.0 compliant real-time ORB, named TAO (The ACE ORB), to the Phasme DBMS and evaluate the performance of the system by using Quantify[4] which is able to analyze performance bottlenecks and identifies sections of code that dominate execution time. Final works are the optimization of the system and the creation of a demonstration as a part of the AHYDS platform.

The remainder of this paper is organized as follows: Section 2 gives an overview of CORBA including shortcomings and related works about CORBA; Section 3 is the overview of related researches; Section 4 describes the he architecture of CORBA-based data delivery system; Section 5 presents the interface between TAO and Phasme DBMS; finally Section 6 concludes and gives a future working plan.

## 2　CORBA Background

### 2.1　Introduction to OMG's CORBA

The Object Management Group (OMG) is a consortium that is working on the object technology. It defines Objects Management Architecture (OMA) which is the multivendor standard for object-oriented distributed computing. In the OMA, the Common Object Request Broker Architecture (CORBA) is included and intended to support the production of flexible and reusable distributed services and applications. It defines middleware that has the potential of subsuming every other form of existing client/server middleware. We can say that CORBA was designed to allow intelligent components to discover each other and interoperate on an object bus.

In CORBA, the services that an object provides are expressed in a contract that serves as the interface between it and the rest of the system. The interface specifications are written in a neutral Interface Definition Language (IDL) that defines a component's boundaries with potential clients. Component written to IDL should be portable across languages, tools, operating systems, and networks. These components should be able to interoperate across multivendor CORBA object brokers too.

To invoke the remote operation at an object, the client is not necessary to know where the object is on the network, or which language the object is implemented in the needs to know only the interface that the server object publishes. We define the component that contains the definitions of all interfaces as *Interface Repository*. It contains the metadata that lets components discover each other dynamically at run time. This make CORBA a self-describing system.

### 2.2　The OMA: Application-Level Integration

Because CORBA connects only objects interface, not application itself, application-level integration is another issue that should be considered. The OMG's Object Management Architecture (OMA) is defined to integrate the application into the system.

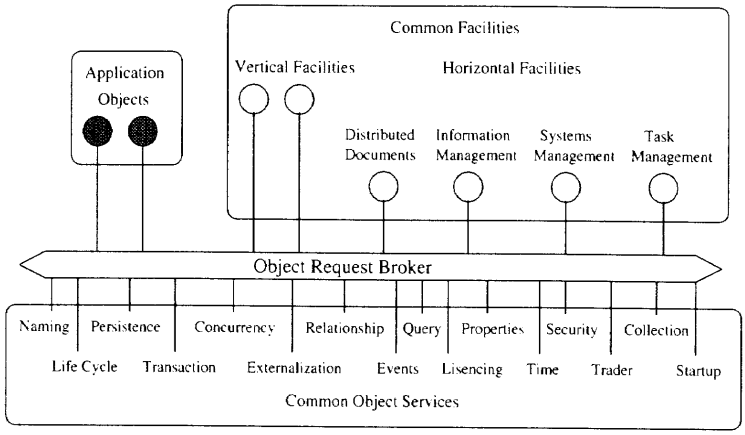



Figure 2 The OMG Object Management Architercture

The OMA components can be divided into 4 major elements: Object Request Broker (ORB), low-level CORBA services, intermediate-level CORBA facilities, and Application objects, as shown in the figure 2.

### 2.2.1 The Object Request Broker

The ORB is the object bus, which lets objects transparently to make requests or to receive responses from other objects located locally or remotely. An ORB is much more sophisticated than other forms of client/server middleware including Remote Procedure Calls (RPCs), Message-Oriented Middleware (MOM), database stored procedures, and peer-to-peer services.

The benefits of using CORBA ORB can be summarized as follows:

- Static and dynamic method invocation: CORBA allows to define statically method invocations at compile time, or to dynamically discover them at run time.
- High-level language bindings: CORBA separates interface from implementation and provides language-neutral data types that make possible to call objects across language and operating system boundaries.
- Self-describing system: Every CORBA ORB supports an Interface Repository that contains real-time information describing the functions that a server provides and their parameters.
- Local/remote transparency: An ORB can run in standalone fashion on a computer or it can be connected with other ORBS by using CORBA2.0's Internet Inter-ORB Protocol (IIOP) services. An ORB can broker inter-object calls within a single process, multiple processes running within the same machine, or multiple processes running across networks and operating systems.
- Built-in security and transaction
- Polymorphic messaging: The same function call will have different effects depending on the objects that receive it. Comparing to RPC, RPC calls have no specificity, all functions with the same name get implemented in the same way.
- Coexistence with existing system: Using CORBA IDL allows writing new application as pure objects and encapsulates existing applications with IDL wrappers.

The structure of CORBA 2.0 ORB is shown in the figure 3.

The components on the client side include:

- The Client itself: This is the program entity that invokes an operation on a Servant that can be remote or co-located.
- The Client IDL Stubs: The stub provides static interfaces to object services. The precompiled stubs define how clients invoke services on the servers. Stub is a local proxy for a remote

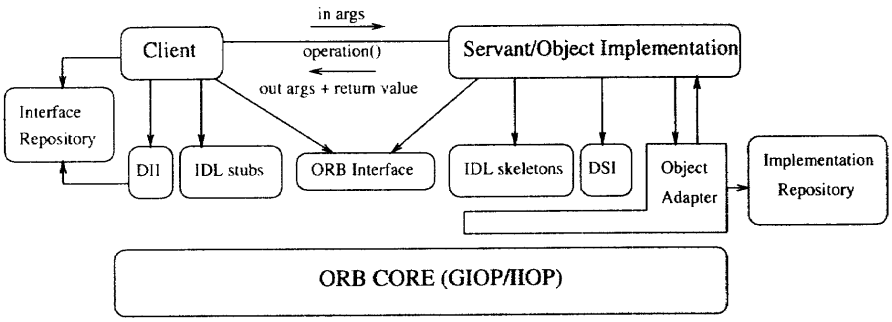



Figure 3　CORBA 2.0 ORB structure

server object, it is generated by the IDL compiler to contain marshaling code.

- The Dynamic Invocation Interface (DII): DII allows a client to directly access to the underlying request mechanisms provided by an ORB. DII is used when the ORB has no compile-time knowledge of the interface it is implementing. To retrieve this information, the ORB must query an Interface Repository.
- The Interface Repository APIs: These APIs allow users to obtain and to modify the parameters and descriptions of all registered interfaces.
- The ORB Interface: The ORB interface provides standard operations that decouple applications from the implementation details of the ORB.

Static invocations are easier to program, faster and self-documenting because of pre-compilation before running program, while dynamic invocations provide maximum flexibility, but they are difficult to program and useful for tools that discover services by querying Interface Repository at run time. The server side cannot separate the differences between static and dynamic invocations. The components of CORBA on the server side includes:

- The Servant or Object Implementation: It defines operations that Implement a CORBA IDL interface. In languages like C++ and Java that support Object-oriented programming, servant are typically implemented using one or more objects.
- The Server IDL Stubs or Skeletons: Skeletons, created by IDL compiler, provide static interfaces to each service exported by the server.
- The Dynamic Skeleton Interface (DSI): This interface provides a run-time binding mechanism for servers that need to handle incoming method calls for components that do not have IDL-base compiled skeletons. The DSI is the server equivalent of a DII. It can receive either static or dynamic client invocations.
- The Object Adapter: it sits on the top of ORB's core communication services and accepts requests for service on behalf of the service's objects. It demulipexes incoming requests to the Servant and dispatches the appropriate operation of that Servant. The adapter also registers the classes and run-time instances with the *Implementation Repository*. The CORBA specifies that each ORB must support a standard adapter called the *Basic Object Adapter* (*BOA*).
- The Implementation Repository: It provides a run-time repository of information about the classes a server supports, the objects that are instantiated, and their ID.
- The ORB Interface: It is the same as the ORB interface on client side.

One of the most important components in this architecture is ORB Core. The ORB Core acts as the mediator between clients and servants. When a client invokes an operation, the ORB Core is responsible for delivering the request to the servant and returning a response to the client. For clients and servers not executing on the same machine, CORBA 2.0 compliant ORB Cores communicate via

the General Inter-ORB Protocol (GIOP) and the Internet Inter-ORB Protocol (IIOP). The ORB Core is typically implemented as a run-time library linked into client and server application.

#### 2.2.2 CORBA services

CORBA services provide basic functionality that almost any objects might need. The CORBA services are used to create a component, name it, and introduce it into the environment. OMG has published standards for 16 object services, such as Naming Service, Event Service, Concurrency Control Service, etc. All of these services enrich a distributed component's behavior and provide robust environment for CORBA. (The details of these are given in [9].)

#### 2.2.3 CORBA facilities

Where CORBA services provide services for objects, the CORBA facilities which are collection of IDL-defined frameworks, provide services for applications directly. The CORBA facilities can be divided into 2 major components: horizontal and vertical facilities.

Horizontal facilities are usable by nearby everyone, they can be divided into 4 basic areas:

- User interface: makes an information system accessible to its user and responsive to their needs.
- Information management: uses for modeling, defining,storing, retrieving, managing, and interchanging information.
- System management: uses for management of complex, multivendor information systems.
- Task management: uses to automate work processes, including both user and system processes.

Vertical facilities are those shared by a number of different applications within a specialized market area such as Healthcare, Telecommunications, Financial Services, Manufacturing and Business Objects.

### 2.3 Implementation of CORBA

As describe before, CORBA supports both static method invocation, via stub and skeleton, and dynamic method invocation, by Dynamic Invocation Interface (DII). In this section, implementation of both of them will be described.

#### 2.3.1 Static Invocation

In order to invoke processes at server, the client performs a request by accessing to an object reference (object ID) and invoking the method that perform the service, whereas in the server side, the server cannot specify if the invocation method is static or dynamic invocation.

Clients see the interfaces through a language mapping or binding, that brings the ORB to the programmer's level. Client programs should be able to work without any changing on any ORB that supports the language binding, and should be able

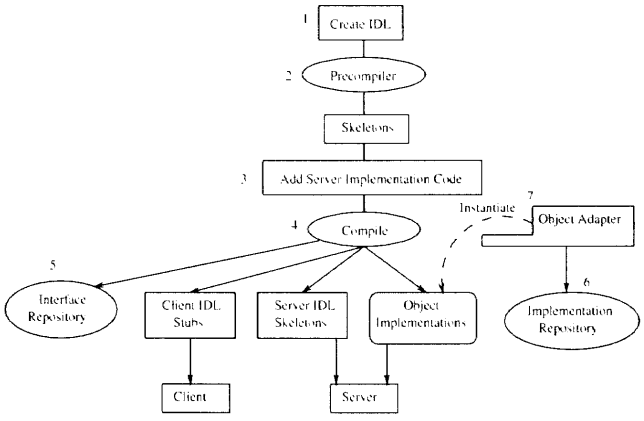



Figure 4 Static metod invocation

to call any object instance that implements the interface.

The static stub interface is bound at compile time and is easier and natural to program because user can call remote method by simply invoking it by name and passing it the parameters.

The following steps are procedures to implement CORBA.

1. *Define object classes using Interface Definition Language (IDL)*
2. *Run the IDL file through a language precompiler*
3. *Add the implementation code to the skeletons*
4. *Compile the code*
5. *Bind the class definitions to the Interface Repository*
6. *Register the rum-time objects with the Implementation Repository*
7. *Instantiate the objects on the server*

The server side program is shown in fig.5, whereas the client side program is shown in fig.6

### 2.3.2 Dynamic Invocation

The Dynamic Invocation Interface (DII) does not need any stubs at run time. It provides a flexible environment because user can add new classes to the system without requiring any changes in the client code. It is very useful for tools that discover services provided at run time.

The client may discover the object reference of remote objects by using the CORBA Naming Service or by the Trader Service (via Interface Repository).

The procedures of dynamic invocation can be summaried as the following step.

1. *Obtain the interface*: name by using get_interface method. This call returns a reference to InterfaceDef object, which is in the Interface Repository that describes the interface.
2. *Obtain the method description from the Interface Repository*: by using InterfaceDef, all of the information about the interface and methods it supports can be obtained.
3. *Create the argument list*: CORBA specifies a self-defining data structure for passing parameters, which it calls the Named Value List. User creates the list by invoking create list, and adds each argument to the list by add_item calls. User can also, alternately, ask ORB to create the list by using create operation_list on the *CORBA:ORB* object.
4. *Create the request*: A request is a CORBA pseudo-object that contains the name of the method, the argument list, and the return value. User creates a request by using create request or user can create the short version of request by invoking _request without requiring parameters.
5. *Invoke the request*: User can invoke a request by 1 of 3 ways: 1) synchronous: the invoke

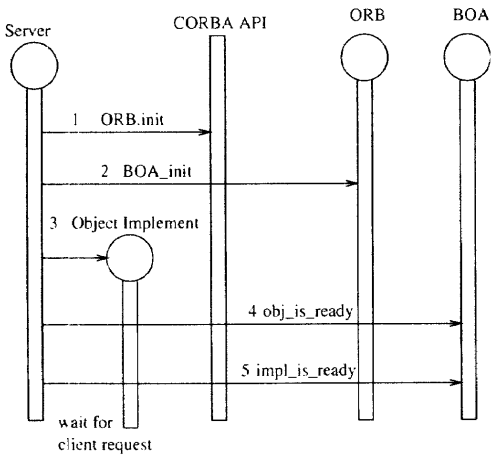



Figure 5 Server side program

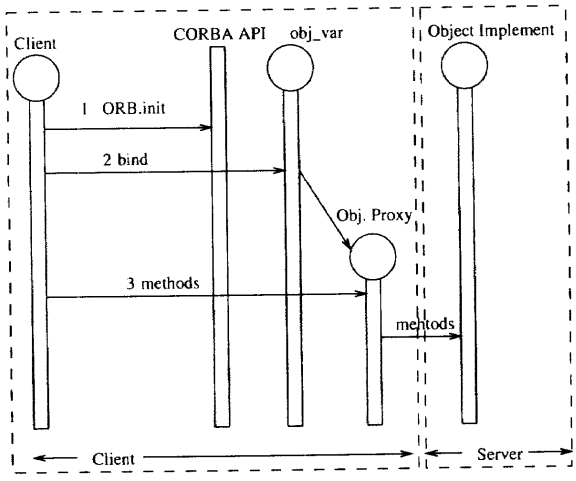



Figure 6 Client side program

call sends the request and obtains the results; 2) deferred synchronous: the send_deferred call returns control to the program, which must poll for the response by issuing poll response or get_response; 3) one-way: the send call can be defined to be a datagram by issuing send_oneway, without requiring any response.

The services which user needs to dynamically invoke an object are part of CORBA core. The methods are in the four interfaces of CORBA module. The methods in these four interfaces are:

1. *CORBA::Object* is a pseudo-object interface that defines operations that every object must support. It is the root interface which includes 3 methods, _interface, create request, and _request, that can be used to construct DII. 2. *CORBA::Request* is a pseudo-object interface that defines the operations on a remote object.
3. *CORBA::NVList* is a pseudo-object interface that helps user construct parameter lists. An NVList object maintains a list of self-describing data items called NameValues. The NVList interface defines operations that let user manipulate the list.
4. *CORBA::ORB* is a pseudo-object interface that defines general-purpose ORB methods. User can invoke the methods on an ORB pseudo-object from a client or server implementation.

The procedures of invocation DII can be concluded into the steps in fig. 8,9, and 10.

### 2.4 Advantages and Shortcoming of CORBA

#### 2.4.1 Advantages of CORBA over low-level interfaces

Although the standard low-level interface such as socket is widely used for performing routines such as locating address information for network services, establishing and terminating connections and sending and receiving data, its design still has several important limitations. Using CORBA concepts to automate common low-level programming tasks enables programmer to concentrate on higher -level details such as performance, reliability, and interface uniformity. The advantages of CORBA over low-level programming such as socket can be described as follow:

- Strong-typed interfaces: All interfaces in CORBA using IDL. A CORBA IDL compiler generates stubs and skeletons that translate IDL into object implementing language. The use of IDL interface allows the transmission of strongly-typed data across network. Strong typing also improves abstraction and eliminates errors common to socket programming. For instance, in case of send and receive operations implemented over socket, it is necessary to convert the typed information into a stream of untyped bytes manually. The sender and receiver software must be tightly coupled to ensure correctness which provides high possibility of errors.
- Parameter marshalling and framing: The stubs and skeletons from IDL compiler ensure correct byte ordering and linearization of all parameters sent via operation calls on CORBA interfaces over network. In case of using socket, manual converting data from host-byte order into network-byte order must be done, and if the bytestream-oriented TCP/IP was used, framing the data correctly at the receiver is also necessary. Marshalling and framing are tedious and error-prone aspects of programming, by using CORBA, these aspects can be eliminated.
- Object location and object activation: CORBA supports location transparency, i.e., services can be located anywhere in a distributed system. Therefore clients can access both local and remote objects. This feature enables independency between clients and implemented objects.

#### 2.4.2 Shortcoming of current ORB implementation

However, current implementation of CORBA over

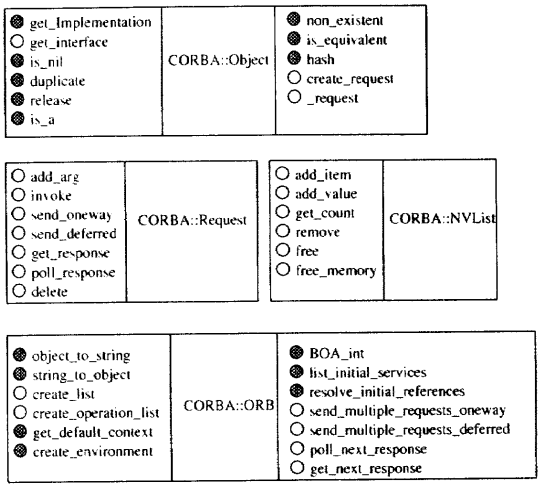



Figure 7　Dynamic Invocation Interface

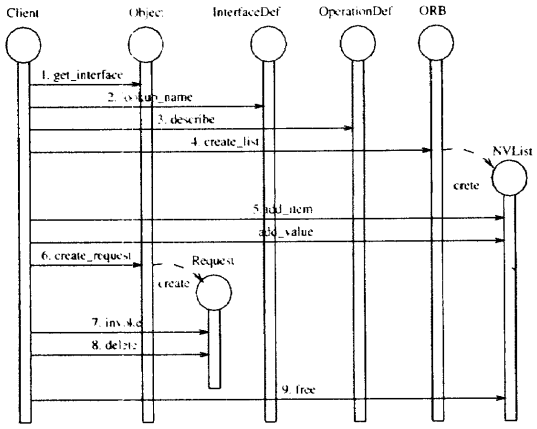



Figure 8　First way of DDI invocation

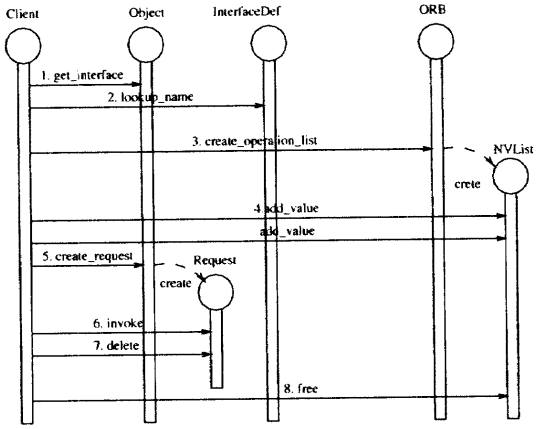



Figure 9　Second way of DDI invocation

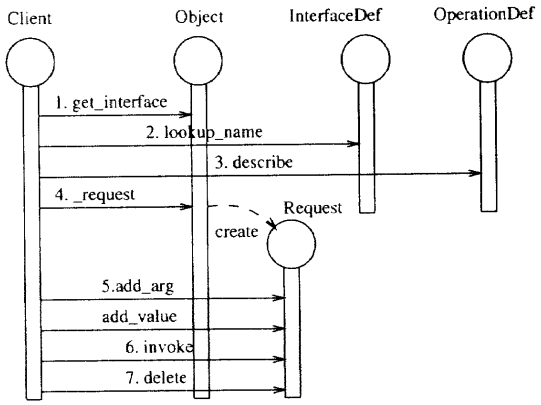



Figure 10　Third way of DDI invocation

high speed network does not achieve the high performance comparing to low-level interface such as socket. As described in [1], CORBA implementation overhead stems from data copying. Although read and write systems calls dominated the execution of CORBA implementations, the highest cost tasks involved data copying and data inspection. The IDL stubs and skeletons copy data multiple times ( from TCP data buffer into a marshalling buffer, and then again into the parameter passed to the send upcall).

Another source of overhead comes from demultiplexing process. Each CORBA request message contains the name of its intended remote operation which is represented as a string. Performing linear search through the list of operation in the IDL interface to demultiplex incoming message causes search time grow linearly with the number of operations in the IDL interface, whereas using hashing search for demultiplexing incoming requests is likely to scale better for large IDL interfaces, but less efficient for small interfaces due to the overhead of computing hash function. Using the suitable strategies can improve the performance of system.

The problem of memory allocation is also one of the sources of overhead in current CORBA implementation. IDL skeletons do not know how user -supplied upcall will use the parameters passed to it from the request message. Therefore, the conservative memory management techniques, that dynamically allocate and release copies of messages before and after an upcall, are used. These memory management policies is suitable for some circumstances such as in a multi-threaded application, but they increase processing overhead for streaming applications that consume their data immediately without modifying it.

In addition to the problems mention above, commercial ORBs such as Orbix or the SunSoft IIOP, which is the standard reference implementation of IIOP written in C++, also has these limitation [8]:

- Lack of key ORB features: Although SunSoft IIOP provides an ORB Core, a robust IIOP protocol engine, and DII and DSI implementation, it lacks an IDL compiler, an Interface Repository, and a Portable Object Adapter (POA).
- Lack of portability: Like most communication software, SunSoft IIOP is directly programmed with low-level networking and OS APIs, therefore these APIs are not portable across OS platforms.
- Lack of configurability: Like many ORBs and other middleware, SunSoft IIOP is statically configured, which makes it hard to extend without modifying its source code directly. Statically configured ORBs are also inefficient in term of time and space. For example, time inefficiency can stem from the inability of a statically configured concurrency model, such as single-threading to be responsive for long -duration requests. Space inefficiency can be attributed to linking many component into an ORB that are unnecessary for many use-cases, which increases the memory footprint and forces applications to pay a space penalty for features they do not require.
- Lack of software cohesion: SunSoft IIOP focuses on solving a specific problem, i.e. implementing an ORB Core and an IIOP protocol engine. It accomplishes this using a tightly-coupled ad -hoc implementation that hard-codes key design decisions.

Many researches have worked involving development of CORBA implementation and alleviation the problems and overhead described above. The next section will describe the relating work concerning performance improvement and the application of CORBA in various fields.

## 3 Related Works

The performance of CORBA and other network programming mechanisms such as sockets on Ethernet and ATM networks is reported in [1]. The results show that for bulk data transfer, the performance overhead of widely used CORBA implementations on high speed ATM networks is much lower than achievable by low-level interface. The paper

also introduces encapsulation low-level network programming interfaces with ACE (the Adaptive Communication Environment) C++ object-oriented wrappers to improve the performance of data transfer. However the ACE wrappers do not address higher-level issued related to the reliability and availability of system, flexible object location and selection, support for transactions, security, and deferred process activation, and the exchange of binary data between different computer architectures.

The ACE C++ wrappers for socket can be integrated with CORBA to enhance the performance of streaming application. The example of system using this concepts is [2], which combines CORBA and the ACE wrappers in a high speed system that transfers 10-40 Mbyte medical images over ATM. In this system, CORBA is used as a signaling mechanism to identify endpoints of communication in a location-independent manner. The ACE wrappers are then used to establish point-to-point TCP connections and transmit bulk data efficiently across the connections.

In [5] describes the reliability of distributed system by using CORBA. The CORBA model itself does not provide solutions to the problem of detecting and reaching to partial failures and to network partitioning. However, a CORBA object can encapsulate internal state and make it accessible through IDL interface. Due to this encapsulation, fault-tolerance techniques like replication and state-checkpointing become easier to implement because the internal state of an object is isolated. The models that can help in building reliable system in this paper include: message queues, Transaction Processing monitors, and Virtual Synchrony. Each model provides different types of reliability. The object-oriented architecture that combines these three models and allows application to pay only for reliability guarantees they need is shown in the figure 11

This architecture consists of the ORBs running on top of a virtually synchronous group communication that supports the abstract of object group [3], meaning that CORBA objects of the same type can be named and accessed as a single entry. Object groups allow run-time replications of stateful CORBA objects and efficient multicast of CORBA requests. The TP monitor and message queue handler are provided in the form of plug-in OMG common object services on top of the ORB. The virtual synchronous ORBs can facilitate both TP monitor and message queue efficiently.

Meanwhile in [6] and [8], the optimization concepts to increase the performance of CORBA and the patterns used for extending dynamically configured ORB middleware are introduced respectively. [6] measures the performance of SunSoft IIOP and optimizes by performing the following 4 steps:

- Inlining to optimize for the common case
- Aggressive inlining to optimize for the common case
- Precomputing, adding redundant state, and passing information through layers
- Eliminating gratuitous waste and specializing generic methods

The results of applying these optimization principles to SunSoft IIOP improve its performance for various kinds data type. And in [8] the pattern technology is used to develop extensible ORBs, for example the Wrapper Facade pattern, the Reactor pattern, the Active Object pattern, and so on. (More details are given in [12].)

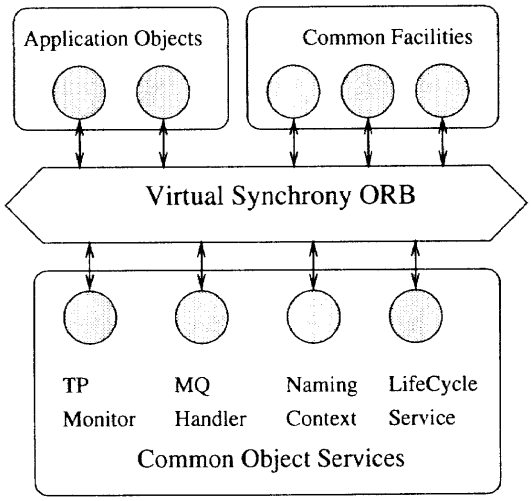



Figure 11　Extended CORBA for reliable systems

All of the patterns mentioned above are implemented in The ACE ORB (TAO) which is a real-time ORB middleware developed by Washington University. The use of the patterns in TAO reported in this paper provides not only the increased extensibility but they also enhance maintenance of ORB middleware and increase portability and reuse of code by the ACE framework.

In this section, we overview research works done around CORBA and research works combining CORBA support and Information System. CORBA research works have tried to improve the performance and to increase dynamically the adaptability of CORBA middleware.

## 4 Architecture of the CORBA-based Data Delivery System

### 4.1 Overview of the Communication

The current multimedia information system consists of various kinds of tools and information. The different kinds of multimedia information, which are linked together can be called the hypermedia information system. However, most of the existing communication systems do not support the functionalities required by the hypermedia information. Therefore, the Active Hypermedia Delivery System (AHYDS) [13] was developed by the Japanese National Center for Science Information System (NACSIS) to provide storage and retrieval a set of hypermedia documents efficiently. The AHYDS uses the application-oriented Phasme DBMS to resolve the problems of uniformity data storage and to support traditional DBMS's services. In addition to this, the AHYDS takes the interoperability between the application into account. Therefore, CORBA is one of the most efficient candidates to cope with this issue.

As mentioned earlier, CORBA acts as the mediator between the application programs which are based on the client-server architecture. To achieve the interoperability among different kinds of platforms and various formats of information, CORBA cores communicate via the General Inter-ORB Protocol (GIOP) and Internet Inter-ORB Protocol (IIOP). However, the services in CORBA are expressed as the interfaces by using IDL as described earlier. Therefore, The design of dynamic, flexible, transparent interfaces is necessary for enabling efficiently using powerful application, like AHYDS, over CORBA in heterogeneous distributed environment.

### 4.2 The Communication Level

The global view of the system is provided in fig.12. It includes the Application level (AHYDS), the Communication level or API (Appilcation Programming Interface), and the middleware level which is CORBA. In our communication mechanism, we use TAO which is the real-time ORB as the mediator to communicate between client and server.

The communication level between TAO and AHYDS can be separated into 2 sub-levels: the MHS (Message Handle System) layer and the IPC (Inter Process Communication) layer.

- *The MHS layer* provides the message management. It defines the procedures to send/receive message, inform the error during the message transfer, subscribe/unsubscribe the event service, including other message management service.
- *The IPC layer* supports various kinds of protocol using for communication in the distributed system. It uses the agents to perform the basic services, such as start/stop the communication, message handling, read-write-synchronization, and so on.

In our architecture, one thing which is taken into account is the flexibility. The interfaces between AHYDS and TAO should be able to adapt the communication protocol dynamically to utilize fully performance of both systems, because sometimes the application may be in form of one-to-one. However, sometimes it may be one-to-multipoint style. In addition to the flexibility, the communication layer should enable the transparency supported by TAO such that from the viewpoint of user, there is no difference to execute the process locally or

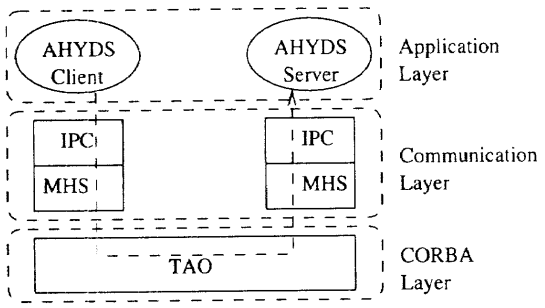



Figure 12　The architecture of the communication mechanism

remotely.

Another characteristic of the communication layer is that, it provides real-time interfaces for the information system which is a DBMS in this case. By using real-time ORB like TAO, this features should be considered when we design the interfaces.

Comparing our design to iBus [11], which is a communication software bus that supports intranet application written in JAVA, our design uses the services from TAO which is more standardized than iBus.

The communication based on TAO provides interoperability among objects implemented in different languages in various kinds of platforms, where as iBus is only applied for the application written in JAVA. Although, JAVA can run on any platforms like CORBA, the speed of operation is not so fast, comparing to TAO which uses Wrapper Facade Pattern to increase the speed.

However, iBus has advantages in that it does not need the IDL compiler or platform specific library and no need to implement marshaling code for the iBus channel. Whereas CORBA still needs the IDL compiler and some marshaling procedures which are the sources of overhead in performance evaluation. However, TAO improves these shortcomings by using pattern technology described in [8].

## 5 API of Data Exchange Bus

In this section, the interface between TAO and Phasme are described. As mentioned in the previous section, TAO provides various kinds of services to enhance the communication, to make it efficient and adaptable, for instance, the real-time and QoS support, the pattern (in ACE framework), and so on; in this section we will also discuss about this functions.

As described in the previous subsection, the Message Handle System is one of the most important things in the communication system. The structure of our interface shown in fig.13. From now on we will discuss about the interfaces of the communication system in details, starting from the MHS object.

### 5.1 The MHS object

The MHS object is defined as the representation of the message sender and receiver. It creates the Comm object, the Receiver object, and the RT_Info object in order to communicating with other MHS by considering the protocol specified in the Protocol object. The MHS object uses the Comm object to send message/event, subscribe/unsubscribe event service, and set the parameter in RT_Info object to maintain the quality of service in the communication process. (The detail of the Comm object, the Receiver object, Protocol object and the RT_Info object will be described later.)

The interfaces of MHS object is composed of:

- void CreateMHS(String name)
- void DelMHS(String name)
- void StartMHS(String name)
- void StopMHS(String name)

The MHS object is created by CreateMHS interface and the StartMHS is used to initialized and to start the MHS object. To stop running the MHS object, the user uses the StopMHS interface to stop it, and uses DelMHS to delete the MHS object.

### 5.2 The Comm object

The Comm (Communication) object is the main object, created by the Message Handle System (MHS), which may be either a client or a server. The communication procedures are performed through the Comm object. Its interfaces include:

- void CreateComm(MHS caller,Protocol ptc)

- void DelComm(MHS caller)
- void StartComm(MHS caller)
- void ShutdownComm(MHS caller)
- void Sendmsg(MHS sender, MHS receiver, Priority p)
- void SendEvent((MHS caller, MHS destination, EventType event)
- void SubscribeEvent(MHS caller, MHS source, EventType event)
- void UnsubscribeEvent(MHS caller, MHS source, EventType event)
- void GetEvent(MHS caller, EventType event)
- void NotifyEvent(MHS owner, EventType event)
- void SetRTvalue(MHS caller, RT_Info value)
- void GetRTvalue(MHS caller, RT_Info value)

To create the Comm object, the MHS which invokes this process and the protocol must be specified as parameters of CreateComm interface, whereas DelComm is used to destroy the Comm object. The Comm object will be started and stopped by using the StartComm and ShutdownComm interfaces, respectively.

In order to send the message, the Comm uses Sendmsg by specifying the MHS which is sender and MHS which is receiver including the priority of the message. The priority of the message is determined by considering the message type, the application which generates the message, and the user profile. Some users, such as director or administrator may have higher priority in sending message than the normal users have. It should be noted that there is no need to perform string copy when the MHS sends the message because TAO supports sending message from memory directly without buffering.

When the Comm will send an event, it will specify to the MHS sender, the MHS receiver, and the event type.

To subscribe or unsubscribe an event service, the MHS will tell the Comm object to subscribe or unsubscribe that event service by specifying to the MHS which invokes the process, the MHS which owns the event, and the event type.

When the MHS wants to know if it subscribed a certain event, it will call GetEvent function via the Comm object.

When there is a new message arriving, the Comm object will notify the message arrival to the MHS by using NotifyEvent.

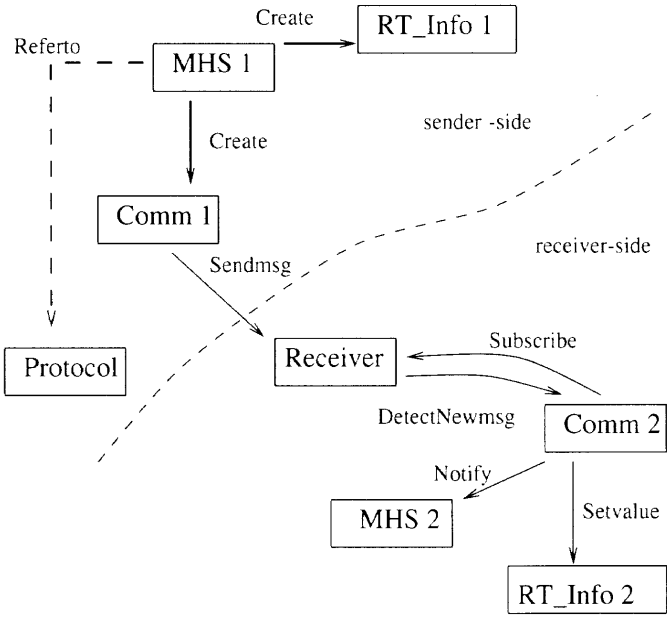



Figure 13 The archiecture of the interface

To maintain the quality of service, the Comm can assign and retrieve the real-time information (RT Info) object, which provided by TAO, by using the SetRTvalue and GetRTvalue interfaces, respectively.

### 5.3 The Receiver object

The Receiver object is used to receive the message or event sending from the Comm object of the sender side. It performs as the proxy of the sender on the receiver side and it will detect the new coming message and tell the Comm object to notify the MHS.

The interfaces of the Receiver object will be described as following.

- void CreateReceiver(MHS caller)
- void DelReceiver(MHS caller)
- void DetectNewmsg(MHS sender, MHS receiver, Protocol ptc)
- void Dectectmsgerror(MHS sender, MHS receiver, String err)
- void DectectNewEvent(MHS sender, MHS receiver, Protocol ptc, EventType event)
- void DectectEventerror(MHS sender, MHS receiver, String err)

When the new message is coming, the Receiver object will perform DetectNewmsg to inform the Comm object. In this interface, the parameters that have to be specified include the message sender, the receiver of the message, and the protocol type. In case there any errors during message transfer, the Dectectmsgerr will specify each error in err string.

For the new arriving event, the Receiver will process the new message in the same way as previously but using DetectNewEvent function instead, and also specify who owns the event type. If some errors occur, the DectectEventerror will specify each error in err string in the same way as message error.

### 5.4 The Protocol object

The Protocol object is used to specify type of protocol used in the system. Since our MHS and IPC system are supported by TAO, the Protocol should be able to change dynamically. By using the patterns (such as Reactor, Proactor, Active object, etc.) provided by ACE framework in TAO, various styles of protocol (such as synchronous, asynchronous, and defer-synchronous) could be achieved. The interface of the Protocol object will be discussed as the following.

- void CreatePtcObject(String pname)
- void DelPtcObject(String pname)
- Protocol ReferTo(MHS caller, String ptc name)
- void S_Client_S_Server(MHS caller, MHS destination)
- void S_Client_M_Server(MHS caller, MHS_set serverset)
- void M_Client_S_Server(MHS caller, MHS_set clientset, MHS destination)
- void M_Client_M_Server(MHS caller, MHS_set clientset, MHS_set serverset)

The Protocol object is created and deleted by CreatePtcObject and DelPtcObject respectively with specifying the name of the protocol. After creating a new instance of Protocol object, it should be registered with the naming service.

The MHS will call the ReferTo interface when it wants to check if the Protocol is supported or not by specifying the caller name and the protocol name.

The S_Client_S_Server interface is for single-client-single-server. The MHS which calls this function and the MHS which is the destination have to be specified when invoking this operation. The other interfaces are single-client-multi-server, multi-client-single-server, and multi-client-multi-server respecively. For single-client, only the caller should be specified, but in case multi-client, the client set should be indicated too. For the single-server protocol, only the destination should be specified, but for the multi-server, the set of server has to be indicated.

### 5.5 The RT_Info object

The RT_Info object is used to maintain the quality of service in the communication system by specify-

ing the real-time parameters. The MHS object uses the Comm object to set and to retrieve the real-time parameters in this object. The information specified in this object includes:

- const String name
- Time MaxExeTime
- Time AvgExeTime
- Time PeriodTime
- Priority priority

The parameter name is the name of the operation invoked by the sender or receiver. The MaxExeTime is the maximum time that the operation requires to finish the task, while the AvgExeTime is the average time that most of the operations require. The PeriodTime is the minimum time between the successive iterations of the operations. And the priority indicates how the operation is important.

## 6 Conclusion and Future Work

The paper provides the specification and the design of interfaces between TAO and AHYDS.

This paper introduced CORBA which is a helpful concept of distributed object computing framework for communication system. CORBA also provides interoperability, reusability, and durability to the heterogeneous communication environment. In this paper, described the necessity of using CORBA and purpose of our study, which is to implement the CORBA 2.0 to the DBMS and optimization the system. Then we overviewed CORBA architecture, the shortcoming of currently implemented CORBA (CORBA1.2), and review of literature were presented. we described the related researches and then the architecture of the CORBA-based Data Delivery System has been.

The implementation of the Data Exchange Bus, the improvement, and the optimization of the system are planned for the first trimester of 1998. Finally, the demonstration of CORBA for distributed and heterogeneous environment of database will be done from March 98.

## 7 Acknowledgement

We thanks for Dr. Douglas C. Schmidt and his group from Washington University for the collaboration and their help.

## References

[1] Schmidt,Douglas C.; Harrison,Tim.; Al-Shaer,Ehab., "Object-Oriented Components for High-Speed Network Programming", *USENIX Conference on Object-Oriented Technologies*, *Monterey*, CA, June 1995.

[2] Pyarali,Irfan.; Harrison,Timothy.; Schmidt,Douglas., "Design and Performance of an Object-Oriented Framework for High-Speed Electronic Medical Imaging", *USENIX COOTS conference*, Toronto, Canada, June 1996.

[3] Landis,Sean.; Maffeis,Silvano., "Building Distributed Systems with CORBA". Theory and Practice of Objecct Systems,April 1997, John Wiley, NewYork

[4] Pure Software. *Quantify User's Guide*, 1996.

[5] Maffeis,Silvano.; Schmidt,Douglas., "Constructing Reliable Distributed Communication Systems with CORBA", *IEEE Communications Magazine*, Vol. 14, No. 2, Feb. 1997.

[6] Gokhale,Aniruddha.; Schmidt,Douglas., "Optimizing the Performance of the CORBA Internet Inter-ORB Protocol Over ATM", *Washington University technical report*, wucs-97-10.

[7] Schmidt,Douglas.; Levine,David.; Mungee, Sumedh., "The Design and Performance of Real-Time Object Request Brokers", *Computer Communications Journal*, Summer 1997.

[8] Schmidt,Douglas.; Cleeland,Chris., "Applying Patterns to Develop Extensible and Maintainable ORB Middleware", Communications of the ACM Special Issue on Software Maintenance, Vol.40, No.12, Dec. 1997.

[9] Siegel,Jon., *CORBA Fundamentals and Programming*, John Wiley & Sons Inc., USA.,1996.

[10] Orfali,Robert.; Harkey,Dan., *Client/Server Programming with JAVA and CORBA*, John Wiley & Sons Inc., USA.,1997.

[11] Maffeis,Silvano., "iBus-The Java Intranet Software Bus", Feb. 1997.

[12] Toranawigrai,T.; Andres,F.; Ono,K., "Data Exchange Communication Protocol for Advanced Media Delivery Systems", Technical report of NACSIS R&D Department, Dec. 1997.

[13] Andres,F.; Ono,K., "The Active Hypermedia Delivery System (AHYDS) using an Application-oriented DBMS", Invited Talk, 4th International HyTime Conference, Montreal, Canada, August 1997.

[14] Andres,F.; Ono,K., "Phasme Un Systeme Parallele de Gestion de Bases de Donnees Griente Application", *Calculateurs Parallels Journal special issue on Parallel and Distributed Database System*, 1997.

学術情報センター紀要　第10号(1998)

研究論文

# Delay Performance Analysis of an ATM Multiplexer

# ATM多重化装置の遅延性能の解析

Weiping ZHAO
National Center for Science Information Systems
学術情報センター　趙　偉平
Shoichiro ASANO
National Center for Science Information Systems
学術情報センター　淺野　正一郎

**ABSTRACT**

ATM technology has been chosen as an efficient and flexible transmission standard for B-ISDN. It is the high flexibility of ATM networks that brings about new problems for network engineers to design and operate the network. Modeling of cell arrival process and its performance analysis for bursty traffics is one of most essential parts of ATM network design. In this paper, we try to study the performance of a multiplexer of bursty traffics. A superposition of sources with bursty traffic is approximated by a two-state MMPP whose parameters can be calculated by statistical values of the original superposition. Thus, the mean cell delay performance of the multiplexed bursty traffic can computed by analyzing two-state MMPP with renewal theory. A superposition of heterogeneous bursty sources is approximated by means of a multi-state MMPP composited by two-state MMPPs each of which represents a superposition of homogeneous sources. The analytic results show that the traffic parameters consisting of only mean and peak bit rates for a general bursty traffic are not sufficient because the performance of superposition of bursty sources are quite different when changing the lasting time of burst while maintaining the same mean and peak bit rates.

**要旨**

広帯域ISDNの転送標準として、ATM技術は有効性および柔軟性を持つ。しかしその反面、ATMネットワークを設計するには、トラヒック制御のためバーストトラヒックの到着過程のモデリングやその性能解析など新しい課題が生じる。そこで、本論文では、多重化したバーストトラヒックの性能解析方法を提案している。本提案では、まず複数均質バーストソースの多重化したトラヒックを2状態MMPPでモデル化し、元トラヒックの統計特性で2-MMPPのパラメータを算出する。複数異質バーストソースの多重化したトラヒックに対して、m-状態MMPPで近似できる。m-MMPPはn($m=2^n$)個の2-MMPPで構成され、各2-MMPPは単一均質ソースの重畳を表す。次に、再生理論を用いてMMPP/D/1キューの平均待ち時間の解析することによって、多重化したトラヒックの遅延性能を究明する。最後に解析結果について検討する。

[Keywords]　ATM network, Analysis, Delay, Multiplexing, Traffic control
[キーワード]　ATM、解析、遅延、多重化、トラヒック制御

## 1 Introduction

The Asynchronous Transfer Mode (ATM) technology has been chosen as an efficient and flexible transmission standard for broadband networks (B-ISDN), which supports a vast variety of services with different bit rates and quality of service (QoS) requirements. In ATM networks all kinds of information including voice, data and video are trans-

mitted in small fixed-length cells. It is the high flexibility of ATM networks that brings about new problems for network engineers to design and operate the network [1].

Traffic control at different level is necessary to avoid possible congestion at each network node and to guarantee the required QoS of multimedia since ATM network adopts the statistical multiplexing technique. Researches on traffic control have been studied widely and many control methods have been proposed [2] [3] [5] [4] [6].

Meanwhile, modeling of cell arrival process and its analysis are essential for designing traffic control methods, especially for designing source policing methods. As well known, the arrival process of cells from multimedia source in ATM environment is not a random one and the analysis based on Poisson process is no longer suitable [7]. Therefore, a lot of researches have been devoted to studying new models for bursty traffic, such as interrupted Poisson process (IPP) [9], Markov modulated Poisson process (MMPP) [8]. For packetized voice sources, each of which is a talk-silence process, a Markov modulated Poisson process (MMPP) is reported to be a good approximation [8] [10]. For videotelephone and videoconferencing information compressed by interframe encoding, auto regressive (AR) is reported to be a suitable model to approximate the performance of their multiplexings [11] [12]. However, it is difficult to find a common model to approximate general bursty traffics.

Long term time-average cell loss probability and cell delay are two important factors adopted as measures of QoS. We choose cell delay as the QoS for our study.

In this paper, we try to study the performance of a multiplexer of bursty traffics. A superposition of sources with bursty traffic is approximated by a two-state MMPP whose parameters can be calculated by statistical values of the original superposition. Thus, the mean cell delay performance of the multiplexed bursty traffic can computed by analyzing two-state MMPP with renewal theory. A superposition of heterogeneous bursty sources is approximated by means of a multi-state MMPP composited by two-state MMPPs each of which represents a superposition of homogeneous sources. The analytic results show that the traffic parameters consisting of only mean and peak bit rates for a general bursty traffic are not sufficient because the performance of superposition of bursty sources are quite different when changing the lasting time of burst while maintaining the same mean and peak bit rates.

The remainer of the paper is organized as follows. After presenting the cell arrival process of a bursty traffic from which the declaring traffic parameters are derived, we approximate the superposition of homogeneous traffics by a two-state MMPP and the superposition of heterogeneous traffics by a multi-state MMPP in Section 2. The computing methods for solving the parameters of a two-state MMPP and the mean waiting time of a MMPP/D/1 queue are reviewed in Appendices. Section 3 gives the numerical examples showing the relationship between the mean cell delay and number of multiplexed calls for several applications. Finally, Section 4 makes some concluding remarks.

## 2 Analysis Model of Traffic

### 2.1 Cell Arrive Mode

An ATM node multiplexes traffics from many sources into a link. We assume that the cell stream from each source follows the following periodic manner as shown in Figure 1, where $T_0$ is the bursty interval during which $X_0$ cells arrive at the peak bit rate $X_0/T_0$, and the rest $(X-X_0)$ cells arrive in

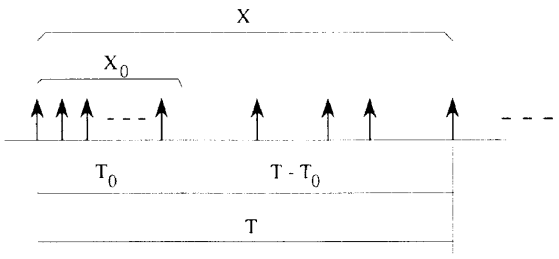



Figure 1 Cell arrival process.

random with mean arrival rate $(X-X_0)/(T-T_0)$ during the interval $(T-T_0)$. The interarrival time of cells arriving in random assumed to follow the exponential distribution. The peak bit rate of such a traffic can easily be obtained by $X_0/T_0$ and the mean bit rate by $X/T(\leqq X_0/T_0)$. Then we represent the traffic by its parameter set $\{T_0, X_0, T, X\}$ with which the user can declare his traffic to the network.

The interarrival time distribution for a single $\{T_0, X_0, T, X\}$ source modeled in Fig. 1 is obtained by

$$F(t)=AU(t-T_m)+B(1-e^{-\lambda_1 t}) \quad (1)$$

where $A=X_0/X$ is the proportion of the number of cells arriving at peak bit rate to the total number of cells; $B=(X-X_0)/X$ is the proportion of the number of cells arriving in random to the total number of cells; $\lambda_1=(X-X_0)/(T-T_0)$ is arrival rate of cells arriving in random; $T_m=T_0/X_0$ is the interval of cells arriving at peak bit rate; $U(t)$ is a step function. The Laplace transform of $F(t)$ is given by

$$F^*(s)=\int_0^\infty e^{-st}dF(t)=Ae^{-sT_m}+B\lambda_1(s+\lambda_1) \quad (2)$$

with the mean cell arrival rate

$$\lambda=-1/F^{*1}(0)=1/(AT_m+B/\lambda_1)=X/T \quad (3)$$

### 2.2 Superposition of Homogeneous Sources

For a superposition of a finite population of $n$ homogeneous sources each of which is declared with $\{T_0, X_0, T, X\}$, we use a two-state MMPP to approximate it. A two-state MMPP is a two-state Markov chain determined by parameter set $\{\lambda_1, \lambda_2, r_1, r_2\}$, where $r_1^{-1}$ and $r_2^{-1}$ are the mean sojourn times in state 1 and 2, $\lambda_1$ and $\lambda_2$ are the cell arrival rates when the chain is in state 1 and 2 respectively.

The parameters $\{\lambda_1, \lambda_2, r_1, r\}$ of the two-state MMPP can be calculated in the following two steps:

1. To calculate the following four statistical values of the superposition of independent $\{T_0, X_0, T, X\}$ sources:

   - the expected cell arrival rate,
   - the deviation of the number of cell arrivals during a short period,
   - the deviations during a long period,
   - the third central moment of the cell arrivals.

   These four values can be numerically computed by the Laplace-Stieltjes transforms of the cell arrival number's first, second and third moments which can be obtained by formula 2 (refer to Appendix A).

2. To calculate the four parameters of the new MMPP by using the above four statistical values (refer to Appendix B).

The performance such as mean delay $d_{mean}$ of the superposition in an ATM node, which can be modeled as a MMPP/D/1 queue, can be computed by means of renewal theory (refer to Appendix C).

### 2.3 Superposition of Heterogeneous Sources

Because a superposition of heterogeneous sources can be consider as a superposition of the aggregates each of which is a superposition of sources of the same type, we can approximate the superposition of heterogeneous sources by means of a multi-state MMPP which is the superposition of two-state MMPPs each of which corresponds to an aggregate of sources of the same type.

For example, let us consider a superposition of sources of two classes. The aggregate of sources of class one is modeled as a two-state MMPP denoted 2-MMPP' with parameter set $\{\lambda_1', \lambda_2', r_1', r_2'\}$, and the aggregate of sources of class two is modeled as another two-state MMPP denoted 2-MMPP" with parameter set $\{\lambda_1'', \lambda_2'', r_1'', r_2''\}$. Then, the superposition of sources of two classes can be modeled as a four-state MMPP denoted 4-MMPP which is the superposition of 2-MMPP' and 2-MMPP". Let $E_j'$

be state $j$ of 2-MMPP' ($j=1, 2$), *$E''_j$ be state $j$ of 2-MMPP''* ($j=1, 2$), $E_j$ be state $j$ of 4-MMPP ($j=1, 2, 3, 4$). Assuming that

$$E_1=E_1'+E_1'', \qquad E_2=E_1'+E_2'',$$
$$E_3=E_2'+E_1'', \qquad E_4=E_2'+E_2''.$$

Then the cell arrival rate $\lambda_j$ with the chain being in state $j$ can be obtained by

$$\lambda_1=\lambda_1'+\lambda_1'', \qquad \lambda_2=\lambda_1'+\lambda_2'',$$
$$\lambda_3=\lambda_2'+\lambda_1'', \qquad \lambda_4=\lambda_2'+\lambda_2''. \tag{4}$$

The infinitesimal generator $Q$ of the chain of 4-MMPP is given by

$$Q=\begin{bmatrix} -(r_1'+r_1'') & -r_1'' & r_1' & 0 \\ r_2'' & -(r_1'+r_2'') & 0 & r_1' \\ r_2' & 0 & -(r_2'+r_1'') & r_1'' \\ 0 & r_2' & r_2'' & -(r_2'+r_2'') \end{bmatrix}, \tag{5}$$

and the stationary distribution $\pi$ for 4-MMPP can be calculated by solving the equilibrium equation in matrix notation

$$\pi Q=0, \quad and, \quad \pi e=1. \tag{6}$$

For the superposition of sources of more than two classes, say $M$ classes, an analysis model $2^M$-state MMPP can be obtained by superposing $M$ two-state MMPPs each of which models the aggregate of sources belonging to one class, and the generator $Q$ and the stationary distribution $\pi$ of the $2^M$-state MMPP can be derived in the same way.

## 3 Numerical Examples and Discussions

Let us consider an application of videoconferencing. The source image information is compressed and multiplexed into a link. We assume that the size of buffer of multiplexer is infinite, the length of cell is 53 bytes, the peak bit rate of compressed source is $4.5Mb/s$, the mean bit rate is $1.5Mb/s$, the bandwidth of link is $156Mb/s$ which result in a fixed cell service time $2.718\mu s$, the parameter $T$ is set to be equal to the frame time $33.33ms$. Therefore, the parameter set $\{T_0, X_0, T, X\}$ will be $\{118\times y\times T_m, 118\times y, 33.33, 118\}$, where the time unit is millisecond, the time interval of cells arriving at peak bit rate $T_m$ is $0.0942ms$ and $y=X_0/X$ called bursty factor is the proportion of the number of cells arriving at peak bit rate.

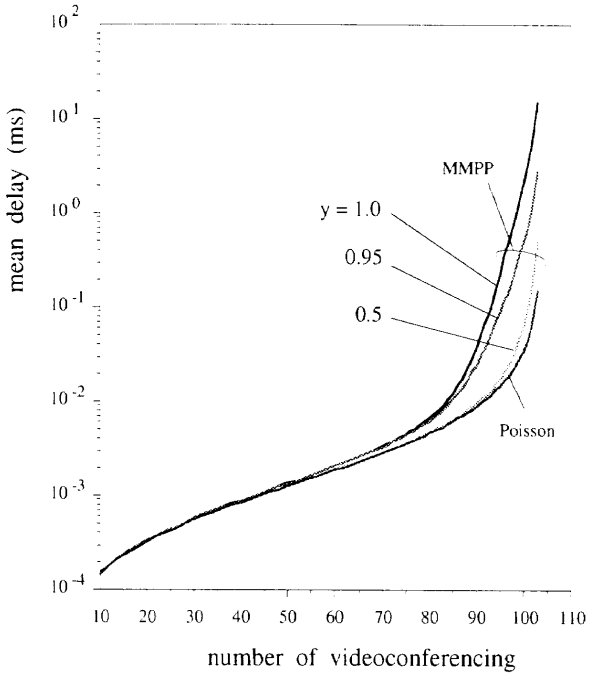



Figure 2 Mean delay for a superposition of videoconferencings.

Figure 2 shows the relation between the mean delay and the number of homogeneous videoconferencing sources with the different bursty factors. Also shown in the figure is the result of the Poisson arrivals. The difference among the performance of the aggregates of bursty traffics and that of Poisson traffic is small when the number of sources is below 60 corresponding to a link utilization of 0.58, which indicates that the burstness of bursty traffics can be absorbed by multiplexing and a bursty traffic can be approximated with a Poissonian traffic. Over 60 sources the difference among the aggregates becomes larger. We note that the mean delay is sensitive to the bursty factor $y$. The performance of the aggregate with $y=1.0$, which means all cells arrive at peak bit rate and the process become an

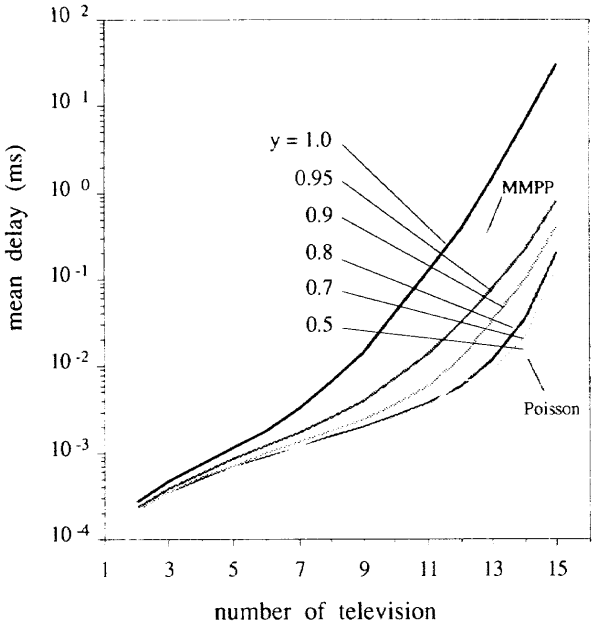



Figure 3　Mean delay for a superposition of televisions.

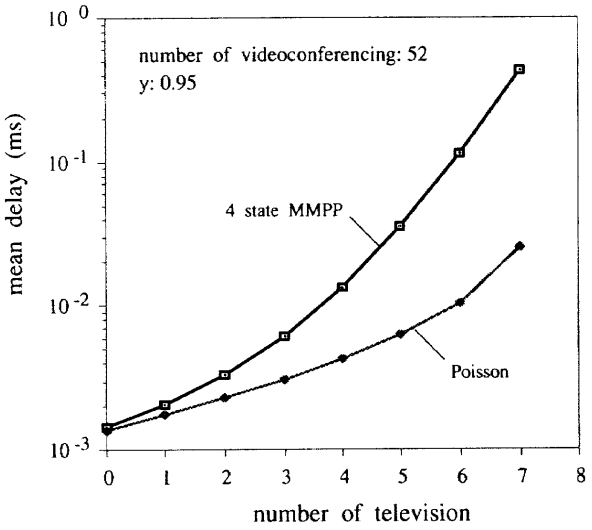



Figure 4　Mean delay for a superposition of videoconferencings (fixed) and TVs.

ON-OFF one, is quite different from that of the aggregate with $y=0.5$, although their peak bit rates and mean bit rates are the same individually. It seems insufficient to describe the burstness of a general bursty traffic only by its peak and mean bit rates.

Next, we pick up an application of television. The peak and mean bit rates of television source are assumed to be $45Mb/s$ and $10Mb/s$ respectively. The link bandwidth is also set to be $156Mb/s$, $T$ be $33.33ms$. Thus, the parameter set $\{T_0, X_0, T, X\}$ will be $\{786\times y\times T_m, 786\times y, 33.33, 786\}$, where $T_m$ is $0.00942ms$. The mean delay as a function of the number of television sources with the different bursty factors is shown in Figure 3.

The difference among the performance of the different aggregates can be seen even at a low link utilization. Comparing with Fig. 2, we can find that the effect of superposition of bursty traffics gets greater when the ratio of link capacity to call bandwidth gets smaller. For example, for a link utilization of 0.45, the ratio of the mean delay of superposition of television with bursty factor 1.0 to that of Poissonian traffic is 2.879 while the responding ratio of videoconferencing is 1.061. Therefore, the effect of superposition of bursty traffic is influenced not only by the bursty characteristics of traffic but also by the link capacity.

Finally, we observe an application of the superposition of above videoconferencing and television traffics. Figure 4 shows the mean delay of mixed cells of both videoconferencing and television traffics. Here we fix the number of videoconferencings at 52 which occupies half of the link capacity and vary the number of televisions.

## 4 Conclusions

In the papaer, we approximated the superposition of homogeneous bursty sources each of which is defined by the parameter set $\{T_0, X_0, T, X\}$ by a two-state MMPP, and approximated the superposition of heterogeneous bursty sources by means of a multi-state MMPP composited by two-state MMPPs each of which represents a superposition of homogeneous sources. It was shown that the traffic parameters consisting of only mean and peak bit rates for a general bursty traffic are not sufficient because the performance of superposition of bursty sources are quite different when changing the lasting time of burst while maintaining the same mean and peak bit rates.

The evaluation of this analysis model by simula-

tion and the cell loss performance analysis of the system with finite buffer are study topics in the future.

### Appendix A Analysis of the statistical characteristics

In this appendix, we use renewal theory to evaluate the mean, variance-mean ratio, and third central moment of cell arrivals in a time interval for a superposition $\{T_0, X_0, T, X\}$ sources.

Because the cell arrival process of a $\{T_0, X_0, T, X\}$ source is a renewal process, the moments of the number of arrivals in an interval can be studied by renewal theory. Let $N(0, t)$ denote the number of arrivals of a renewal process in the interval $(0, t)$, let

$$M_r(t)=E[N^r(0,t)],$$
$$M_r^*(s)=L[M_r(t)]$$

where $L(\cdot)$ denote the Laplace transform. Then it is known [14] that

$$M_1^*(s)=\lambda/s^2 \tag{7}$$

$$M_2^*(s)=\frac{\lambda}{s^2}\left(\frac{1+F^*(s)}{1-F^*(s)}\right) \tag{8}$$

$$M_3^*(s)=\frac{\lambda}{s^2}\left(\frac{1+4F^*(s)+(F^*(s))^2}{(1-F^*(s))^2}\right) \tag{9}$$

where $F^*(s)$ is the Laplace transform of the inter-arrival time distribution and $\lambda$ is the mean arrival rate. It is also known [15] that the variance-mean ratio satisfies

$$\lim_{t\to\infty}\frac{var(N(0,t))}{M_1(t)}=\frac{var(X)}{E^2(X)}.$$

Applying these results to Formula 2, we obtain

$$M_1(t)=tX/T \tag{10}$$

$$\lim_{t\to\infty}\frac{var(N(0,t))}{M_1(t)}=\frac{C^2}{A}+\frac{2A}{A-1}(1-C)^2-1, \tag{11}$$

where $A=X_0/X$ and $C=T_0/T$. The second and third moments of the number of arrivals in a finite time interval can be obtained by numerical transform inversion [16] [17] [13] of $M_2^*(s)$ and $M_3^*(s)$ at the desired time.

For a superposition of n identical independent $\{T_0, X_0, T, X\}$, let $N_i(0, t)$ denote the number of cell arrivals during $(0, t)$ from the $i$th source. Then the number of cell arrivals for the superposition is given by

$$N^S(0,t)=\sum_{i=0}^{n}N_i(0,t).$$

Clearly,

$$M_1^S(t)=E[N^S(0,t)]=nM_1(t) \tag{12}$$

and the variance-mean ratio is given by

$$\frac{var[N^S(o,t)]}{E[N^S(0,t)]}=\frac{var[N(0,t)]}{E[N(0,t)]} \tag{13}$$

The third central moment of the superposition process

$$\mu_3^S(0,t)=E[N^S(0,t)-E(N^S(0,t)]^3$$

can be reduced to

$$\mu_3^S(0,t)=n[M_3(t)-3M_2(t)M_1(t)+2M_1^3(t)] \tag{14}$$

where $M_2(t)$ and $M_3(t)$ are obtained from Laplace transform inversion (8) and (9) of at the desired time, and $M_1(t)$ is obtained from (12).

As a conclusion, the mean cell arrivals in $(0, t)$ for the superposition is obtained by (12), the variance-mean ratio by (13), the third central moment by (14) and the variance-mean ratio over an infinite time interval by (11).

### Appendix B Analysis of the parameter of two-state MMPP

In this appendix, we review the computing method for solving the parameters of a 2-state

MMPP by using the statistical characteristics of the superposition approximated by this two-state MMPP. The results presented here were already published [14] [8].

Let $r_a$ denote the cell arrival rate for the superposition, $d_t$ denote the deviation of the number of cell arrivals of the superposition during $(0,t)$, $d_\infty$ denote the deviation of the number of cell arrivals of the superposition during $(0,\infty)$, and $\mu^{(3)}(0,t)$ denote the third central moment of the number of arrivals of the superposition during $(0,t)$. From Appendix A, $r_a$, $d_t$, $d_\infty$ and $\mu^{(3)}(0,t)$ are given by

$$r_a=\frac{nM_1(t)}{t}, \tag{15}$$

$$d_t=\frac{M_2(t)-M_1^2(t)}{M_1(t)}, \tag{16}$$

$$d_\infty=\frac{F^{*\prime\prime}(0)-(F^{*\prime}(0))^2}{(F^{*\prime}(0))^2}, \tag{17}$$

$$\mu^{(3)}(0,t)=n[M_3(t)-3M_2(t)M_1(t)+2M_1^3(t)]. \tag{18}$$

On the other hand, the probability generating function of the number of arrivals in a time interval of length $t$ for a MMPP is given by

$$g(z,t)=\pi exp\{[Q+(z-1)\Lambda]t\}e \tag{19}$$

where $\pi$ denotes the equilibrium probability vector for the Markov chain, $Q$ denotes the generator for the chain. For the 2-state MMP,

$$\pi=\frac{1}{r_1+r_2}(r_1,\ r_2),\quad e=(1,1)^T,$$
$$Q=\begin{bmatrix}-r_1 & r_1\\ r_2 & -r_2\end{bmatrix},\quad \Lambda=\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}.$$

Defining $N_t$ as the number of arrivals from the MMPP in $(0,t)$, $\bar{N}_t(t)$ as the mean value of $N_t$, $var\ (N_t)/\bar{N}_t$ as the deviation of $N_t$ and $E\ [((N_t)/\bar{N}_t)]$ as the third central moment of $N_t$, we will have

$$\bar{N}_t=E[N_t]=\frac{\lambda_1r_2+\lambda_2r_1}{r_1+r_2}t, \tag{20}$$

$$\frac{var(N_t)}{\bar{N}_t}=1+\frac{2(\lambda_1-\lambda_2)^2r_1r_2}{(r_1+r_2)^2(\lambda_1r_2+\lambda_2r_1)}-\frac{2(\lambda_1-\lambda_2)^2r_1r_2}{(r_1+r_2)^3(\lambda_1r_2+\lambda_2r_1)t}\cdot(1-e^{-(r_1+r_2)t}), \tag{21}$$

$$E[(N_t-\bar{N}_t)^3]=g^{(3)}(1,t)-3\bar{N}_t(\bar{N}_t-1)\frac{var(N_t)}{\bar{N}_t}-\bar{N}_t(\bar{N}_t-1)(\bar{N}_t-2), \tag{22}$$

where $g^{(3)}(1,t)$ is given by

$$g^{(3)}(1,t)=\frac{6}{r_1+r_2}\left(\frac{A11}{6}t^3+\frac{A_{21}}{2}t^2+A_{31}t+A_{21}te^{-(r_1+r_2)t}+A_{41}(1-e^{-(r_1+r_2)t})\right) \tag{23}$$

and $A_{ij}s$ are given by

$$A_{11}=\frac{(\lambda_1r_2+\lambda_2r_1)^3}{(r_1+r_2)^2},$$
$$A_{21}=\frac{2r_1r_2(\lambda_1-\lambda_2)^2\ (\lambda_1r_2+\lambda_2r_1)}{(r_1+r_2)^3},$$
$$A_{31}=\frac{r_1r_2(\lambda_1-\lambda_2)^2}{(r_1+r_2)^4}\ [\lambda_1r_1+\lambda_2r_2-2(\lambda_1r_2+\lambda_2r_1)],$$
$$A_{41}=\frac{-2r_1r_2(\lambda_1-\lambda_2)^3(r_1-r_2)}{(r_1+r_2)^5},$$
$$A_{12}=\frac{r_1r_2(\lambda_1-\lambda_2)^2(\lambda_1r_1+\lambda_2r_2)}{(r_1+r_2)^4}$$

Solving Equation (15)～(23), the parameters of the two-state MMPP cab be obtained by

$$r_1=\frac{d}{2}\left(1+\frac{1}{\sqrt{4e+1}}\right), \tag{24}$$

$$r_2=d-r_1, \tag{25}$$

$$\lambda_1=\frac{K}{r_1-r_2}+\lambda_2, \tag{26}$$

$$\lambda_2=\left(\frac{r_a d}{r_2}-\frac{K}{r_1-r_2}\right)\left(\frac{r_2}{r_1+r_2}\right) \tag{27}$$

where $d=(r_1+r_2)$ is determined by

$$d=\frac{1}{t_1}\left(\frac{d_\infty-1}{d_\infty-d_{t_1}}\right)(1-e^{-dt_1}),$$

$K=(\lambda_1-\lambda_2)(r_1-r_2)$ is obtained by solving

$$\begin{aligned}g^{(3)}(1,t_2)=&r_a^3t_2^3+3r_a^2(d_\infty-1)t_2^2\\&+\frac{3r_a(_\infty-1)}{d}\cdot\left[\frac{K}{d}-r_a\right]t_2\\&+\frac{3r_a(_\infty-1)}{d^2}(K+r_ad)\,t_2e^{-dt_2}\\&-\frac{6r_a(_\infty-1)}{d^3}\cdot K(1-e^{-dt_2}),\end{aligned}$$

$t_1$ and $t_2$ can freely be chosen and

$$e=\frac{(d_\infty-1)\,r_ad^3}{2K^2}.$$

## Appendix C Analysis of mean delay of MMPP/G/1

The computing method for mean delay of a MMPP/G/1 system, which was already published [14] [8], is summarized in this appendix.

Assuming that arrivals from a m-state MMPP with infinitesimal generator $\boldsymbol{Q}$ join a first-in-first-out (FIFO) single-server queue, and the service times are independent and identically distributed with distribution function $\bar{H}(\cdot)$. Let the $i$th moment of the service times be $\mu^{(i)}$ and $\Lambda$ be a diagonal matrix with the element $\lambda_j$ along the diagonal . Let the vector of distribution functions $\vec{\boldsymbol{W}}(x)$ have components $\widetilde{W}_j(x)$ where $\widetilde{W}_j(x)$ is the joint probability that at an arbitrary time the MMPP is in phase $j$ and that a cell's waiting time is than or equal to $x$ before entering service. Then, the waiting time distribution can be given by $\vec{\boldsymbol{W}}(x)\boldsymbol{e}$. Let $\boldsymbol{\pi}$ be the stationary distribution of MMPP and $\boldsymbol{\lambda}$ be the vector with $j$th component $\lambda_j$. The algorithm for computing the delay distribution and the first two moments of the waiting time can be obtained by:

$$\begin{aligned}E(W)=&\frac{1}{2(1-\rho)}[2\rho+\mu^{(2)}\boldsymbol{\pi\lambda}-2\mu^{(1)}\\&(\boldsymbol{y}_0+\mu^{(1)}\pi\boldsymbol{\Lambda})(\boldsymbol{R}+\boldsymbol{e\pi})^{-1}\boldsymbol{\lambda}]\end{aligned} \tag{28}$$

$$\begin{aligned}E(W^2)=&\frac{1}{3(1-\rho)}[3\mu^{(1)}[2\mu^{(1)}\boldsymbol{W}'(0)\boldsymbol{\Lambda}\\&-2\boldsymbol{W}'(0)-\mu^{(2)}\boldsymbol{\pi\Lambda}](\boldsymbol{R}+\boldsymbol{e\pi})^{-1}\lambda\\&-3\mu^{(2)}\boldsymbol{W}'(\boldsymbol{0})\lambda+\mu^{(3)}\boldsymbol{\pi\lambda}]\end{aligned} \tag{29}$$

where

$\rho$: $\boldsymbol{\pi\lambda}\mu^{(1)}$(the traffic intensity),

$\boldsymbol{W}'(\boldsymbol{0})$: $\mu^{(1)}\boldsymbol{\pi\Lambda}(\boldsymbol{R}+\boldsymbol{e\pi})^{-1}+\boldsymbol{y}_0(\boldsymbol{R}+\boldsymbol{e\pi})^{-1}-E(W)\boldsymbol{\pi}-\boldsymbol{\pi}$,

$(y_0)_j$: the stationary probability of the system being empty and the phase of the Markov chain beibg in phase $j$ at an arbitrary point time

$\boldsymbol{y}_0$: the vector composed by $(y_0)_j$.

Further, $\boldsymbol{y}_0$ can be computed by the following procedure:

- To compute the stochastic matrix $G$ with the $(i,j)$ component being the probability that a busy period starting with the MMPP in phase $i$ ends in phase $j$ by the iterative procedure given by

$$\begin{aligned}\boldsymbol{H}_{n+1,k}=&[\boldsymbol{I}+\theta^{-1}(\boldsymbol{R}-\boldsymbol{\Lambda})]\boldsymbol{H}_{n,k}\\&+\theta^{-1}\boldsymbol{\Lambda}\boldsymbol{H}_{n,k}\boldsymbol{G}_k,\\&n=0,\ 1,\ 2,\ \cdots\cdots,\\\boldsymbol{G}_{k+1}=&\sum_{n=0}^{\infty}\gamma_n\boldsymbol{H}_{n,k}\\\boldsymbol{G}_0=&\boldsymbol{0},\qquad k=0,\ 1,\ 2,\ \cdots\cdots,\end{aligned}$$

where $\boldsymbol{H}_{0,k}=\boldsymbol{I}$, $\theta=max(\lambda_j-R_{jj})$ and $\lambda_0=e^{-\theta d}$, $\gamma_n=(\theta d/n)\gamma_{n-1}$, $d$ is mass of the $\widetilde{\boldsymbol{H}}(\cdot)$.

- To compute the stationary probability distribution of the Markov chain with transition matrix $\boldsymbol{G}$ from

$$g=(g_1,g_2)=\frac{1}{G_{12}+G_{21}}(G_{21},G_{12}).$$

- To compute $\boldsymbol{A}=\int_0^\infty e^{Rt}\,d\widetilde{\boldsymbol{H}}(\cdot)$. $A_{ij}$ is the probability that a service time ends with the MMPP in phase $j$ given that the service began in phase $i$. In the two-state case we have that $\boldsymbol{A}$ is given by

$$\boldsymbol{A}=\boldsymbol{e\pi}-\frac{H(r_1+r_2)}{r_1+r_2}\boldsymbol{R}$$

where $H(s)$ is the LST of $\widetilde{\boldsymbol{H}}(\cdot)$.

- To compute $\boldsymbol{U}=(\boldsymbol{\Lambda}-\boldsymbol{R})^{-1}\boldsymbol{\Lambda}$ *where* $\boldsymbol{U}$ keeps track of the phase during an idle period. That is, $U_{ij}$ is the probability that a busy period arrives with the MMPP in phase $j$ given that the last departure from the previous busy period departed with MMPP in phase $i$.

- To compute $\boldsymbol{\beta}=\mu^{(1)}(\boldsymbol{\pi\lambda})\boldsymbol{e}+(\boldsymbol{R}+\boldsymbol{e\pi})^{-1}(\boldsymbol{A}-\boldsymbol{I})\boldsymbol{\lambda}$ where $\beta_j$ is the expected number of arrivals during a service that began in phase $j$.

- To compute $\boldsymbol{\mu}=(\boldsymbol{I}-\boldsymbol{G}+\boldsymbol{eg})[\boldsymbol{I}-\boldsymbol{A}+\boldsymbol{eg}-\boldsymbol{\beta g}]^{-1}\boldsymbol{e}$ where $\mu_j$ is the expected number of departures during a busy period that began in phase $j$.

- To compute $\boldsymbol{d}$ such that $\boldsymbol{dUG}=\boldsymbol{d}$, $\boldsymbol{de}=1$. It is clear that $d_j$ is the stationary probability of ending a busy period in phase $j$.

- To compute $\boldsymbol{x}_0=(\boldsymbol{dU\mu})^{-1}\boldsymbol{d}$. $(x_0)_j$ is the stationary probability that a departure leaves the system empty with the MMPP in phase $j$. This is just the stationary probability of being in phase $j$ at successive epochs which leave the system empty divided by the expected number of departures between such epochs.

- To compute $\boldsymbol{y}_0=(\boldsymbol{\pi\lambda})\boldsymbol{x}_0(\boldsymbol{\Lambda}-\boldsymbol{R})^{-1}$.

## References


[1] Minzer,S,E. "Broadband ISDN and asynchronous transfer mode (ATM)", *IEEE Commun. Mag.*, Vol.27, No.9, Sept. 1989.

[2] Jain,R., "Congestion control and traffic management in ATM networks: recent advances and a survey", *Computer and ISDN Systems*, Vol.28, No.13, pp.1723-1737, 1996.

[3] Nogami,S., "Quality control at the cell level and its characteristics in the ATM network", *Trans. IEICE*, J73-B-I, No9, pp.671-680, 1990.

[4] Drury,D,M., "ATM traffic management and the impact of ATM switch design", *Computer and ISDN Systems*, Vol.28, No.4, pp.471-480, 1996.

[5] Lee,T.; Lai,K.; Duann,S., "Design of a real-time call admission controller for ATM networks", *IEEE/ACM Trans.* on Networking, Vol.4, No.5, pp.758-765, 1996.

[6] Zhao,W.; Runggeratigul,S.; Asano,S., "An admission procedure based on the virtual bandwidth of calls", *1996 International Conference on Communication Technology (ICCT'96) Proceedings*. pp.1-4, Beijing, May 1996.

[7] Paxson,V.; Floyd,S., "Wide-area traffic: the failure of Poisson modeling", *Proceedings of SIGCOMM'94*, pp.257-268, London, August 1994.

[8] Heffes,H.; Lucantoni,D,M., "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance", *IEEE J. Select. Areas Commun.*, SAC-4, No.6, pp.856-868, 1986.

[9] Kuczura.A., "The interrupted Poisson process as an overflow process", *Bell Syst. Tech. J.*, Vol.52, No.3, pp.437-448, 1973.

[10] Nagarajan,R.; Kurose,J,F.; Towsley,D., "Approximation techniques for computing packet loss in finite-buffered voice multiplexers", *IEEE J. Select. Areas Commun.*, SAC

-9, No.3, pp.369-377, 1991.

[11] B.Maglaris et al.: "Performance analysis of statistical multiplexing for packet video sources", *IEEE Trans. on Commun.*, COM-36, No7, pp.834-843, (1988).

[12] Nomura,M.; Fujii,T.; Ohta,N., "Basic characteristics of variable rate interframe video coding", *Trans. IEICE*, J71-B, No.10, pp.1110-1118, 1988.

[13] Zhao,W., "Researches on control of qualities of service in ATM networks", dissertation, Tokyo University, December 1992.

[14] Takacs,L., "Introduction to the theory of queues", New York: Oxford Univ. Press, 1962.

[15] Cox,D,R.; Lewis,P,A,W., "The statistical analysis of series of events", London, Methuen, 1966.

[16] Jagerman,D,L., "An inversion technique for the Laplace transform with application to approximation", *Bell System Technical J.*, Vol.57, No.3, pp.669-710, 1978.

[17] Jagerman,D,L., "An inversion technique for the Laplace transform", *Bell System Technical J.*, Vol.61, No.8, pp.1995-2002, 1982.

学術情報センター紀要　第10号(1998)

研究論文

# 多重化離散時間発生バーストパケット入力待ち行列システムの再生近似による性能解析法

# A Performance Analysis Method by a Renewal Approximation for the Queueing System with the Input of Multiplexed Burst Packets by a Discrete Time Generation

学術情報センター　阿部　俊二
Shunji ABE
National Center for Science Information Systems
学術情報センター　淺野　正一郎
Shoichiro ASANO
National Center for Science Information Systems

**要旨**

本論文は、ON-OFFバーストに基づくバーストパケットの多重化流を入力とする待ち行列システムの性能を再生近似により解析する手法に関係する。この解析手法の一つとして、著者の一人は、既に待ち行列システムの平均残余稼働時間$R_b$とパケット発生数に関する分散指数$I(t)$を用いて多重化トラヒックを再生近似する性能解析方法を提案した。しかし、そこで扱われたバーストパケットソースは、ON区間長及びOFF区間長が共に指数分布で、且つON区間中はポアソン分布に基づく連続時間パケット発生のみであった。本論文では、既に提案した解析手法を基本に、より多くの種類のバースト発生パターンのモデル化を念頭に置き、ON区間でのパケットがある一定の時間間隔で離散的に発生し、OFF区間長が$k$-アーラン分布、超指数分布に従うバーストモデルに拡張する。拡張したバーストモデルの$I(t)$が非常に複雑であり、$t$が大きい場合に膨大な計算時間が必要で実用的計算に不向きであることから、再生過程をなす'Doubly Stochastic Poisson Process'の分散指数を適用した$I(t)$の近似法を提案する。平均待ち時間の計算機シミュレーション結果と近似$I(t)$を用いた平均待ち時間の比較から本提案法の有効性を示す。

**ABSTRACT**

This paper is concerned with a renewal process approximation for a queueing system at which a multiplexed ON-OFF burst traffic is offered. As one of the renewal approximation, one of authors proposed the method in which the multiplexed burst traffic was approximated by using the mean residual life time for the busy period of the queue and the index dispersion for counts(IDC) $I(t)$ of ON-OFF burst. The ON-OFF burst model in the method was handled as each ON-OFF period was an Exponential distribution and packets on the ON-period were generated with continuous time according to a Poisson process. In order to model for many kind of burst patterns in this paper, the ON-OFF burst model is extended. As the extended model, we consider that packets on the ON-period are generated with the discrete time of a fixed $T$, and furthermore $k$-Erlang and Hyperexponential are assumed as the OFF-period distribution. The $I(t)$ of the extended model is very complicated and its computation time is very large for large $t$. To reduce computation time, a new approximation for the $I(t)$ by using IDC of Doubly Stochastic Poisson Process is proposed. The validity of the proposed method is shown by comparison with results of computer simulation for the mean waiting time of the queue.

[キーワード]　バーストトラヒック、多重化バースト、再生過程近似、ATMトラヒック
[Keywords]　Burst Traffic, Multiplexed Burst Traffic, Renewal Approximation, ATM Traffic

## 1　まえがき

ATM技術を基本としたLAN/WAN等に代表されるバースト情報の通信を実現する網構築およびその設備設計や、品質保証のためのトラヒック制御方式の実現には、バーストトラヒックが混在した場合の遅延、廃棄、スループットなどの性能評価が重要である。バーストトラヒックを混在(多重化)させた場合の性能解析/評価法は、これまで多くの手法が提案されている。本論文は、特に、多重化バーストトラヒックを再生近似して性能解析を行う方法に関係する。

再生近似以外の近似法として、MMPP(Markov Modulated Poisson Process)[4]が良く知られている。MMPP近似では、少なくとも4つのパラメタの決定が必要となる。これに対して、再生近似では、高々3つのパラメタの決定で済む。本論文では、パラメタ決定の数の複雑度を考慮し、再生近似によるアプローチを行うものである。

再生近似による性能解析法として、(平均)発生率と平方変動係数を用いたツー・パラメタによるQNA[7]が良く知られている。著者の一人は、多重化バーストトラヒックの性能解析に関し、発生率、平方変動係数、歪み度を用いたスリー・パラメタによる再生近似法を提案し、その近似精度が、QNAを多重化バーストトラヒックの性能解析に適用した場合の近似精度より高いことを示した[1]。平方変動係数と歪み度の決定に、シングルバーストソースが$(0,t]$に発生するパケット数の分散指数(IDC:Index of Dispersion for Counts)$I(t)$と一つのバーストソースが無限待ち室の待ち行列に加わった時の平均残余稼働時間($R_b$)を用いることで、近似の高精度化を実現した。

文献[1]の扱う多重化されるバーストソースは、ONおよびOFF区間長分布をそれぞれ指数分布とし、ON区間中のパケット発生を連続時間に基づいたポアソン分布に従うバーストソースの多重のみを扱っており†、扱うことのできるバーストソースの範囲が狭くなっている。将来、どの様なサービスが出現し、その情報発生パターンがどの様なるかは、予想はできないが、極力種々の情報発生をモデル化できるように準備する必要があると考える。

本論文では、再生近似手法は文献[1]を基本に、バーストソースが再生過程で扱える範囲内で、バーストソースのモデル化の範囲を極力広げることを考慮して以下のような拡張を行う。

(1) ON区間中のパケット発生を一定時間間隔に基づく離散発生とする(ただし、ON区間中に発生するパケット数は幾何分布に従うものとする)。

(2) OFF区間長分布として、$k$-アーラン分布、超指数分布を扱う。

例えば、ATM通信でのバーストトラヒックを、情報を運ぶパケット(セル)が発生するON区間とセルが全く発生しないOFF区間とするON-OFFモデルとして扱う場合を考えると、ATMのセル長が固定であることからON区間でのセルは、セル化時間$T$を一定時間間隔とした離散発生となる。このことから、拡張(1)によってATM通信のバーストトラヒックにより近いモデル化が可能となる。

拡張(2)は、OFF区間長分布の指数分布から$k$-アーラン分布と超指数分布への拡張であるが、これにより、OFF区間長分布として、より一般的な分布の取り扱いが可能となる。これは、OFF区間長分布の平方変動係数($C_z^2$)の大きさにより、$k$-アーラン分布($0<C_z^2<1$)、指数分布($C_z^2=1$)、超指数分布($C_z^2>1$)にそれぞれモーメントマッチして、OFF区間分布を近似できるからである。

以上のように拡張したバーストソースを多重化した時の性能評価を文献[1]に従い行う場合、拡張したバーストソースの分散指数$I(t)$と平均残余稼働時間$R_b$を求める必要がある。特に、拡張したバーストソースの$I(t)$は非常に複雑であり、且つ$t$が大きい場合に膨大な計算時間が必要であり、実用的計算には不向きである。そこで、計算時間を削減するために、再生過程をなす特殊な 'Doubly Stochastic Poisson Process(DSPP)' [8]の分散指数を応用した拡張バーストソースの$I(t)$の近似法を提案する。

† このバーストソースモデルは再生過程となる。再生過程をなすソースの多重化過程をまた再生過程で近似して性能解析を行っている。

まず、拡張(1)と(2)を考慮した場合の離散時間発生バーストソースモデルを示し、本モデルの分散指数$I(t)$の厳密解の導出した後、再生過程をなすDSPPの分散指数を用いた離散時間発生バーストソースの$I(t)$の近似式を導く。次に、一つの離散時間発生バーストソースが無限待ち室の待ち行列に加わった場合の平均残余稼働時間をスペクトル分解法(spectrum factorization)により導出する。

近似した分散指数と平均残余稼働時間から、離散時間発生バーストを$n$多重した場合のトラヒック流が待ち行列に加わった場合の平均待ち時間を、文献[1]の方法により求める。この結果と計算機シミュレーション結果との比較を行い、分散指数の近似法の有効性を示す。

## 2　離散時間発生バーストモデル

離散時間発生バーストソースのモデルとして(図1)、固定長のパケットがある決まった一定間隔($T$)で発生し、その後ある分布に従う期間、発生しないと言うパケットの発生パターンを考える。一定間隔で離散的に発生する区間をバーストパケットのON区間とする。また、ON区間でのパケット発生数($k$)は幾何分布に従うものとする。すなわち、$P(k)=qp^{k-1}$とする。ただし、$p$はパケットが発生する確率、$q(=1-p)$は発生しない確率とする。このON区間の後のパケットが発生しない期間をOFF区間とする(特に、本節のモデル化においては、OFF区間分布として一般的分布でも成立するため、OFF区間分布に特別な分布の仮定をしない)。

パケットの発生時間間隔分布を$F(t)$とすると、次式で求まる。

$$F(t)=pD(t,T)+qF_z(t)\bigstar D(t,T) \tag{1}$$

ただし、$D(t,T)$は平均$T$の一定分布(単位分布)、$F_z(t)$は平均$\beta^{-1}$のOFF区間分布、★はたたみ込み積分を示す。また、ON区間長の平均$\alpha^{-1}$は、幾何分布に従うパケット発生数の平均が$1/q$となるので、$\alpha^{-1}=T/q$となる。

パケット発生間隔分布$F(t)$のラプラス変換を$f^*(s)$とすると次式となる。

$$f^*(s)=\{p+qf_z^*(s)\}e^{-sT} \tag{2}$$

ただし、$f_z^*(s)$はOFF区間長分布のラプラス変換である。

パケット発生間隔分布のラプラス変換$f^*(s)$から、平均$\lambda^{-1}$($\lambda$は発生率)、平方変動係数$C_a^2$(分散/平均$^2$)、歪み度$S_k$(3次中心モーメント/分散$^{3/2}$)は以下で求まる。

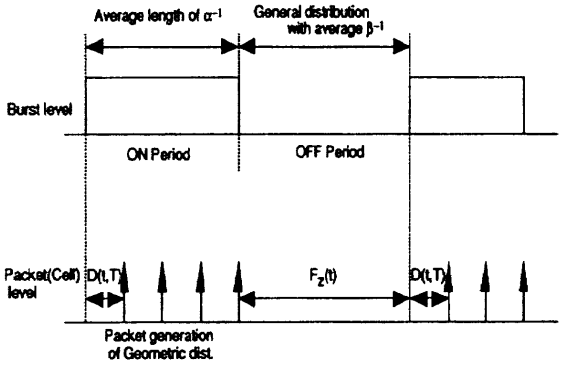



図 1　離散時間発生バーストモデル

$$\left.\begin{aligned}\lambda^{-1}&=T+\frac{q}{\beta}\\ C_a^2&=\frac{(C_z^2+1-q)q}{(q+\beta T)^2}\\ S_k&=\frac{C_z^3S_{kz}+3C_z^2(1-q)+(1-3q+2q^2)}{q^{1/2}(C_z^2+1-q)^{3/2}}\end{aligned}\right\} \tag{3}$$

ただし、$C_z^2$、$S_{kz}$は、$F_z(t)$の平方変動係数と歪み度を表す。

## 3　離散時間発生バーストの分散指数

時間区間(0,t]までに発生するパケット数の平均を$M_1(t)$、2次モーメントを$M_2(t)$、分散を$\sigma_N^2(t)$として、分散指数を$I(t)\equiv\sigma_N^2(t)/M_1(t)$で定義する。各統計量は、パケット発生間隔分布$F(t)$、式(1)、のラプラス変換$f^*(s)$を用いて、文献[4]の次式より求まる。

$$\left.\begin{aligned}M_1(t)&=\lambda t\\ I(t)&=\frac{\sigma_N^2(t)}{M_1(t)}=\frac{M_2(t)-M_1(t)^2}{M_1(t)}\\ M_2(t)&=\mathcal{L}^{-1}\left[\frac{\lambda(1+f^*(s))}{s^2(1-f^*(s))}\right]\end{aligned}\right\} \tag{4}$$

ただし、記号$\mathcal{L}^{-1}$はラプラス逆変換を表す。

式(4)から分散指数$I(t)$を求めるため、$M_2(t)$のラプラス変換$M_2^*(s)$を$f^*(s)$のべき級数展開を行い、式(2)を代入すると次式が得られる。

$$M_2^*(s)=\frac{\lambda}{s^2}+2\lambda\sum_{n=1}^{\infty}e^{-nsT}\sum_{i=0}^{n}{}_nC_ip^{n-i}q^i\frac{f_z^*(s)^i}{s^2} \tag{5}$$

ここで、OFF区間の分布$F_z(t)$として、ラプラス変換$f_z^*(s)$が有理関数で表すことができる分布を仮定し、

$$g_i^*(s)\equiv f_z^*(s)^i/s^2=N(s)^i/\{s^2D(s)^i\} \tag{6}$$

と置き(ただし、$g_0^*(s)=1/s^2$)、$g_i^*(s)$が部分分数に展開できるものとする(部分分数展開は$s^2D(s)^i$のゼロ点から容易に求めることができる)。$g_i^*(s)$の逆ラプラス変

換を$g_i(t)$とすると、$M_2(t)=\mathcal{L}^{-1}[M_2^*(s)]$は以下となる。

$$M_2(t)=\lambda t+2\lambda\sum_{n=1}^{\infty}\sum_{i=0}^{n}{}_nC_i p^{n-i}q^i g_i(t-nT) \quad (7)$$

式(7)と式(4)の関係を用いると分散指数$I(t)$は次式で求まる。

$$I(t)=1-\lambda t+\frac{2}{t}\sum_{n=1}^{\infty}\sum_{i=0}^{n}{}_nC_i p^{n-i}q^i g_i(t-nT) \quad (8)$$

$F_z(t)$を、指数分布、超指数分布、$k$-アーラン分布とした場合の$g_i(t)$を以下に示す。

(1) $F_z(t)$が指数分布の場合

$F_z(t)$が指数分布の場合のラプラス変換は、

$$f_z^*(s)=\beta/(s+\beta) \quad (9)$$

となるので、式(6)の$s^2D(s)$のゼロ点は$s=0$(2重根)、$s=-\beta$($i$重根)となり、$g_i(t)$は以下となる。

$$g_i(t)=\left[t-\frac{i}{\beta}+\sum_{k=1}^{i}\frac{k\beta^{i-k-1}t^{i-k}}{(i-k)!}e^{-\beta t}\right]U(t) \quad (10)$$

ただし、$U(t)$は$t\geq 0$のとき1となるステップ関数である。

(2) $F_z(t)$が超指数分布の場合

$F_z(t)$として超指数分布($H_2$)、$F_z(t)=1-C_1e^{-\beta_1 t}-C_2e^{-\beta_2 t}$(ただし、$C_1+C_2=1$)を考えると、ラプラス変換は以下となる。

$$f_z^*(s)=\{f_z(0)s+\beta r_z\}/\{(s+\beta_1)(s+\beta_2)\} \quad (11)$$

ただし、

$$\left.\begin{aligned} f_z(0)&\equiv C_1\beta_1+C_2\beta_2\\ r_z&\equiv C_1\beta_2+C_2\beta_1\\ \beta&=\beta_1\beta_2/(C_1\beta_2+C_2\beta_1)\end{aligned}\right\} \quad (12)$$

である。したがって、式(6)の分母のゼロ点は、$s=0$(2重根)、$s=-\beta_1$($i$重根)と$s=-\beta_2$($i$重根)なる。これを基に部分分数展開を行い、ラプラス逆変換すると、$g_i(t)$は以下となる。

$$g_i(t)=\left[t-\frac{i}{\beta}+\sum_{j=1}^{i}B_{1j}\frac{t^{i-j}}{(i-j)!}e^{-\beta_1 t}+\sum_{j=1}^{i}B_{2j}\frac{t^{i-j}}{(i-j)!}e^{-\beta_2 t}\right]U(t) \quad (13)$$

ただし、$2\leq j\leq i$、$u=1$のとき$v=2$、$u=2$のとき$v=1$として、

$$\left.\begin{aligned} B_{u1}&=\frac{\{\beta r_z-f_z(0)\beta_u\}^i}{\beta_u^2(\beta_v-\beta_u)^i}\\ B_{uj}&=\frac{1}{(j-1)!}\sum_{r=0}^{j-1}C_r\\ &\times\frac{i!f_z(0)^{j-1-r}\{\beta r_z-f_z(0)\beta_u\}^{i-j+1+r}}{(i-j+1+r)!}\\ &\times\left[\sum_{l=0}^{r}{}_rC_l(-1)^{l-2}\frac{(r-l+1)!(i+l-1)!}{(i-1)!\beta_u^{r-l+2}(\beta_v-\beta_u)^{i+l}}\right]\end{aligned}\right\} \quad (14)$$

(3) $F_z(t)$が$k$-Erlang分布の場合

平均が$\beta^{-1}$で、平方変動係数$C_z^2=k^{-1}$となる$k$-Erlang分布の場合のラプラス変換は以下である。

$$f_z^*(s)=\{\beta k/(\beta k+s)\}^k \quad (15)$$

式(6)の分母のゼロ点は、$s=0$(2重根)、$s=-\beta k$($ki$重根)となるのである。これを考慮して$g_i(t)$を求めると以下となる。

$$g_i(t)=\left[t-\frac{i}{\beta}+\sum_{j=1}^{ik}\frac{j(\beta k)^{ik-j-1}t^{ik-j}}{(ik-j)!}e^{-\beta k t}\right]U(t) \quad (16)$$

### 3.1 分散指数の近似

離散時間発生バーストの任意時間$t$における分散指数$I(t)$を計算する場合、ステップ関数$U(t-nT)$($t\geq nT$のとき、ステップ関数は1となり、それ以外はゼロである)を含むため、$\bar{N}=\lfloor t/T\rfloor$($\lfloor x\rfloor$は、$x$を越えない最大の整数とする)として、$n$に関する和の上限$\infty$を$\bar{N}$に代えて$I(t)$を計算することができる。しかし、$I(t)$の計算量は、$\bar{N}^2/2$のオーダより大きくなり、大きな時間$t$では、計算時間が大きく実用的ではない。そこで、計算時間の削減に着目した分散指数の近似法について示す。

ON区間中幾何分布でパケットが発生し、OFF区間が指数分布に従う場合の分散指数の近似として、超指数分布($H_2$)の分散指数$I_H(t)$を用いた近似法が提案されており、また高い近似精度が得られることも確認されている[2]。

$$\left.\begin{aligned} I_{ap}(t)&=1-\lambda t && (0\leq t\leq T)\\ &=I_H(t-T)-\lambda T && (t>T)\end{aligned}\right\} \quad (17)$$

ただし、

$$\left.\begin{aligned} I_H(t)&=C_{aH}^2-(C_{aH}^2-1)^2/\{2I'_H(0)t\}\\ &\times\{1-Exp[-\frac{2I'_H(0)t}{C_{aH}^2-1}]\}\end{aligned}\right\} \quad (18)$$

で、$I'_H(0)=\{I(2T)-I(T)\}/T$、$C_{aH}^2=C_a^2$で近似する。

ON区間およびOFF区間がそれぞれ指数分布に従い、

ON区間でのパケットの発生過程がポアソン分布の場合のパケット発生過程は、断続ポアソン過程(Interrupted Poisson Process(IPP))となる。IPPのパケット発生間隔分布は、超指数分布に等しく、再生過程であることが良く知られている[6]。したがって、ON区間長が指数分布でポアソン発生する場合の分散指数を式(17)の様に適用すれば、ON区間中幾何分布発生するパケットの分散指数を旨く近似できることが分かる。

そこで、OFF区間が超指数分布や$k$-アーラン分布の場合の離散発生パケットの分散指数の近似についても、ON区間が指数分布、OFF区間が一般分布で、ON区間でのパケット発生がポアソンである確率過程(再生過程をなすDSPPとなる)場合の分散指数で旨く近似できることが期待できる。

### 3.2 $F_z(t)$が超指数分布の場合の近似

まず、ON区間長を平均$\alpha^{-1}$の指数分布、OFF区間長分布($F_z(t)$)を平均$\beta^{-1}$の一般分布、ON区間でのパケット発生を発生率$\nu$のポアソン分布とする。このときの、パケット発生間隔分布の密度関数$a(t)$のラプラス変換を$a^*(s)$とすると、$a^*(s)$は次式で与えられる(文献[8]のP.275)。

$$a^*(s)=\nu/\{s+\nu+\alpha-\alpha f_z^*(s)\} \tag{19}$$

$f_z^*(s)$として超指数分布のラプラス変換を用いて、式(4)の$f^*(s)$の代わりに$a^*(s)$を適用すると、容易に分散指数を求めることができる。

$$\left.\begin{aligned} I_{he}(t)&=C_{ahe}^2+\frac{a_{13}+b_1e^{-s_1t}+b_2e^{-s_2t}}{t}\\ a_{13}&=-(2m_3m_1-3m_2^2)/(6m_1^3)\\ b_1&=-Y(-s_1)/\{s_1^3(s_2-s_1)\}\\ b_2&=-Y(-s_2)/\{s_2^3(s_1-s_2)\}\\ Y(s)&=s^3+\{f_z(0)+r_z+2\nu+\alpha\}s^2\\ &\quad+[2\nu\{f_z(0)+r_z\}+(\alpha+\beta)r_z]s\\ &\quad+2\nu\beta r_z \end{aligned}\right\} \tag{20}$$

$$\left.\begin{matrix}s_1\\ s_2\end{matrix}\right\}=\{\alpha+f_z(0)+r_z\}/2 \mp\sqrt{\{\alpha+f_z(0)+r_z\}^2-4(\alpha+\beta)r_z}/2 \tag{21}$$

ただし、$C_{ahe}^2$, $m_1$、$m_2$、$m_3$、はそれぞれ$a(t)$の平方変動係数、平均、2次モーメント、3次モーメントである。

さらに、分散指数の原点($t=0$)の一般的性質[3]、$I_{he}(0)=1$、$I'_{he}(0)=a(0)-m_1^{-1}$を用いると次の関係が得られる。

$$a_{13}+b_1+b_2=1 \tag{22}$$

$$C_{ahe}^2-b_1s_1-b_2s_2=1 \tag{23}$$

$$I'_{he}(0)=\frac{b_1s_1^2+b_2s_2^2}{2}=\nu\left(1-\frac{\beta}{\alpha+\beta}\right) \tag{24}$$

式(24)から、$I'_{he}(t)$の原点における増分は、常に正の値で有ることがかる。一方、離散時間発生バーストの分散指数の原点における増分は$-\lambda$であり、$t=T$まで分散指数値は線形に減少し、$t>T$となって初めて値が増加する。このため、$I_{he}(t)$そのままでは近似できない。そこで、式(17)と同様に、$t>T$の範囲で$I_{he}(t)$を適用する。このとき、$I'_{he}(0)=\{I(2T)-I(T)\}/T$、$C_{ahe}^2=C_a^2$と置き、式(23)、式(24)、式(22)を用いて、$b_1$、$b_2$、$a_{13}$を近似する。

$$\left.\begin{aligned} b_2&=\{2I'_{he}(0)-(C_a^2-1)s_1\}/(s_2^2-s_1s_2)\\ b_1&=(C_a^2-1-b_2s_2)/s_1\\ a_{13}&=-b_1-b_2 \end{aligned}\right\} \tag{25}$$

$s_1$と$s_2$は、ON区間の平均長$\alpha^{-1}$とOFF区間の分布にのみ依存するので、式(21)をそのまま用いる。

### 3.3 $F_z(t)$が$k$-アーラン分布の場合の近似

$f_z^*(s)$として$k$-アーラン分布のラプラス変換を適用し、$a(t)$の分散指数を求めると以下となる。

$$\left.\begin{aligned} I_{ek}(t)&=C_{aek}^2+\frac{1}{t}\left\{a_{13}+\sum_{j=1}^{k}b_je^{-s_jt}\right\}\\ a_{13}&=-(2m_3m_1-3m_2^2)/(6m_1^3)\\ b_j&=-\frac{1}{s_j^3\prod_{i=1,i\neq j}^{k}(s_i-s_j)}X(-s_j)\\ X(s)&=(s+2\nu+\alpha)(s+k\beta)^k-\alpha(k\beta)^k \end{aligned}\right\} \tag{26}$$

ただし、$C_{aek}^2$、$m_1$、$m_2$、$m_3$、はそれぞれ$a(t)$の平方変動係数、平均、2次モーメント、3次モーメントであり、$s_j$、$(j=1,\cdots,k)$は、$(s+\alpha)(s+k\beta)^k-\alpha(k\beta)^k=0$を満たす、$s=0$以外の$k$個の根である。これらの根は、Muller法などの逐次計算で求めるのが一般的であるが、$\bar{N}^2/2$のオーダを越える計算量から比べれば、遙かに計算量は少なくてすむ。

分散指数$I_{ek}(t)$についても、$I'_{ek}(0)$が常に正になることが確認できる。したがって、OFF区間が$k$-アーラン分布の場合の離散時間発生バーストの分散指数の近似として、$I_{ek}(t)$をそのままでは適用できない。また、前節で述べた分散指数の原点における性質を用いる$t>T$の範囲の近似は、$k=2$では前節と同様に近似できる。しかし、$k>2$では、未定変数の数に対して、条件式の

数が少ないため、同様な方法が適用できない。

このため、$k>2$の場合には、新たな近似法を考える必要がある。離散時間発生バーストモデルと再生DSPPとの大きな違いは、ON区間でのパケットの発生過程にある。特に、再生DSPPの分散指数の原点における増分は、$\nu-m_1^{-1}$であり、ON区間でのパケット発生率($\nu$)と再生DSPPのパケット発生率($m_1^{-1}$)に関係している。そこで、ON区間でのパケット発生率$\nu$を離散時間発生バーストモデルの特徴パラメタから、近似することを考える。

$F_z(t)$を$k$-アーラン分布とする再生DSPP、$a(t)$、の平方変動係数$C_{aek}^2$は、

$$C_{aek}^2=\{\alpha\nu(C_z^2+1)+(\alpha+\beta)^2\}/(\alpha+\beta)^2$$

となる。これを$\nu$について解くと、次式が得られる。

$$\nu=(C_{aek}^2-1)(\alpha+\beta)^2/\{\alpha(C_z^2+1)\} \qquad (27)$$

式(27)において、$C_{aek}^2=C_a^2$と置いて、$\nu$を近似して、式(26)と式(17)を適用すれば分散指数の近似ができる。

以上の、近似法を用いた場合の分散指数の数値例を図2に示す。破線は近似法の値を示している。$C_z^2=1$のときは式(17)、$C_z^2=2$のときは式(17)の$I_H(t)$の代わりに式(25)による$I_{he}(t)$、$C_z^2=0.25$のときは$I_H(t)$の代わりに式(27)による$I_{ek}(t)$をそれぞれ適用する。式(27)を適用する$k=4$の近似は、他に比較して精度は若干悪いが、IDC≧6で0.5程度の誤差であり、実用上大きな問題とならないと考えられる。

## 4　平均残余稼働時間

一つの離散時間発生バーストソースが無限待ち室の待ち行列に加わった場合の平均残余稼働時間を求める。平均残余稼働時間を求めるには、パケットを処理する時間分布(サービス分布)を決める必要がある。$I(t)$の近似手法の有効性を次節の平均待ち時間より評価するため、サービス分布の影響が評価に影響しないようにする必要がある。サービス分布の平方変動係数が小さい程、$I(t)$の近似精度が平均待ち時間の近似精度に大きく影響することが報告されていることから[1]、サービス分布の平方変動係数がゼロである一定分布を仮定する。このときの平均サービス時間を$1/b$とする。

平均残余稼働時間$R_b$は、稼働時間の平均$m_b$と2次モーメント$m_{b2}$とより、$R_b=m_{b2}/(2m_b)$で求めることができる[1]。パケット発生間隔時間の最小値が$T$となることから、$T\geq 1/b$の場合と$T<1/b$の場合の$1/b$に分けて考

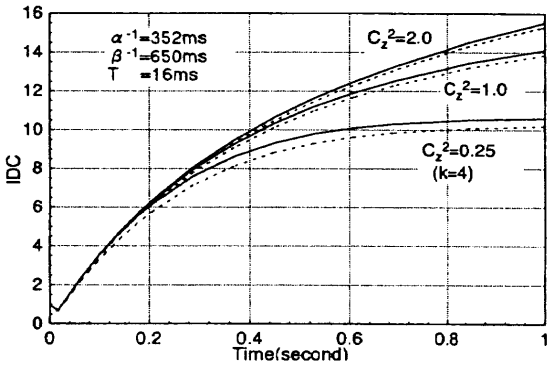



図 2　分散指数の近似例

える。これは、$\rho_T\equiv\beta/(\alpha+\beta)$と置き($\rho=\lambda/b$で、$1/b=T$とした場合の値)、加わる負荷$\rho$が$\leq\rho_T$場合と$\rho>\rho_T$の場合に分けて考えるのと等価である。

### 4.1　1/bがT以下の場合

$T\geq 1/b(\rho\leq\rho_T)$を満たす$1/b$の範囲では、パケットが待ち行列へ到着した時にサービスを待っているパケットが無いので(待ち率がゼロである)、到着パケットは即座にサービスされる。したがって、稼働時間の平均と2次モーメントは、サービス分布のそれに等しくなる。一定サービス分布の平均と2次モーメントは、それぞれ$1/b$、$1/b^2$となるので、平均残余稼働時間は以下となる。

$$R_b=m_{b2}/(2m_b)=1/(2b)\quad(\rho\leq\rho_T) \qquad (28)$$

### 4.2　1/bがTより大きい場合

$T<1/b(\rho>\rho_T)$を満たす$1/b$の範囲では、待ち率がゼロとならないため、若干の複雑な計算を要する。スペクトル分解法を適用した稼働期間分布のラプラス変換$\gamma(z)$の解析法およびその解が、文献[5]で述べられている。$\gamma(z)$から稼働期間の平均と2次モーメントを求めることで、平均残余稼働時間が導出できる。ここでは、結果のみを示す(尚、$\gamma(z)$から導出する基本的な考え方については、文献[1]を参照されたい)。また、離散時間発生バーストモデルの$F_z(t)$として、指数分布、超指数分布、$k$-アーラン分布の場合についての平均残余稼働時間を示す。

(1) $F_z(t)$が指数分布の場合

$$R_b=C_a^2/\{2(1-\rho)^2b\}\quad(\rho_T<\rho) \qquad (29)$$

(2) $F_z(t)$が超指数分布の場合

$$R_b=\frac{1}{2}\frac{C_a^2}{(1-\rho)^2 b}-\frac{s'_1(1,z=0)}{s_1} \qquad (\rho_T<\rho) \quad (30)$$

ただし、

$$\left.\begin{aligned} &s'_1(1,z=0)=X/Y\\ &X=\frac{1}{b}[\beta r_z-\{f_z(0)+r_z p\}s_1+ps_1^2]e^{(T-1/b)s_1}\\ &Y=-\{f_z(0)+p(r_z-2s_1)\}e^{(T-1/b)s_1}\\ &+[\beta r_z-\{f_z(0)+pr_z\}s_1+ps_1^2](T-1/b)\\ &\times e^{(T-1/b)s_1}+\{f_z(0)+r_z-2s_1\} \end{aligned}\right\}$$

で、$s_1$は、次の方程式の$Re(s)>0$における唯一の解である。

$$[p(\beta_1-s)(\beta_2-s)+q\{\beta r_z-f_z(0)s\}]e^{(T-1/b)s}-(\beta_1-s)(\beta_2-s)=0 \quad (31)$$

(3) $F_z(t)$が$k$-アーラン分布の場合

$$R_b=\frac{1}{2}\frac{C_a^2}{(1-\rho)^2 b}-\sum_{i=1}^{k-1}\frac{s'_i(1,z=0)}{s_i} \qquad (\rho_T<\rho) \quad (32)$$

ただし、

$$\left.\begin{aligned} &s'_1(1,z=0)=X/Y\\ &X=\frac{1}{b}[p(k\beta-s_i)^k+q(k\beta)^k]e^{(T-1/b)s_1}\\ &Y=k(k\beta-s_i)^{k-1}+[-kp(k\beta-s_i)^{k-1}\\ &+(T-1/b)\{p(k\beta-s_i)^k+q(k\beta)^k\}]e^{(T-1/b)s_1} \end{aligned}\right\}$$

である。また、$s_i$は、次の方程式を満たす$Re(s)>0$平面の$k$-1個の解となる。

$$\{p(k\beta-s)^k+q(k\beta)^k\}e^{(T-1/b)s}-(k\beta-s)^k=0 \quad (33)$$

## 5　多重化時の平均待ち時間近似

離散時間発生バーストパケットソースを$n$多重したトラヒック流が待ち行列に加わった場合の平均待ち時間を、文献[1]の分散指数と平均残余稼働時間を用いた多重化トラヒックを再生過程近似する手法を適用して、$GI/G/1$モデルより評価する方法を述べる。$n$多重されたトラヒック流を再生過程で近似するための、平均、平方変動係数、歪み度をそれぞれ、$\lambda_n^{-1}$、$C_{an}^2$、$S_{kn}$とする。離散時間発生バーストパケットソースの平均が$\lambda^{-1}$であるので、$\lambda_n^{-1}=(n\lambda)^{-1}$となる。平方変動係数は、文献[1]の式(9)を用いて、$C_{an}^2=I(R_b/n)$で近似する。

$C_{an}^2=1$の時は、$M/G/1$モデルの指数入力過程の平均に、$\lambda_n$をモーメントマッチさせる。$C_{an}^2<1$の時は、$E_k/G/1$モデルの$k$-アーラン入力過程の平均と平方変動係数に、$\lambda_n$と$C_{an}^2$をそれぞれモーメントマッチさせる。ただし、$k=\lfloor 1/C_{an}^2 \rfloor$とする。$C_{an}^2>1$の場合は、$H_2/G/1$の超指数入力過程に$\lambda_n$、$C_{an}^2$、$S_{kn}$を文献[6]の式(3.101)を用いてモーメントマッチさせる。歪み度$S_{kn}$は次のように求める。文献[1]の式(41)の各パラメタにサフィックス$n$を付け、文献[1]の式(45)の関係式に代入すると次式が得られる。

$$S_{kn}=\frac{1}{2C_{an}^3}\left[\frac{3(C_{an}^2-1)^2}{f_{0n}/\lambda_n-1}+3C_{an}^4+1\right] \quad (34)$$

ただし、$f_{0n}$は、$C_{an}^2>1$の時の$n$多重化トラヒック流を再生近似した場合の再生過程($H_2$)の原点における密度

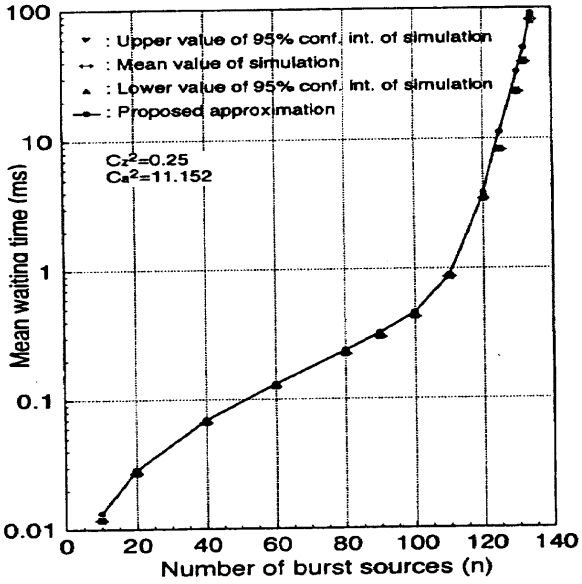



図 3　平均待ち時間の数値例($C_z^2=0.25$)

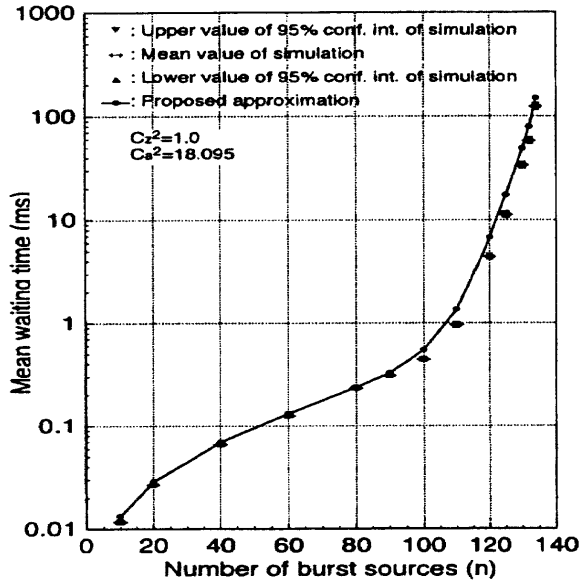



図 4　平均待ち時間の数値例($C_z^2=1.0$)

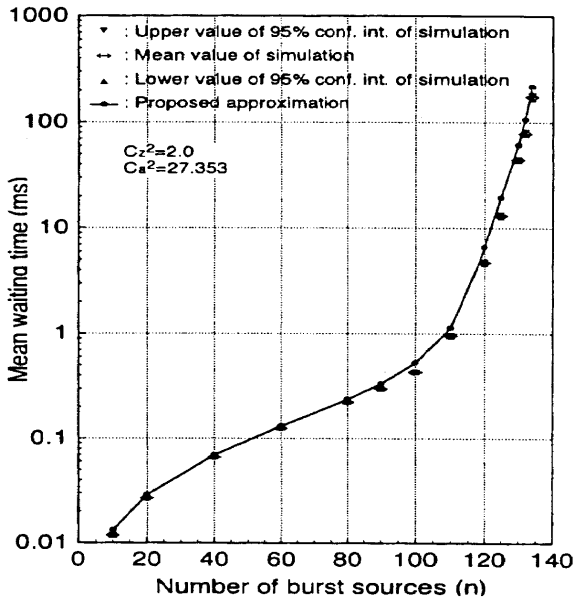



図 5　平均待ち時間の数値例($C_z^2=2.0$)

を表す。したがって、$f_{0n}/\lambda_n-1=(f_{0n}-\lambda_n)/\lambda_n$は、多重化流の近似再生過程($H_2$)の分散指数の原点における増分を$\lambda_n$で規格化した量となる。

$f_{0n}/\lambda_n-1$を文献[1]の分散指数の原点における増分に関する規格化増分等価法の考えに基づいて、次の様に近似する。

$$f_{0n}/\lambda_n-1=\frac{C_{an}^2-1}{C_a^2-1}(f_0/\lambda-1) \qquad (35)$$

ここで、$f_0/\lambda-1$は、バーストソースの分散指数の原点における増分を$\lambda$で規格化した量を表している。$H_2$を$n$多重した場合の分散指数$I_{Hn}(t)$は$H_2$の分散指数$I_H(t)$と等しくなる[1,6]ことを考慮して、$f_0/\lambda-1$として$I'_H(0)/\lambda$を用いる。これは、離散時間発生バーストパケットのモーメント、式(3)、を$H_2$にモーメントマッチさせれば求まる。すなわち、文献[6]の式(3.101)を用いると以下で表すことができる。

$$f_0/\lambda-1=\{k_H\ \lambda_1+(1-k_H\ )\lambda_2\}/\lambda-1 \qquad (36)$$

ただし、$\lambda_1$、$\lambda_2$は文献[6]の式(3.101)を用い、また本式中の$k$を$k_H$ で表している。

以上のように再生近似で得られた待ち行列モデル$M/G/1$、$E_k/G/1$、$H_2/G/1$の平均待ち時間は、文献[6]のスペクトル分解法で求めることができる。

図3-図5に数値例を示す。平均サービス時間$1/b=1/3ms$の一定分布、$\alpha^{-1}=352ms$、$\beta^{-1}=650ms$、$T=16ms$、1バーストソースの負荷を0.00732erlとした。OFF区間分布として、指数分布、$k$-アーラン分布、超指数分布を用い、$C_z^2$をそれぞれ、1、0.25、2.0とした。$C_{an}^2$は、提案した分散指数の近似式を用い、$n$多重した時の負荷と等価な負荷($\rho$)の$R_b$から計算した。図中のシミュレーション結果は、1回の実行に関して、1ソース当たり約8万パケットを発生させ、50回行ったものである。

加わる負荷の大きいところでは、近似精度が若干悪くなるが、全体を通して安全側の評価となっており、概ね良好な近似が得られていることが分かる。

表1に平均待ち時間を$I(t)$の厳密解と近似式を用いて評価した場合の比較を示す。特に、表中の平均待ち時間が評価されていない部分は($I(t)$の厳密解を用いた場合の"-")、ワークステーション(SUN/ss20)を用いた30分以上の倍精度計算で結果が得られなかったものである。比較のできる範囲で、何れの$C_z^2$の大きさに対しても、多重数$n$の大きい(加わる負荷の大きい)部分で、$I(t)$の近似式を用いた方が、小さめの平均待ち時間評価を与えるが、その差は小さいことが分かる。このことから、平均待ち時間に対する$I(t)$の近似式は、ほぼ良好であることが分かる。

## 6　むすび

バーストパケットトラヒックモデルとして、ON区間でのパケット発生数分布を発生間隔が一定の幾何分布、パケットが発生しないOFF区間長を$k$-アーラン分布、超指数分布に拡張した。この離散時間発生バーストパケットを$n$多重したトラヒック流が待ち行列システムに加わった場合の性能を、分散指数と平均残余稼働時間を用いて再生過程近似して解析する場合、分散指数が非常に複雑であり、実用的な計算に不向きであった。この分散指数の計算の簡便化を図るために、再生DSPPによる分散指数の近似法を提案した。提案方法の有効性を平均待ち時間のシミュレーション結果との比較評価により行った。負荷の大きいところで、若干近似精度が悪くなるが、全体を通して安全側の近似であり、良好な近似であることを明らかにした。

今回のバーストソースのモデルの拡張は、再生過程の範囲にとどまっているが、より多くの情報発生パターンソースのモデル化に対応するには、非再生過程の範疇への拡張が今後重要である。また、非再生なソースが多重化されるシステム系の性能解析法も今後の重要な課題である。

表 1　平均待ち時間評価による$I(t)$近似精度の比較

| Number of Sources($n$) | Approximation of the mean Waiting time(ms) | | | | | |
|---|---|---|---|---|---|---|
| | $C_z^2=0.25$ | | $C_z^2=1.0$ | | $C_z^2=2.0$ | |
| | Exact $I(t)$ | Approx. $I(t)$ | Exact $I(t)$ | Approx. $I(t)$ | Exact $I(t)$ | Approx. $I(t)$ |
| 10 | 0.013161 | 0.013161 | 0.013161 | 0.013161 | 0.013161 | 0.013161 |
| 20 | 0.028579 | 0.028579 | 0.028579 | 0.028579 | 0.028576 | 0.028579 |
| 40 | 0.068987 | 0.068987 | 0.068987 | 0.068987 | 0.068987 | 0.068987 |
| 60 | 0.13049 | 0.13049 | 0.13049 | 0.13049 | 0.13049 | 0.13049 |
| 80 | 0.23542 | 0.23542 | 0.23542 | 0.23542 | 0.23542 | 0.23542 |
| 90 | 0.32164 | 0.32164 | 0.33448 | 0.32293 | 0.33438 | 0.33199 |
| 100 | 0.47587 | 0.45680 | 0.55804 | 0.54134 | 0.52755 | 0.52248 |
| 110 | 0.92694 | 0.89084 | 1.3727 | 1.3453 | 1.1371 | 1.1188 |
| 120 | 4.1325 | 3.8138 | 6.8378 | 6.6720 | 6.6914 | 6.5301 |
| 125 | 11.913 | 11.100 | 17.569 | 17.214 | 19.640 | 19.294 |
| 130 | 33.836 | 32.367 | 49.486 | 48.586 | - | 60.385 |
| 132 | - | 49.132 | 79.509 | 78.042 | - | 105.74 |
| 134 | - | 90.236 | - | 147.34 | - | 215.5 |

**参考文献**


[1] 阿部，“多重化バーストパケット入力待ち行列システムの再生過程近似による性能評価法”，信学論(B-I)，Vol. J79-B-I, No.6, pp.383-395, 1996.

[2] 阿部，宗宮，“ATM 通信におけるセル品質評価法の考察”，信学技報，SSE91-118, pp.31-36, 1991.

[3] Okuda,T.; Akimaru, H.; Sakai, M.“A simplified performance evaluation for packetized voice systems”,*IEICE*, *Tran.*, Vol. E73, No.6, pp.936-941, 1990.

[4] Heffes,H.; Lucantoni,D,M., “A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Multiplexer Performance”, *IEEE JSAC*, Vol. SAC-4, No.6, pp.856-868, 1986.

[5] Rice,O,S., “Single Server System - II Busy Periods,” *BSTJ*, No.1, 1962.

[6] 秋丸，川島，情報通信トラヒック，電気通信協会，1990.

[7] Whitt,W., “The Queueing Network Analyzer,” *BSTJ*, Vol.62, No.9, 1983.

[8] Neuts,M.F., Structured Stochastic Matrices of M/G/1 Type and Their Applications, Marcel Dekker, Inc., 1989.

学術情報センター紀要　第10号(1998)

研究論文

# Analytical Performance Study of Traffic Shaping Mechanisms

# 解析手法によるトラヒックシェーピング機構の性能評価

Yusheng JI
National Center for Science Information Systems
学術情報センター　計　宇生

**ABSTRACT**

Traffic shaping can reduce delay and loss in the network nodes by smoothing the traffic before letting it entering the network. Rate control algorithms such as Leaky Bucket and Moving Window are proposed for traffic shaping. Although many efforts have been made to analyze the performance of Leaky Bucket as a rate control scheme, few results can be found for Moving Window-liked mechanisms.

In this paper, we build a general purposed analytical model for Moving Window scheme, by dealing it as a G/D/m queueing system, give the resolution for the Poisson and batch Poisson arrival, and by showing the results obtained through analytical models for Moving Window and Leaky Bucket, and results of simulation, give comparison of the performance of these two schemes when they are used for traffic shaping.

[Keywords]　traffic shaping, rate control, queueing system, Moving Window, Leaky Bucket

## 1 Introduction

In order to provide the quality of services while maintaining the level of network resource utilization, it is necessary to regulate source traffic by using traffic shaping mechanisms. A traffic shaping mechanism usually acts at the source side working with a buffer to prevent too many data rushing into the network.

In [1], Radhakrishnan et al. proposed a shaping mechanism called the Shift Register Traffic Shaper (SRTS). SRTS is a extension of the Moving Window scheme, with more than one adjustable windows. In [1], simulation results show that with more flexibility to short term bursts, SRTS causes smaller packet delay and loss when comparing with Leaky Bucket shaping. However, comparisons are conducted only from the sight of traffic passing the shaper, no results about the shaping effect, i.e. the performance improvement in the network have been produced.

Shaping can reduce delay and loss in the network nodes by smoothing the traffic before letting it enter the network. Buffering delay in the shaper will occur with non-conforming packets. Intuitively, with the same source traffic, the larger the delay in the shaper, the bigger the performance improvement can be gained in the network, but the larger the total delay [2].

In a previous study [3], we evaluated a packet scheduling scheme, called the Virtual Rate-Based Queueing (VRBQ), which is an extension to the basic Processor Sharing algorithm. By applying VRBQ, at least a requested rate (called the virtual rate) can be guaranteed for each source. We also gave the bound of queueing delay in a VRBQ multiplexer for traffic shaped by Moving Window or Leaky Bucket rate control schemes.

In a more recent study [4], we gave a performance comparison of two commonly proposed algorithms, the Leaky Bucket and the Moving Window, produced by simulation results. Our results shown that for traffic with large bursts, Leaky Bucket algorithm have more preferable feature than the other one.

The efforts to provide the performance model analytically for Leaky Bucket [5] rate control algorithm have been made from various aspects. Those results can be found in literature, such as [6]-[12].

On the other hand, few analytical results can be found on performance issues for Moving Window scheme. In this paper, we build a general purposed analytical model for Moving Window scheme, by adopting it as a G/D/m queueing system, give the solution for the Poisson and batch Poisson arrival, and by showing the results obtained through analytical models for Moving Window and Leaky Bucket, and results of simulation, give comparison of the performance of these two schemes.

In Section 2, a new analytical model for Moving Window scheme is given. We also adopt the queueing analysis on Leaky Bucket from other study into our system. In Section 3, we give results produced by simulation, and compare them with the analytical results obtained by our models. The conclusion and issues for future study will be given in Section 4.

## 2 Analytical Models

Shaping is a commonly used traffic control mechanism in ATM networks. The purpose of shaping is, by giving constraints on peak cell rate, average cell rate and burstyness of source traffic, to guarantee the quality of services, and at the same time, to obtain reasonable network bandwidth utilization. As well as for Usage Parameter Control (UPC) or policing, some rate control algorithms have been adopted for shaping. To be different with policing, in which violating cells are discarded or marked, however, cells which can not pass a shaper immediately will be dropped or queued in the buffer till the time when they can. Limiting the buffer length in shapers will result on cell loss, but at the same time, be possible to limit the maximum queueing delay in shapers. For more generality and simplicity, we used the term "packet" here to refer to a fix-sized basic data unit, which is similar as a cell in ATM networks.

One of the performance measure to a shaping mechanism is the latency caused by the shaper itself when source traffic passes through it. Along with delay, delay variation has more impact to real-time applications, with which it is usually desirable to preserve the characteristics of the input process. Limiting the length of buffer in shapers can limit the maximum delay, but may result in packet loss. When the queueing space for packets in shapers is limited, packets arrived at full queues are discarded by shapers immediately.

Our discussion is focused on the shaping mechanism used for enforcing average transmission rate and the burstyness. The peak transmission rate can be formalized by a packet spacer as in [13], or limited by the bandwidth of the output line of the shaper. In following discussion, we assume that shapers have limited output bandwidth of $\lambda_p$. We also use "tick" as the basic time unit in a discrete time system. The basic time unit tick is determined so that the peak transmission rate $\lambda_p$ is 1 packet/tick, equal to the bandwidth of output links of shapers.

### 2.1 Moving Window

Without losing generality, let us assume that the number of packets waiting to be transmitted at the queue of a Moving Window shaper is 0 at time $t \leq 0$, and $A(t)$ is the accumulated number of packet departing from the shaper during time period $(0, t]$. Let $A(t_1, t_2) = A(t_2) - A(t_1)$ for $t_2 > t_1$, the Moving Window constraint can be expressed simply as:

$$\forall t, \quad A(t, t+I) \leq \lambda_s I, \tag{1}$$

where $\lambda_s$ is a parameter controlling the average transmission rate, and $I(>0)$ is the averaging interval (called the time window in [14]), being a parameter which is able to control the burstyness of traffic.

Using constraint (1), if the source indiscriminately send as many packets as possible, the number of packets which can pass through a Moving Window is at most

$$B_{MW}=\lambda_s I. \tag{2}$$

This is the maximum burst size allowed by the Moving Window scheme.

Since the Moving Window scheme constrains the maximum number of packets which can pass the shaper during any time period of length $I$ to be $\lambda_s I$, such system can be considered as a single queue served by $m=\lambda_s I$ servers, and each server has the constant service time of $I$. Therefore, we can use a G/D/m queueing system to model a Moving Window shaper with generally distributed input traffic pattern, averaging interval of $I$, shaping average rate of $\lambda_s$, and associated by infinite buffer space. When the buffer associated has limited length of $B$ packets, the model becomes G/D/m/B.

Towards the G/D/m/B problem, Chu [15] produced a solution with slotted time constraint, by assuming that the constant service time can be only initiated at the beginning of a time slot, and finished at the termination of the same slot. Although the service time of each server, which is assumed to be the time slot interval in the literature, can be expected much shorter than in our model, in which the service time of each server should be a whole averaging interval $I$, we can still extend their model to fit our case.

With a Moving Window scheme, packets arrived in the system can be served at time $t$ as soon as the number of packets which have been served during time interval $[t-I,\ t)$ becomes less than $\lambda_s I$. Therefore, if we consider the service time $I$ as a basic time slot, every server may start service at not only the beginning point of a time slot, but any time during the time slot. But if we look at the system once every $I$ time units, what we can see is that there are at most $\lambda_s I$ packets which have finished service during each time we look at it. So we can still assume an averaging interval as a basic time slot, as if those packets were served at the beginning of each slot and hence finished at the end of it.

Let $p_k$ be the probability of $k$ packets which are left in the system (including packets which are in service). In our system, $p_k$ corresponds to the probability, that the number of packets which pass the shaper during an averaging interval plus the number of packets which are left in the queue at the end of the interval, equals to $k$. When $k\leq m$, where $m$ is the number of the server and $m=\lambda_s I$, $p_k$ is the probability that there are $k$ packets which pass the shaper during an averaging interval, and no packets left in the queue at the end of the interval. When $k>m$, $p_k$ is the probability that there are $m$ packets which pass the shaper during an averaging interval, and $k-m$ packets left in the queue at the end of the interval. We also define $a_m$ as the equilibrium probability that there are no more than $m$ packets which pass the shaper during an averaging interval, i.e.,

$$a_m=\sum_{k=0}^{m} p_k.$$

Let $\pi_k$ be the probability that there are $k$ packets which arrive during an averaging interval, the balance equations among these equilibrium probabilities at the termination of adjacent averaging intervals can be written as same as in [15]:

$$\begin{aligned}
&p_0=a_m\pi_0\\
&p_1=a_m\pi_1+p_{m+1}\pi_0\\
&\quad\vdots\\
&p_k=a_m\pi_k+p_{m+1}\pi_{k-1}+p_{m+2}\pi_{k-2}+\cdots+\\
&\qquad p_{m+k-1}\pi_1+p_{m+k}\pi_0, \qquad \text{for } k\leq B \qquad (3)\\
&\quad\vdots\\
&p_k=a_m\pi_k+p_{m+1}\pi_{k-1}+p_{m+2}\pi_{k-2}+\cdots+\\
&\qquad p_{m+B-1}\pi_{k+1-B}+p_{m+B}\pi_{k-B},\\
&\qquad\qquad \text{for } B<k\leq m+B-1
\end{aligned}$$

Recall that $B$ is the buffer length. The first equation in (3) describes the case in which there is no packet which pass the shaper during an averaging interval, if no more than $m$ packets were served during the last interval and no arrivals occur during this interval. The second equation describes the case in which one packet which pass the shaper

during an averaging interval, if no more than $m$ packets passed the shaper during last interval and one packet arrives during this interval, or there were $m+1$ packets in the system at the end of last interval and no packet arrives during this interval, etc. Due to limited buffer size, $p_k=0$ when $k>m+B$. From the conservation of probability, we also have

$$\sum_{k=0}^{m+B} p_k=1.$$

Assume that the average arrival rate is $\lambda_a$. Since the expected number of packets which arrive during an averaging interval $I$ is $\lambda_a I$, we can get $\pi_k$ for Poisson arrivals simply as

$$\pi_k=\frac{(\lambda_a I)^k e^{-\lambda_a I}}{k!}. \tag{4}$$

As a more bursty source, we consider $\pi_k$ for the batch Poisson arrival. In a batch Poisson arrival process, every arrival contains a burst (a batch) of data instead of an unit of data (a packet) in a Poisson process, and the burst length is geometrically distributed. An On-Off arrival process with geometrically distributed On and Off periods also becomes a batch Poisson arrival when the peak arrival rate (the arrival rate in On periods) is infinite.

As described above, here the averaging interval $I$ is a basic time slot. If the expected inter-arrival time of batches is $I_a$, then the expected number of batches which arrive during an averaging interval $I$ is $I/I_a$. If the overall average arrival rate is $\lambda_a$ as we assumed above, the expected number of packets in each batch arrival is $\lambda_a I_a$. Hence, the probability mass functions of random variables of the number of packets in a batch arrival, $X$, and the number of batch arrivals during an averaging interval, $Y$, are:

$$f_X(l)=\frac{1}{\lambda_a I_a}\left(1-\frac{1}{\lambda_a I_a}\right)^{l-1}, \quad l=1, 2, \cdots \tag{5}$$

$$f_Y(n)=e^{-\frac{I}{I_a}}\left(\frac{I}{I_a}\right)^n \frac{1}{n!}, \quad n=0, 1, 2, \cdots \tag{6}$$

The total number of packets which arrive during an averaging interval is a random sum and equals to

$$S=\sum_{i=0}^{Y} X_i,$$

where $X_i$ is a random variable distributed as (5), and $Y$ a random variable distributed as (6). By expressing $S$ in terms of the characteristic function as in [16], the probability mass function of $k$ packets arriving during an averaging interval, $\pi_k$, is a compound Poisson distribution which can be expressed as:

$$\pi_k=\begin{cases} e^{-\frac{I}{I_a}}, & k=0 \\ \displaystyle\sum_{i=0}^{k}\binom{k-1}{i-1}\left(\frac{I}{\lambda_a I_a^2}\right)^i\left(1-\frac{1}{\lambda_a I_a}\right)^{k-i}\frac{e^{-\frac{I}{I_a}}}{i!}, & k=1, 2, \cdots \end{cases} \tag{7}$$

Note that the computational complexity for obtaining $p_k$ largely depends on the maximum number of packets allowed in the system (especially in the case of batch Poisson arrivals). The number of packets which can pass the shaper during an averaging interval can be computed from probabilities $p_k$ as:

$$T_{MW}=\sum_{k=1}^{m-1} kp_k+m\sum_{k=m}^{m+B} p_k. \tag{8}$$

Since the expected number of packets which arrive during an averaging interval is $\lambda_a I$, the packet loss probability, which is the expected fraction of the number of lost packets, can be obtained as

$$L_{MW}=\frac{\lambda_a I-T_{MW}}{\lambda_a I}=1-\frac{T_{MW}}{\lambda_a I}. \tag{9}$$

### 2.2 Leaky Bucket

If $A(t_1, t_2)$ is the accumulated number of packets which can pass a shaper during time period $(t_1, t_2]$, the Leaky Bucket constraint can be expressed as

$$\forall t, x\geq 0, \quad A(t, t+x)\leq \lambda_s x+\sigma, \tag{10}$$

where $\sigma$ is the maximum number of tokens (or bucket depth) in unit of packets, and $\lambda_s$ the token generation rate.

Let us also assume that the number of packets waiting at the Leaky Bucket to be transmitted is 0 and the number of token is $\sigma$ at time $t \leq 0$. Note that generality will not be lost by the second assumption because the number of tokens will reach to the maximum within a limited period if the packet arrival rate $\lambda_a$ is smaller than the token generation rate $\lambda_s$. The traffic pattern with maximum bursts can be obtained by repeating the following sequence: Transmit packets at the peak rate until the bucket is empty, then stop transmission until the tokens fill the bucket. In this case, the maximum number of packets which can pass a Leaky Bucket with peak rate $\lambda_p$ is equal to

$$B_{LB}=\frac{\lambda_p \sigma}{\lambda_p-\lambda_s}. \tag{11}$$

Since the maximum burst size (MBS) is a burstyness controlling parameter of shapers, it can be used for comparing the performance of time window-based and credit-based schemes as in [4]. When MBS's under Moving Window scheme (Eq. (2)) and Leaky Bucket scheme (Eq. (11)) are equal, we obtain the relationship between the maximum number of tokens in the bucket (the bucket depth) $\sigma$ and averaging interval $I$ as

$$\sigma=\lambda_s I\left(1-\frac{\lambda_s}{\lambda_p}\right). \tag{12}$$

Many studies have been done concerning of Leaky Bucket-based rate control. Simulation results can be found in [17], [1] and [13]. An analytical model of Leaky Bucket scheme has been established for Poission arrival by Sidi et al. in [7]; Let $p_k$ be the probability that the number of tokens used plus the number of packets waiting in the buffer is $k$. When $0 \leq k \leq \sigma$, there are $\sigma-k$ tokens left in the bucket and no packets waiting in the buffer. When $k>\sigma$, there are no tokens left, and $k-\sigma$ packets in the buffer. Consider of a slotted time axis where a new token is generated at each slot boundary. When $\pi_k$ is the probability that there are $k$ packets arrived during a slot interval, which is a token generation interval $\lambda_s$, the steady-state equations can be written as:

$$p_k=p_0\pi_k+\sum_{i=0}^{k}p_{i+1}\pi_{k-i}, \quad \text{for} 0 \leq k \leq \sigma+B-1 \tag{13}$$

where $B$ is the buffer size, and $p_k=0$ when $k>\sigma+B$. From the conservation of probability, we still have

$$\sum_{k=0}^{\sigma+B}p_k=1.$$

With Poisson arrival process, the probability of $k$ packets that arrive during a token generation interval is

$$\pi_k=\frac{\left(\frac{\lambda_a}{\lambda_s}\right)^k e^{-\frac{\lambda_a}{\lambda_s}}}{k!}. \tag{14}$$

On the other hand, for batch Poisson arrivals, if $I_a$ is the average inter-arrival time of batches, $\lambda_a$ is the overall average arrival rate, then the average number of packets in each batch arrival is $\lambda_a I_a$, and the probability mass function of $X$, i.e., the number of packets in a batch arrival, remains the same as (5). Since the average number of batch arrivals during a token generation interval is $1/\lambda_s I_a$, the probability mass function of $Y$, the inter-arrival time, becomes to

$$f_Y(n)=\frac{e^{-\frac{1}{\lambda_s I_a}}}{(\lambda_s I_a)^n n!}. \tag{15}$$

Therefore, the probability of $k$ packets that arrive during a token generation interval is

$$\pi_k=\begin{cases} e^{-\frac{\lambda_a}{\lambda_s}} \\ \displaystyle\sum_{i=0}^{k}\binom{k-1}{i-1}\frac{\left(1-\frac{1}{\lambda_a I_a}\right)^{k-i}e^{-\frac{1}{\lambda_s I_a}}}{(\lambda_s \lambda_a I_a^2)\, i!}, \end{cases} \qquad k=1,\ 2,\ \cdots \tag{16}$$

Similarly with Moving Window scheme, the number of packets which can pass a Leaky Bucket shaper during a token generation interval is [7]

$$T_{LB}=p_0\left(\sum_{k=1}^{\sigma-B}k\pi_k+(\sigma+B)\bar{\pi}_{\sigma-B}\right)+\sum_{i=1}^{\sigma+B}p_i*\left(\sum_{k=1}^{\sigma+B-i+1}k\pi_k+(\sigma+B-i+1)\bar{\pi}_{\sigma+B-i+1}\right), \tag{17}$$

where $\bar{\pi}_k=1-\sum_{i=0}^{k}\pi_i$. Since the average number of arrivals during a token generation interval is $\lambda_a/\lambda_s$, the packet loss probability is

$$L_{LB}=1-\frac{T_{LB}\lambda_s}{\lambda_a}. \tag{18}$$

## 3 Numerical Results

Figure 1 shows a queueing model with multiple traffic sources; Each source is enforced by a traffic shaper before entering the network, and merges into a switching node in the network. In order to limit loss probability, each shaper is associated by a buffer to queue packets which can not pass a shaper immediately. The output of queues in shapers is constrained by relations (1) or (10).

Shaping parameter $\lambda_s$ in (1) and (10), as average rate or token generation rate, is chosen to be 0.1 packet/tick. Packets generated from a Poisson source have average arrival rate $\lambda_a$ of 0.067 packet/tick. With batch Poisson arrivals, the average length of a batch arrival is 10 packets, and the average length of inter-arrival time is 150 ticks. Therefore, as same as the Poisson arrival, the average arrival rate $\lambda_a$ from each batch Poisson source is also 0.067 packet/tick. Hence, the over-dimensioning factor, which is the ratio of shaping average rate to the average arrival rate ($\lambda_s/\lambda_a$), in both cases is 1.5. This allows shaped traffic to maintain a certain level of burstyness.

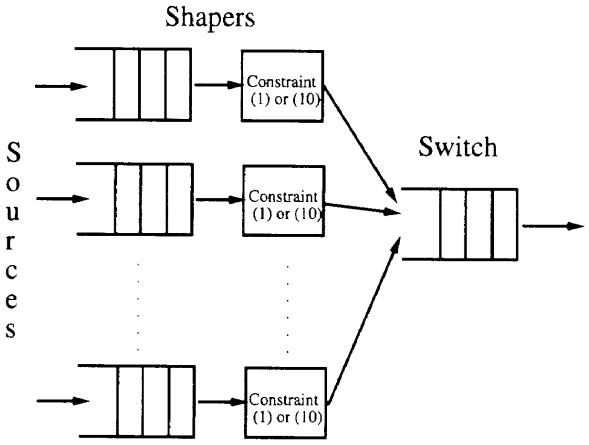



Figure 1 The queueing model.

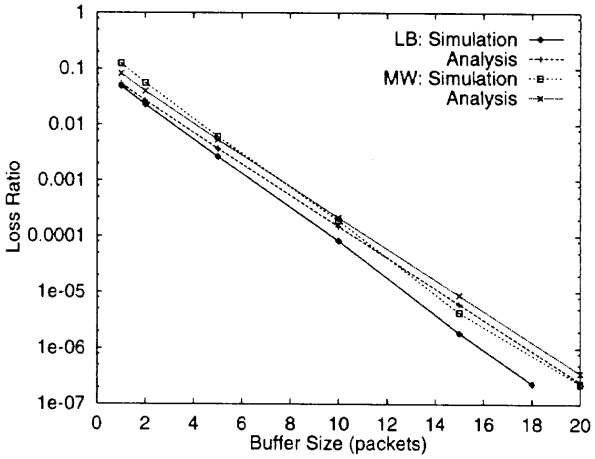



Figure 2 Packet loss vs. buffer size: Poisson arrival.

Table 1 Pockets loss probability for batch Poisson arrival.

| Buf. Size (packet) | Moving Window | | Leaky Bucket | |
|---|---|---|---|---|
| | Simul. | Analysis | Simul. | Analysis |
| 10 | 0.3413 | 0.1663 | 0.3630 | 0.1860 |
| 20 | 0.1438 | 0.1066 | 0.1829 | 0.1187 |
| 50 | 0.02967 | 0.03266 | 0.04565 | 0.03632 |
| 100 | 0.004166 | 0.005357 | 0.006545 | 0.005935 |
| 150 | 6.508e-4 | 9.183e-4 | 0.001017 | 0.001016 |
| 200 | 1.005e-4 | 1.585e-4 | 1.556e-4 | 1.754e-4 |
| 250 | 1.613e-5 | 2.741e-5 | 2.801e-5 | 3.033e-5 |

When shapers have limited buffer space, which limits the maximum queueing delay, the relationships between loss ratio and buffer length for Poisson arrival and batch Poisson arrival are shown in Figs. 2 and 3 respectively, with the Y-axes logarithmically scaled.

In case of Poisson arrival in Fig. 2, the averaging interval $I$ for Moving Window scheme is set to 20 ticks. Since the shaping average rate $\lambda_a$ is 0.1 packet/ticks, the maximum burst size becomes to 2

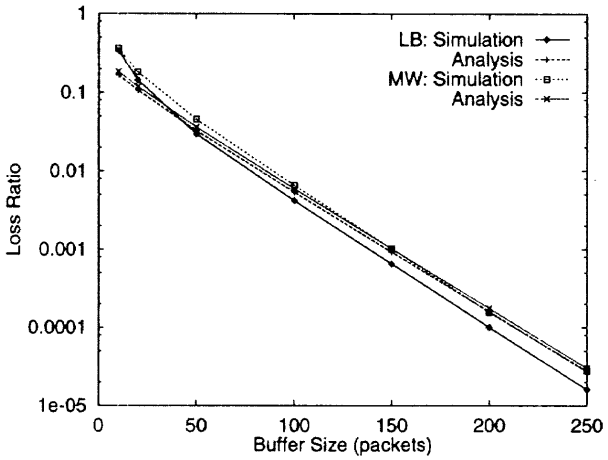



Figure 3　Packet loss vs. buffer size: batch Poisson arrival.

packets. For Leaky Bucket scheme, the maximum number of tokens is set to 2. With the same $\lambda_a$, it corresponds to about $I=23$ ticks. We can see from simulation results that packet loss reduces exponentially when buffer size increases, although Leaky Bucket shaping has smaller loss probability than Moving Window with the same buffer size. Analytical results have the same tendency as simulation results and show good precision.

For batch Poisson arrival in Fig. 3, the maximum burst size is set to 20 packets. Since batch Poisson arrival has more burstyness, it shows larger loss probability comparing with Poisson arrival, even with larger MBS. Simulation shows that loss ratio reduces near exponentially when buffer size increases, and again Leaky Bucket shaping has smaller loss probability. On the other hand, the figure shows that the analytical results of loss probability in both schemses are closer to exponential decrease with buffer length in crement, and has smaller difference between two schemes, because analytical results of Leaky Bucket scheme have larger difference with simulation results. But from the results in Table 1 we can still see that Leaky Bucket has smaller loss probability than Moving Window, as we found in our previous study.

## 4 Conclusions

As two most familiar rate control schemes, we have dealt with the performance issues of Leaky Bucket and the Moving Window when they work as shaping mechanisms. The main contribution of this paper is to build an analytical model for Moving Window rate control scheme. Based on other studies on Leaky Bucket, we obtained a method to compare these two schemes analytically. Through the comparison, we also learned that the Leaky Bucket algorithm have better feature than the other one as a shaping mechanism.

Here we only gave performance feature in shapers, which is from the user's point of view. From our previous study, which made more extensive comparisons of these two schemes by simulation from not only the user's but also the network's point of view, we can gain a total vision on the performance of these two schemes.

Input source dealt here are only Poisson and batch Poisson arrivals. Also the buffer space assumed in our models are limited. In such cases, the computational complexity increases with buffer length. For solutions of very large buffer space, approaches for resolving infinite buffer can be expected with less complexity and more accuracy. Therefore, further studies are needed to explore solutions for more general sources and system descriptions.

## References

[1] Radhakrishnan,S.; Raghavan,S.; Agrawala,A., "A Flexible Traffic Shaper for High Speed Networks: Design and Comparative Study with Leaky Bucket," *Computer Networks and ISDN Systems*, Vol.28, pp. 453-469, 1996.

[2] Li,S-q.; Chong,S., "Fundamental Limits of Input Rate Control in High Speed Network," *Proc. IEEE INFOCOM'93*, pp. 662-671, 1993.

[3] Ji,Y.; Asano,S., "Virtual Rate - Based Queueing: A Generalized Queueing Discipline for Switches in High-Speed Networks,"

*IEICE Trans. Commun.*, Vol. E77-B, No. 12, pp.1537-1545, 1994.

[4] Ji,Y., "The Characteristics of ATM Traffic Constained by Different Shaping Mechanisms" (in Japanese) *Research Bulletin of the National Center for Science Information Systems*, Vol.9, pp.189-194, 1997.

[5] Turner,J., "New Directions in Communications (or Which Way to the Information Age?)," *IEEE Commun.*, Vol.24, No.10, pp. 8-15, 1986.

[6] Bala,K., Cidon,I.; Sohraby,K., "Congestion Control for High Speed Packet Switched Networks," *Proc. IEEE INFOCOM'90*, pp.520-526, 1990.

[7] Sidi,M.; Liu,W.-Z.; Cidon,I.; Gopal,I., "Congestion Control Through Input Rate Regulation," *IEEE Trans. Commun.*, Vol. 41, No. 3, pp.471-477, 1993.

[8] Berger,A., "Performance Analysis of a Rate Control Throttle where Tokens and Jobs Queue," *IEEE JSAC*, Vol.9, No.2, pp. 165-170, 1991.

[9] Ahmadi,H.; Guerin,R.; Sohrabi,K., "Analysis of a Rate-Based Access Control Mechanism for High-Speed Networks," *IEEE Trans. Commun.*, Vol.41, No.6, pp. 940-950, 1993.

[10] Sohraby,K.; Sidi,M., "On the Performance of Bursty and Correlated Sources Subject to Leaky Bucket Rate-Based Access Control Schemes," *Proc. IEEE GLOBECOM'91*, pp.0426-0434, 1991.

[11] Logotheits,D.; Trivedi,K., "Transient Analysis of the Leaky Bucket Rate Control Scheme Under Poisson and ON-OFF Sources," *Proc. IEEE INFOCOM'94*, pp. 490-497, 1994.

[12] Lee,D.C., "Effects of Leaky Bucket Parameters on the Average Queueing Delay: Worst Case Analysis," *Proc. IEEE INFOCOM'94*, pp.482-489, 1994.

[13] Patel,B.; Bisdikian,C., "End-Station Performance under Leaky Bucket Traffic Shaping," *IEEE Network*, Vol. 10, No.5, pp.40-47, 1996.

[14] Faber,T.; Landweber,L.; Mukherjee, A., "Dynamic Time Windows: Packet Admission Control with Feedback," *Proc. ACM SIGCOMM'92*, pp.124-135, 1992.

[15] Chu,W., "Buffer Behavior for Poisson Arrivals and Multiple Synchronous Constant Outputs", *IEEE Trans. Computers*, Vol. C-19, No.6, pp.530-534, 1970.

[16] Chu,W., "Buffer Behavior for Batch Poisson Arrivals and Single Constant Output" *IEEE Trans. Commun.*, Vol.COM-18, No. 5, pp.613-618, 1970.

[17] Rathgeb,E., "Modeling and Performance Comparison of Policing Mechanisms for ATM Networks", *IEEE JSAC*, Vol.9, No. 3, pp.325-334, 1991.

学術情報センター紀要　第10号(1998)

研究論文

# FTP冗長トラフィックを削減するための探索ドメインモデル

# Search Domain Model for Reducing Redundant FTP Traffics

学術情報センター　藤野　貴之
Takayuki FUJINO
National Center for Science Information Systems

**要旨**

現在のarchie,検索エンジンといったファイル資源の検索システムは、実際的なネットワークトポロジを反映しない。そのため、時としてユーザはより遠方のanonymous ftpサーバからファイル資源を取得してしまい、結果として本来不要なはずの冗長なFTPトラフィックを発生させてしまっていた。

本稿では、新たに探索ドメインという概念を提案し、それを利用することによってネットワーク的により近いanonymous ftpサーバからファイル資源を取得できるようにすることを試みる。

**ABSTRACT**

Current file resource searching systems, such as archie,search engines don't reflect actual network topology. Because of this, user sometimes gets some files from anonymous ftp located in farther site and it occures redundant FTP traffics.

This paper discusses about new scheme 'search domain' and we attempt to make user to get files from topological nearby site.

[キーワード]　インターネット、冗長トラフィック、archie、検索エンジン、帯域の効率的利用
[Keywords]　Internet, redundant traffics, archie, search engine, efficient bandwidth use

## 1　はじめに

インターネットの普及とともに利用者が急増し、ネットワークトラフィックが急激に増大してきている。一般に、インターネットにおいて主に帯域を消費しているものは、HTTP[1]、FTP[2]、NNTP[3]等のTCP[4]上に実装されるプロトコルであると考えられる。

帯域の圧迫は、需要が供給を上回っているために生じる現象であり、各方面においてより効率的なネットワークの運用法が提案されている。代表的なものとして、HTTPトラフィックをローカルドメイン内にキャッシュさせることによる効率化を図るもの[5]、HTTPアクセスの集中するサーバをまとめてサーバファームを形成し、その場所へ高速回線をつなぎこむことで均一にトラフィックを分散させる仕組み[6]などが挙げられる。

本稿では、FTPを用いたデータ転送トラフィック(以下、FTPトラフィックと記す)の中で、本来は不要であるはずの冗長なトラフィックを削減する手法を提案する。一般に、ユーザがFTPを用いて何らかのデータ、プログラムを取得する場合、その配布場所がわからない場合には、何らかの手段でファイル資源の提供場所を探索しなければならない。現在では、archie[7]と呼ばれる資源検索システムを利用するか、あるいはgoo[8]、Altavista[9]といったロボット収集系の検索エンジンを使用するのが主流となっている。ここで、多くの場合、不幸なことに検索の結果はネットワークトポロジを反映させることができない。このため、ユーザの近辺に求める対象物があるにも関わらず、より遠方のFTPサーバにアクセスしてしまい、結果として冗長なFTPトラフィックを発生させてしまうことがあった。

この問題を解決するため、本稿では探索ドメインと呼ばれる概念を新たに導入し、インターネット全域に

おけるファイルの探索範囲、探索手順を階層化することを試みる。探索範囲を階層化することにより、はじめにユーザの近辺に限定した探索ドメインを探索し、その後により広い探索ドメインにおいて同様の対象物を検索する、といった段階的なファイルの検索が可能となる。探索ドメインを適宜設定することにより、従来の手法に比べてファイル検索の結果にネットワークトポロジを反映することが可能となるため、従来のような冗長なFTPトラフィックを削減することが可能となる。

本稿は初めに現在のファイル資源検索の状況およびその問題点について触れ、次に階層化された探索ドメインモデルを提案する。最後に階層化された探索ドメインにおけるファイル検索の実装例を紹介し、実際に所定の動作を行なえることを示す。

## 2　ファイルの検索と冗長なFTPトラフィック

### 2.1　ファイルの検索

anonymous ftpサーバ[10]は、インターネットを通して不特定多数のユーザに対し、様々なプログラム、データ、文書といった情報資源をFTPを通して提供することを目的として構築されたサーバである。ここで、anonymous ftpサーバが保持するコンテンツを“ファイル資源”と呼び、本稿全体を通してこの用語を用いることにする。

ユーザはFTPクライアントを用いてanonymous ftpサーバにアクセスし、求めるファイル資源を入手することができる。ネットワークニュース(NNTP)を用いた各種プログラムのソースファイル配布、ホームページを通した様々なファイル資源の配布なども行なわれている現在においても、FTPを利用したファイル資源の公開／配布は依然として有効な手段であり、インターネット社会において重要な位置を占めている。

インターネットの拡大と共にanonyous ftpサーバの数も増大し、各サーバが保持するファイル資源の量も膨大なものとなったため、ユーザが求めるファイル資源を検索するものとしてarchieが開発された。archieサーバは定期的にanonymous ftpサーバよりファイル資源情報を取得し、それを基にデータベースを作成、管理する。ユーザはarchieサーバにtelnet等の手段によって直接ログインするか、PROSPEROプロトコル[7]をサポートするarchieクライアントを通してarchieサーバに照会を行ない、求めるファイル資源に関する検索結果を入手することができる。検索結果は、

- ユーザが求めるファイル資源を保持するanonymous ftpサーバのホスト名
- ファイル資源が置かれているディレクトリパス

という組合せで出力される。図1は、archieサーバにarchie.kuis.kyoto-u.ac.jpを指定し、GNU Cコンパイラgcc-2.7.2.3.tar.gzの所在場所を検索した例である。

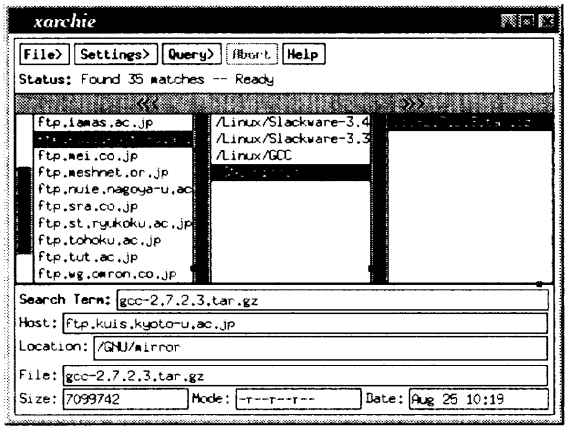



図1　archie検索例

その後、数多くのホームページの情報を検索するための検索エンジンが開発され、インターネット上の資源検索の中心的存在となった。ファイル資源の検索においても例外ではなく、欲しいファイル資源を検索エンジンで探す事例は増えてきている。図2は、先の例と同じgcc-2.7.2.3.tar.gzをgoo(http://www.goo.ne.jp)を用いて検索した例である。

### 2.2　冗長なFTPトラフィック

現在存在するarchie、検索エンジンといったファイル資源の探索手法では、ネットワークトポロジを反映した検索結果を得ることができない。すなわち、ユーザが求めるファイル資源を保持するanonymous ftpサーバはネットワーク的に遠くても近くても全く等価な結果としてユーザに渡される。そのため、ユーザの近辺に求めるファイル資源があるにも関わらず、より遠方のanonymous ftpサーバにアクセスしてしまい、結果として冗長なFTPトラフィックを発生させてしまうことがある。冗長なFTPトラフィックは、AS(Autonomous System)内で発生する形態、AS間で発生する形態に分類することができる。本節では、はじ

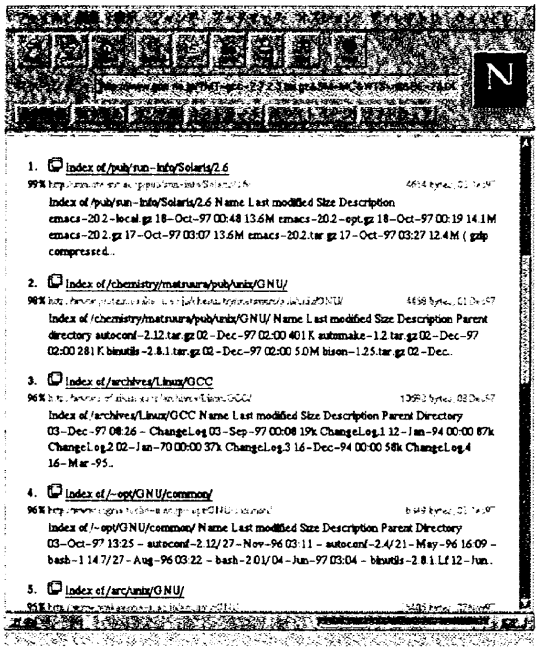


図 2　検索エンジンによる検索例

めに Autonomous System について簡単に概略を説明した後、冗長な FTP トラフィックに関する幾つかの例を挙げる。

### 2.2.1 Autonomous System

Autonomous System は、主に経路制御の分野において用いられるインターネットの管理単位の1つであり、一般に AS と略記される。AS は単一の管理権限により制御されるネットワークとルータの集合である。

図3は AS を概略的に示したものである。AS は NODE, NOC, POP 等と呼ばれる一つ以上のネットワーク集積ポイントを持つ(本稿では以下 NODE 表記を用いる)。NODE にはルータが置かれ、ネットワーク参加組織の回線を収容する。NODE 間はバックボーンと呼ばれる高速な専用回線で接続される。NODE に設置される各ルータにおいて、ネットワーク参加組織の経路情報を RIP[11][12]、OSPF[13]等の経路制御プロトコルによって交換する。各 NODE に置かれたルータ同士は自身の持つ経路情報を OSPF、IS-IS[14]等のプロトコルを用いて交換する。AS 内の全ての経路情報は、BGP[15]と呼ばれる AS 間経路制御プロトコルによって他 AS に向けて広告される。

AS は AS 番号と呼ばれる唯一の識別番号[16][17]を持つ。AS 間経路制御においては、AS 番号が重要な要素となる。AS の例として、SINET(AS2907)、JOIN

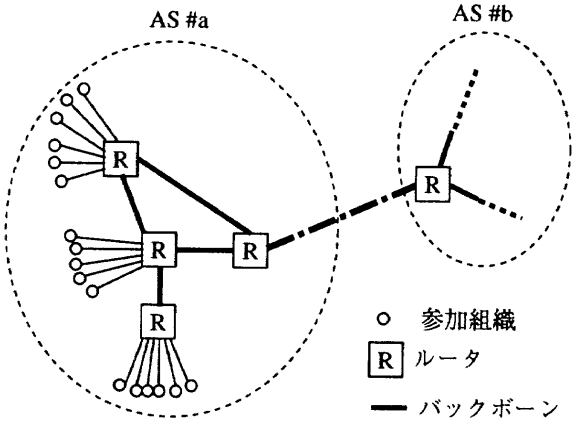



図 3　Autonomous System

(AS2498)、OCN(AS4713)等が挙げられる。AS 間経路制御は各 AS の経路制御技術管理者同士の協調によって成り立つため、各 AS の管理組織は自 AS に隣接している AS の AS 番号を知っていなければならない。インターネットトポロジは、AS 同士が互いに接続されている形と考えることができる。

### 2.2.2 AS 内冗長 FTP トラフィック

本節では、先に述べた AS の内部における冗長 FTP トラフィックの例を示す。

**例 1**　図4のような構成のネットワークを考える。すなわち、A 大学、B 大学はそれぞれインターネットに接続されており、互いに到達性が存在する。また、両大学は anonymous ftp サーバ ftp-A, ftp-B をそれぞれ構築しており、両サーバともファイル資源 X を保持していると仮定する。ここで、A 大学のユーザがファイル資源 X を入手しようとする場合、自大学内のサーバ ftp-A にアクセスすれば問題ないが、ftp-B からファイル資源を入手した場合、点線で示すような、本来は不要なはずの冗長 FTP トラフィックが発生する。

### 2.2.3 AS 間冗長トラフィック

本節では、AS 間に発生する二種類の冗長 FTP トラフィック例を示す。

**例 2**　図5のような構成のネットワークを考える。A 大学は anonymous ftp サーバを持たず、B 大学、C 大学はそれぞれ ftp-B、ftp-C を持ち、どちらもファイル資源 X を保持しているとする。ここで A 大学のユー

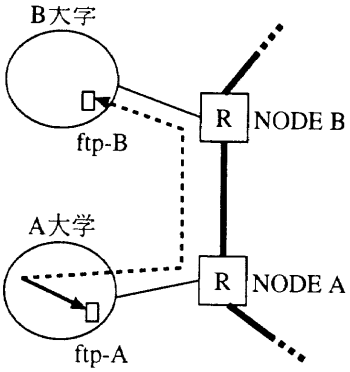



図 4　冗長 FTP トラフィック(1)

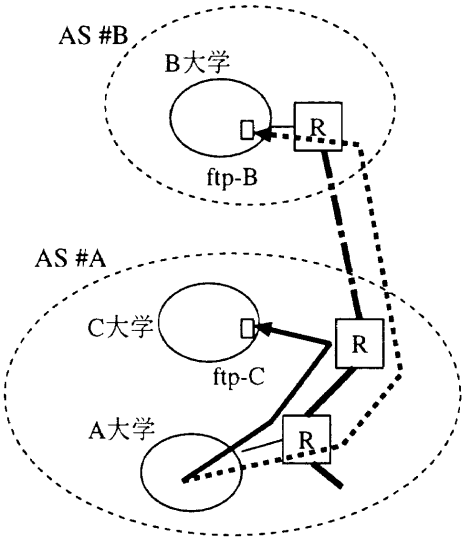



図 5　冗長 FTP トラフィック(2)

ザがファイル資源 X を入手しようとする場合、同一 AS 内にある ftp-C にアクセスすれば問題ないが、別の AS ＃B に属する ftp-B からファイル資源を入手した場合、点線で示すような、AS ＃A、AS ＃B 間の冗長な FTP トラフィックが発生する。

**例 3**　図6のような構成のネットワークを考える。A 大学、B 大学、C 大学の環境は先の例2に準じる。ここで、A 大学のユーザがファイル資源 X を入手しようとする場合、近隣の AS ＃B 内にある ftp-B にアクセスすれば問題ないが、AS ＃C 内にある ftp-C からファイル資源を入手した場合、点線で示すような AS ＃B、AS ＃C 間の冗長な FTP トラフィックが発生する。

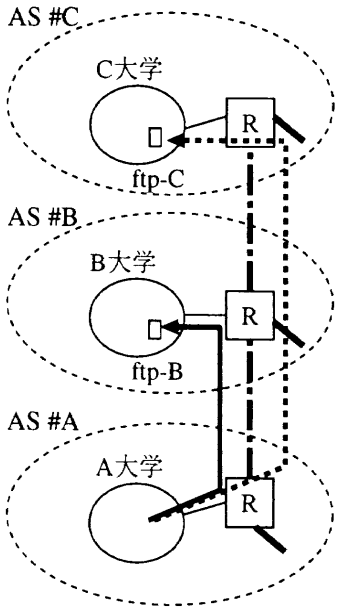



図 6　冗長 FTP トラフィック(3)

## 3　ファイル資源探索領域の階層化

冗長な FTP トラフィックを削減するためには、ファイル資源の検索結果に完全ではないにしてもある程度ネットワークトポロジを反映させる必要がある。本節では、本稿において新たに提案する探索ドメインと、階層化の手法について述べる。

### 3.1　探索ドメイン

探索ドメインとは、1つ以上の anonymous ftp サーバを含む物理的なある範囲である。例えば、考えられる最小の探索ドメインはある anonymous ftp サーバが置かれているネットワークのサブネットだけであり、最大の探索ドメインは全世界となる。

実際的なネットワーク運用を考える場合には、ファイル探索ドメインは以下のようなものが考えられる。

・ある大学、企業などの組織
・ある NODE に接続された参加組織全て
・ある AS に属する参加組織全て
・ある AS 及びその AS に隣接する AS の集合

探索ドメインには、必ず1つ以上のファイル資源情報サーバが存在しなければならない。逆接的に言えば、探索ドメインとは、ファイル資源情報サーバの管理者が管理する範囲と定義される。

ファイル資源情報管理サーバは、自らが管理する探

索ドメイン内に置かれているanonymous ftpサーバが持つファイル資源情報全てを管理し、ユーザからの問い合わせに対して自らのファイル資源情報を検索し、検索結果をユーザに返すものである。

**例4**　ある大学内だけを探索ドメインとするファイル資源情報管理サーバについて考える。学内LANの管理者がarchieサーバを構築し、学内に存在する全てのanonymous ftpサーバの保持する情報を管理すれば、そのarchieサーバはファイル資源情報管理サーバとなる。

### 3.2　探索ドメインの階層化

前節の例4のような運用においては、例1に示したような冗長FTPトラフィックは発生を防ぐことが可能となる。しかし、ユーザは学内のファイル資源探索、学外のファイル資源探索でいちいちarchieサーバを切替えなければならず、また学外においては依然としてネットワークトポロジを反映しない検索結果が出てしまう問題が残される。

この問題を解決させるために、ファイル資源情報管理サーバにある種のProxy機能が必要となる。図7は、その場合のファイル資源サーバの振る舞いを示している。すなわち、自らが保持するファイル資源情報にユーザの求めるファイル資源が存在しない場合には、ユーザの代わりに別のファイル資源情報サーバ(あるいは全世界を対象とするarchieサーバ)にアクセスし、その結果をユーザに返すというものである。

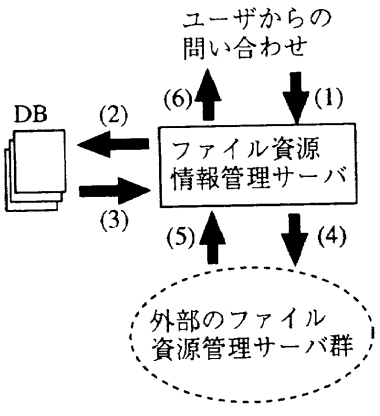



図 7　外部への問い合わせ代行

ファイル資源情報管理サーバがユーザの代理で外部に問い合わせを行なう場合、別のファイル資源情報管理サーバをひとつ指定し、そこに問い合わせを行なう。ここで、問い合わせを行なうものを“下位サーバ”、問い合わせを受けるものを“上位サーバ”と呼ぶ。上位サーバの振る舞いは図7と全く同様である。すなわち、下位サーバの問い合わせを受けて自身の保持するデータベースを検索し、そこに情報がなければ下位サーバの代理で自分の上位サーバに問い合わせを行なう。

ASの管理者が以下のようにファイル資源情報管理サーバを配置し、上位サーバの設定を行なうとする。

表 1　AS内部のサーバ設置例

| 管理範囲 | 上位サーバ |
|---|---|
| 各NODE接続組織<br>AS参加組織全て | AS全体を管理するサーバ<br>archie(全世界対象) |

この場合、AS参加組織は自組織内にファイル資源情報管理サーバを構築し、上位サーバとして最寄りのNODEのファイル資源情報管理サーバをポイントするだけで、

1. 組織内
2. 組織が属するNODEに接続される他組織
3. 組織が属するASに含まれる全組織
4. 全世界

といった、探索範囲、探索手順の階層化が可能となる。このような構成においては、AS内の冗長FTPトラフィックを相当削減することが期待できる。

## 4　ファイル資源情報管理サーバの実装例

前節において提案したファイル情報管理サーバをWWW上のCGIで実装した。このサーバが管理するanonymous ftpサーバはring.nacsis.ac.jpのみであり、上位サーバとしてarchie.kuis.kyoto-u.ac.jpを指定している。

図8は検索画面からgccを入力したところである。gccはring.nacsis. ac.jpが保持しているため、自探索ドメイン内に存在する(図9)。同様にして、ring.nacsis.ac.jpが保持しないetherfindを検索した場合には、上位サーバに問合せを行ない、その結果を表示している(図10)。どちらの場合においても、検索結果画面からファイルをダウンロードすることができる。

## 5　おわりに

本稿は現在のarchie、検索エンジンを使用したファイル資源入手に関する問題点について述べ、それを解

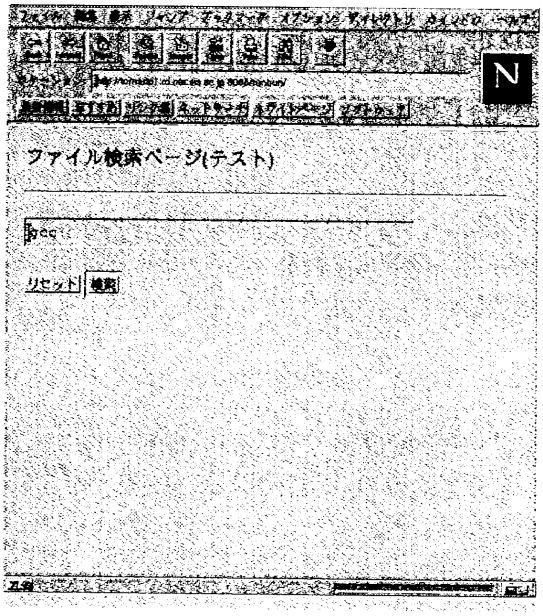



図 8　検索画面


図 9　自探索ドメイン内での検出

決するための階層化された探索ドメインのモデルを提案した。最後にCGIで実装したファイル資源情報管理サーバの例を示し、実際に所定の動作が行なわれることを示した。

現在までの実装では、例2、例3に示したようなAS間の冗長トラフィックを削減することができない。AS間冗長トラフィックを削減するためには、複数の隣接

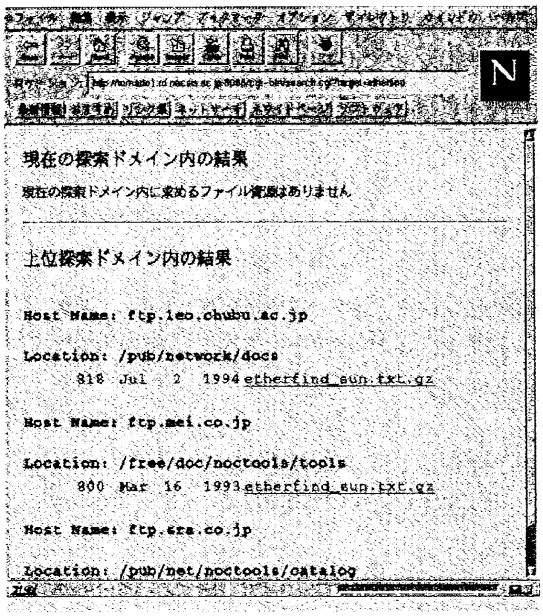



図 10　上位サーバへの問合せ結果

したASのうち、どこに求めるファイル資源があるのかを調べなければならない。

ひとつの解決法として、隣接する全てのASに置かれているファイル資源情報管理サーバに問合せを行なうことが考えられるが、それはかえって無用なトラフィックの増大を招く可能性がある。ある程度問合せのデータをキャッシュしながら、過去にヒットした隣接ASに問合せを行なうなど、戦略的な方法も検討中であるが、それらの実装は今後の課題である。

## 参考文献

[1] Fielding,R.; Gettys,J.; Mogul,J.; Frystyk, H.; Berners-Lee T., Hypertext Transfer Protocol - HTTP/1.1, RFC2068, 1997.

[2] Postel, J.; Reynolds, J., File Transfer Protocol, STD9, RFC959, 1985.

[3] Kantor,B.; Lapsley,P., Network News Transfer Protocol: A Proposed Standard for the Stream-Based Transmission of News, RFC977, 1986.

[4] Postel,J., Transmission Control Protocol, STD7, RFC793, 1981.

[5] http://squid.nlanr.net/Squid/.

[6] http://www.mfeed.ad.jp/.

[7] Emtage,A.; Deutsch,P., archie - An Electronic Directory Service for the Internet,

*USENIX Winter* 1992 *Technical Conference Proceedings*, pp.93-110, 1992.

[8] http://www.goo.ne.jp/.

[9] http://altavista.digital.com/.

[10] Deutsch,P.; Emtage,A.; Marine,A., How to Use Anonymous FTP, FYI24, RFC1635, 1994.

[11] Hedrick,C., Routing Information Protocol, STD34, RFC1058, 1988.

[12] Malkin,G., RIP Version 2 Carrying Additional Information, RFC1723, 1994.

[13] Moy,J., OSPF Version 2, RFC2178, 1997.

[14] Oran,D., OSI IS-IS Intra-domain Routing Protocol, RFC1142, 1990.

[15] Rekhter,Y.; Li,T., A Border Gateway Protocol 4 (BGP-4), RFC1771,1995.

[16] Hawkinson,J.; Bates,T., Guidelines for creation, selection, and registration of an Autonomous System (AS), BCP6, RFC1930, 1996.

[17] http://www.nic.ad.jp/jpnic/ipaddress/as-numbers.txt

学術情報センター紀要　第10号(1998)

研究論文

# 項目反応パタンとロジスティックモデル

# Item Response Patterns and Logistic Models

学術情報センター　孫　媛
Yuan SUN
National Center for Science Information Systems

**要旨**

項目反応理論における項目特性曲線として現在最も広く用いられているのは、ロジスティック関数である。項目特性曲線は当初、累積正規分布関数としてLordによって導かれたが、累積正規分布関数と近い形状をしている上に、数学的にさまざまなよい性質を持っているロジスティック関数をかわりに使うことをBirnbaumが提案したのである。本研究においては、IRTを前提とせず、項目反応パタンに基づいて直接にロジスティック・モデルを導くことを試みる。能力特性値に対して新たな定義を与え、被験者の能力と被験者・項目の相互作用を平均情報量と関係づける。これにより、項目反応理論に新たな角度から光が当てられると考える。

**ABSTRACT**

The first item response model was the normal-ogive model which postulated normal cumulative distribution function as a response function for the item. Because logistic cdf is very close to normal cdf but has more mathematical advantages over it, Birnbaum suggested to replace the normal-ogive model by the logistic model, which is now one of the most popular unidimensional item response models. The purpose of this paper is to obtain logistic models directly from item response patterns without the assumption of IRT models. Meanwhile, in this paper a statistic definition for subject's latent traits, the relationship between subject's and items' parameters with the average amount of information will be made based on the item response patterns.

[キーワード]　項目反応理論、項目反応パタン、ロジスティックモデル、項目特性関数、個人特性関数、項目困難度、能力特性値、平均情報量

[Keywords]　IRT, item response patterns, logistic models, item response function, person response function, item difficulty, ability, average amount of information.

## 1　従来のロジスティックモデル

テスト項目に対する被験者の正答確率が、被験者特性の関数として表されるというアイディアは、Binetにまで遡ることができる(Binet & Simon, 1916 [1])。Binetの知能検査において注目された被験者特性は〝年齢〟であったが、彼の素朴なアイディアは、その後、Lawley (1943 [5], 1944 [6])やTucker (1946 [15])などによって、数学的モデルとして洗練されていく。Tuckerは、被験者特性と正答確率の関数関係を、項目特性関数と名付けている。項目反応理論が、テスト理論の領域で今日のように盛んに研究される基礎を作ったのは、Lord (1952) [7]である。Lordは当初、項目特性関数をつぎのような累積正規分布関数で表した(Lord, 1953a [8], 1953b [9], Lord & Novick, 1968 [11])。

$$P_j(\theta)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a_j(\theta-b_j)} exp\left(-\frac{1}{2}z^2\right)dz$$

Birnbaum[2]はこの項目特性関数を、ロジスティック累積分布関数に置き換え、つぎのような2パラメータ・ロジスティック・モデルを提案した。

$$P_j(\theta)=\frac{1}{1+e^{-Da_j(\theta-b_j)}}$$

ここで、$D$は能力特性値$\theta$を累積正規モデルによる尺度と対応づけるための定数で、$D=1.7$とおけば、2パラメータ累積正規モデルと2パラメータ・ロジスティック・モデルにおける$P_j(\theta)$の値の差の絶対値を、$\theta$の全範囲にわたって、0.01より小さくすることができる(Haley, 1952 [3])。$a_j$および$b_j$は項目$j$の特徴を表すパラメータで、それぞれ項目の識別力と困難度と呼ばれる。項目反応理論の出発点は累積正規モデルであったが、ロジスティックモデルの方が数学的な取り扱いが容易であるために、現在ではロジスティックモデルが使われることの方が多い(Hambleton & Swaminathan, 1985 [4])。

## 2 IRTを前提としないロジスティックモデル

上に見たように、従来のロジスティックモデルは、累積正規モデルの代替として提起されたものである。以下では、ある一定の能力を持つ被験者が多数の項目に繰り返し反応した場合に得られる項目反応パタンから、項目反応理論を前提としないロジスティックモデルが導かれることを示す。

### 2.1 項目反応パタンに基づくモデルの導出

いま、ある一定の能力を持つ被験者が、$m$個の項目のそれぞれに対して$Y_j(j=1,2,\cdots,m)$回の反応をしたデータを想定する。このとき、項目$j$に対する第$i$回目の反応を$U_j^i$と表す($j=1,2,\cdots,m$；$i=1,2,\cdots,Y_j$)と、反応パタンを以下のようにまとめることができる。

| | 項目 | | | | | |
|---|---|---|---|---|---|---|
| 反応 | 1 | 2 | $\cdots$ | $j$ | $\cdots$ | $m$ |
| 1 | $U_1^1$ | $U_2^1$ | $\cdots$ | $U_j^1$ | $\cdots$ | $U_m^1$ |
| 2 | $U_1^2$ | $U_2^2$ | $\cdots$ | $U_j^2$ | $\cdots$ | $U_m^2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $i$ | $U_1^i$ | $U_2^i$ | $\cdots$ | $U_j^i$ | $\cdots$ | $U_m^i$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $U_2^{Y_2}$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\vdots$ | $U_1^{Y_1}$ | | $\vdots$ | $\vdots$ | $\vdots$ | $U_m^{Y_m}$ |
| $Y_j$ | | | | $U_j^{Y_j}$ | | |
| 正答回数 | $X_1$ | $X_2$ | $\cdots$ | $X_j$ | $\cdots$ | $X_m$ |

ここで、

$$U_j^i=\begin{cases}1 & (\text{正答のとき})\\ 0 & (\text{誤答のとき})\end{cases}$$

とすれば、項目$j$に対する正答回数$X_j$はつぎのように表される。

$$X_j=\sum_{i=1}^{Y_j}U_j^i,\quad j=1,2,\cdots,m \tag{1}$$

さて、項目$j$に対する$Y_j$回の反応において生じうる1-0(正誤)パタンは、$Y_j, X_j$を与えたとすると、

$$C_{Y_j}^{X_j}=\frac{Y_j!}{(Y_j-X_j)!X_j!}$$

通りある。さらに、被験者の各項目に対する反応が独立であることを仮定すれば、$m$個の項目に対して可能な1-0パタン数は、

$$\Omega=\prod_{j=1}^{m}\frac{Y_j!}{(Y_j-X_j)!X_j!} \tag{2}$$

となる。各項目への正答回数$X_j$によって$\Omega$の値が変わることは明らかであるが、$X_j$は被験者の特性と項目の特性によって規定される。すなわち、項目の難しさが同じであれば能力の高い人ほど正答回数は多くなり、解答者の能力が同じならば易しい項目において正答回数が多くなる。ある一定($\theta_0$)の能力を持つ被験者が、$m$項目のそれぞれに対して反応する回数$Y_j$が十分大きいときには、各項目への正答回数が安定し、正答回数の総和が一定の値($N_{(\theta_0)}$)になることが期待される。すなわち、

$$N_{(\theta_0)}=\sum_{j=1}^{m}X_j \tag{3}$$

また、各項目の正答数と困難度の積和が一定($B_{(\theta_0)}$)であること、すなわち、

$$B_{(\theta_0)}=\sum_{j=1}^{m}X_jb_j \tag{4}$$

も仮定する(後に従来のモデルと比較することを考えて、本モデルでも従来と同じく、被験者能力を$\theta$、項目の難しさを$b$と表すことにする)。

このような2つの仮定の下で$\Omega$を最大にするような$X_j$を求めれば、被験者の$Y_j$回の応答のうち、正答数が$X_j$回である可能性が最も大きいという意味になる。式(2)の両辺の対数をとると、

$$\ln\Omega=\sum_{j=1}^{m}[\ln Y_j!-\ln(Y_j-X_j)!-\ln X_j!] \tag{5}$$

が得られる。$Y_j\gg1$, $X_j\gg1$であれば、Stirlingの公式$n!\approx\sqrt{2\pi n}n^ne^{-n}$を用いることができる。その対数をとれば、

$$\ln n! \approx n(\ln n-1)+\frac{1}{2}\ln 2\pi n$$

であるが、$n$が十分に大きいとき、第2項は第1項と比べて無視できるほど小さくなるので、さらに

$$\ln n! \approx n(\ln n-1)$$

という近似式を得る。これを用いると、式(5)は以下のようになる。

$$\begin{aligned}\ln\Omega &= \sum_{j=1}^{m}[\,Y_j(\ln Y_j-1)-(Y_j-X_j)\\ &\quad\times(\ln(Y_j-X_j)-1)\\ &\quad -X_j(\ln X_j-1)\,]\\ &= \sum_{j=1}^{m}[\,Y_j\ln Y_j-(Y_j-X_j)\\ &\quad\times\ln(Y_j-X_j)-X_j\ln X_j\,] \qquad (6)\end{aligned}$$

ここで、上記の2つの制約条件の下で$\ln\Omega$を最大化するために、Lagrangeの未定乗数$\beta$と$\alpha$を用い、

$$\begin{aligned}&\delta\Big(\ln\Omega-\beta\Big(\sum_{j=1}^{m}X_j-N\Big)\\ &\quad-\alpha\Big(\sum_{j=1}^{m}X_jb_j-B\Big)\Big)=0 \qquad (7)\end{aligned}$$

を解くことを考える。$N$と$B$は定数だから、$\delta N=\delta B=0$となり、方程式(7)はつぎのようになる。

$$\begin{aligned}&\sum_{j=1}^{m}\delta[\,Y_j\ln Y_j-(Y_j-X_j)\ln(Y_j-X_j)\\ &\quad -X_j\ln X_j-\beta X_j-\alpha(X_jb_j)\,]=0\end{aligned}$$

さらに整理すると、

$$\sum_{j=1}^{m}[\ln(Y_j-X_j)-\ln X_j-\beta-\alpha b_j]\delta X_j=0 \qquad (8)$$

$\delta X_j$がどんな値をとっても方程式が成立するためには、以下の式が成立しなければならない。

$$\ln\frac{Y_j-X_j}{X_j}-\beta-\alpha b_j=0 \qquad (9)$$

つまり、

$$\frac{X_j}{Y_j}=\frac{1}{1+e^{\beta+\alpha b_j}} \qquad (10)$$

という関係が成り立つとき、$\Omega$は最大化される。$Y_j$が十分大きいとき、被験者が項目$j$に正答する確率$P_{j(\theta_0)}$は$\frac{X_j}{Y_j}$で近似することができる。したがって、

$$P_{j(\theta_0)}=\frac{1}{1+e^{\beta+\alpha b_j}} \qquad (11)$$

となり、項目反応理論を前提とせず、項目反応パタンに基づいて直接にロジスティクモデルが導かれることが示されている。

### 2.2　能力特性値の新たな定義

式(11)を$b(-\infty,\ +\infty)$の関数と見なすと、$P_{\theta_0}$を$\theta_0$の個人の正答確率の変化を表すロジスティック個人反応関数(Person Response Function: PRF)と解釈することができる。図1は、個人反応関数の例($\alpha=2$, $\beta=-2$)である。この関数は困難度$b$の単調減少関数であり、困難度$b$の値が大きいほど、被験者$\theta_0$が正答する確率$P_{(\theta_0)}$は低くなる(ただし、$\alpha>0$)。式(11)を書き直すと、

$$P_{\theta_0}(b)=\frac{1}{1+e^{-\alpha(-\frac{\beta}{\alpha}-b)}}$$

となり、$b=-\frac{\beta}{\alpha}$とおくと$P_{(\theta_0)}=0.5$となることがわかる。そこで、被験者の能力特性値を、正答確率が50%となる項目困難度$b$の値

$$\theta_0\equiv-\frac{\beta}{\alpha} \qquad (12)$$

によって定義することにする。それによって、被験者が項目$j$に正答する確率は以下のように表される。

$$P_{\theta_0}(b)=\frac{1}{1+e^{-\alpha(\theta_0-b)}} \qquad (13)$$

曲線の変曲点$b=\theta_0$における傾きは、式(13)を$b$について微分し、$b=\theta_0$を代入した値、すなわち、

$$\left.\frac{\partial P_{\theta_0}(b)}{\partial b}\right|_{b=\theta_0}=-\frac{1}{4}\alpha \qquad (14)$$

に等しくなる。つまり、パラメータ$\alpha$の値は曲線の変曲点$\theta_0=b$における傾きと関連し、$\alpha$の値が大きいほど、$\theta_0=b$における傾きが大きくなる。$\alpha\to\infty$のとき、$P(\theta_0>b)=1$、つまり、被験者は自分の能力$\theta_0$より小さい困難

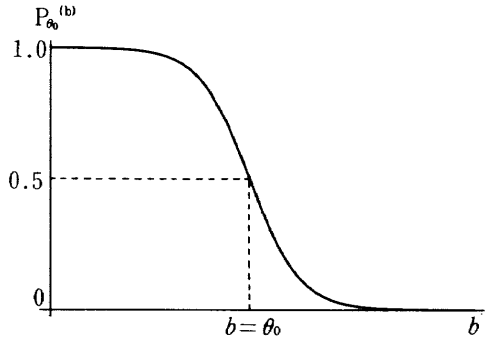



図 1　個人反応曲線の例示

度をもつ項目すべてに正答し、逆に$P(\theta_0<b)=0$、つまり、自分の能力より難しい項目に正答することがまったく不可能ということになる。

一方、同様に能力特性値の関数として、ある項目jへの正答確率を表すと、以下のように従来の項目特性関数(Item Response Function: IRF)に相当した関数が得られる。

$$P_j(\theta)=\frac{1}{1+e^{-\alpha'(\theta-b_j)}} \tag{15}$$

この関数は$\theta$の単調増加関数であり、能力特性値$\theta$の値が大きいほど、項目に正答する確率$P_j(\theta)$は高くなる。パラメータ$\alpha'$と$(\theta-b_j)$の積が項目特性関数の指数部分にあり、$\alpha'$が能力$\theta$と困難度$b_j$との差による正答確率の変化に影響を与えていることが明らかである。

$\alpha$と$\alpha'$は、被験者と項目の相互作用に関連するパラメータと考える(孫・芝, 1990[13]；芝, 1991[12]；孫, 1997[14])。被験者が項目に反応するとき、能力特性値$\theta$がテスト項目とは関係なく、その被験者の固有のものとして一点に固定されていると仮定すれば、モデル(15)が従来の2パラメータ・ロジスティツクIRTモデルと一致することがわかる。

## 3　個人および項目パラメータと平均情報量との関係

式(3)と式(4)に式(10)を使うと、

$$N_{(\theta_0)}=\sum_{j=1}^{m}X_j=\sum_{j=1}^{m}\frac{Y_j}{1+e^{\beta+\alpha b_j}} \tag{16}$$

$$B_{(\theta_0)}=\sum_{j=1}^{m}X_jb_j=\sum_{j=1}^{m}\frac{b_jY_j}{1+e^{\beta+\alpha b_j}} \tag{17}$$

ここで、以下のようなGreen関数

$$\Xi_j=[1+e^{-(\beta+\alpha b_j)}]^{-Y_j} \tag{18}$$

$$\Xi=\prod_{j=1}^{m}\Xi_j=\prod_{j=1}^{m}[1+e^{-(\beta+\alpha b_j)}]^{-Y_j} \tag{19}$$

を定義すると、

$$\ln\Xi=\sum_{j=1}^{m}\ln\Xi_j=-\sum_{j=1}^{m}Y_j\ln[1+e^{-(\beta+\alpha b_j)}] \tag{20}$$

となり、

$$-\frac{\partial}{\partial\beta}\ln\Xi=N_{(\theta_0)} \tag{21}$$

$$-\frac{\partial}{\partial\alpha}\ln\Xi=B_{(\theta_0)} \tag{22}$$

となることがわかる。したがって、

$$\alpha dB=-\alpha d\left(\frac{\partial}{\partial\alpha}\ln\Xi\right) \tag{23}$$

ここで、

$$d\ln\Xi=\frac{\partial}{\partial\beta}\ln\Xi d\beta+\frac{\partial}{\partial\alpha}\ln\Xi d\alpha$$

を用いると、以下の式が導かれる(付録Aを参照)。

$$\alpha dB=d\left(\ln\Xi-\beta\frac{\partial}{\partial\beta}\ln\Xi-\alpha\frac{\partial}{\partial\alpha}\ln\Xi\right)-\beta dN \tag{24}$$

一方、情報理論に基づいて、まず情報量と平均情報量のエントロピー(Entropy)を以下のように定義する。

$$I(P_k)=-\ln P_k$$

$$H(P)=-c\sum_{k=1}^{\Omega}P_k\ln P_k$$

ただし、$\Omega$は可能な1-0パタン状態総数、$P_k$は各パタンが得られる確率、$c$は定数である。各パタンが等しい確率で現れると仮定するため、$P_k=\frac{1}{\Omega}$であるがわかる。それにより、

$$H(P)=-c\sum_{k=1}^{\Omega}\frac{1}{\Omega}\ln\frac{1}{\Omega}=-c\ln\frac{1}{\Omega}=c\ln\Omega$$

となり、式(6)を代入すると、

$$H(P)=c\sum_{j=1}^{\Omega}[\,Y_j\ln Y_j-(Y_j-X_j)\times\ln(Y_j-X_j)-X_j\ln X_j \tag{25}$$

となる。さらに、

$$H(P)=c\left(\ln\Xi-\beta\frac{\partial}{\partial\beta}\ln\Xi-\alpha\frac{\partial}{\partial\alpha}\ln\Xi\right) \tag{26}$$

となることが示される(付録Bを参照)。式(24)と式(26)を比較すると、以下の式が成立する。

$$\alpha dB=\frac{1}{c}dH-\beta dN \tag{27}$$

これより、被験者の能力特性値、項目の困難度、被験者・項目の相互作用と反応パタンにおける平均情報量との関係が結びつけられた。さらに、

$$\left.\frac{\partial B}{\partial N}\right|_H=-\frac{\beta}{\alpha}$$

であり、

$$\theta=\left.\frac{\partial B}{\partial N}\right|_{H} \tag{28}$$

となることがわかる。これによって、能力$\theta$の定義についてパタンの平均情報量からも考察、解釈できることが示された。

## 参考文献


[1] Binet,A.; Simon,T., *The development of intelligence in young children*, The Training School, 1916.

[2] Birnbaum,A., "Some latent trait models and their use in inferring an examinee's ability". In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

[3] Haley,D.C., Estimation of the dosage mortality relationship when the dose is subject to error (*Technical Report No. 15*). Stanford, CA: Stanford University, Applied Mathematics and Statistics laboratory, 1952.

[4] Hambleton,R.K.; Swaminathan,H., *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers, 1985.

[5] Lawley,D.N., On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edingburgh*, Vol.61, pp.273-287, 1943.

[6] Lawley,D.N., The factorial analysis of multiple item test. *Proceedings of the Royal Society of Edingburgh*, Vol.62, pp.74-82 1944.

[7] Lord,F.M., A theory of test scores. *Psychometric Monograph No.7*, Psychometric Society, 1952.

[8] Lord,F.M., An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika*, Vol.18, pp.57-75, 1953a.

[9] Lord,F.M., The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, Vol.13, pp.517-548, 1953b.

[10] Lord,F.M., *Application of item response theory to practical testing problems*. New York: Lawrence Erlbaum Associates, 1980.

[11] Lord,F.M.; Novick,M.R., *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

[12] 芝祐順編,「項目反応理論――基礎と応用――」, 東京大学出版会, 1991

[13] 孫媛・芝祐順,「特異な反応パタンを示す被験者の能力推定――一般項目反応理論の適用」, 教育心理学研究, Vol.38, No.4, pp.360-368, 1990.

[14] 孫媛,「多次元データに対する項目反応モデル」, 学術情報センター紀要, No.9, pp.103-112, 1997.

[15] Tucker,L.R., Maximum validity of a test with equivalent items. *Psychometrika*, Vol.11, pp.1-13, 1946.


## 付録A

等式(24)の右辺の3つの項は、それぞれ

$$d\ln\Xi=\frac{\partial}{\partial\beta}\ln\Xi\,d\beta+\frac{\partial}{\partial\alpha}\ln\Xi\,d\alpha$$

$$d\left(-\beta\frac{\partial}{\partial\beta}\ln\Xi-\alpha\frac{\partial}{\partial\alpha}\ln\Xi\right)$$

$$=-\frac{\partial}{\partial\beta}\left(\beta\frac{\partial}{\partial\beta}\ln\Xi\right)d\beta-\frac{\partial}{\partial\alpha}\left(\beta\frac{\partial}{\partial\beta}\ln\Xi\right)d\alpha-\frac{\partial}{\partial\beta}\left(\alpha\frac{\partial}{\partial\alpha}\ln\Xi\right)d\beta-\frac{\partial}{\partial\alpha}\left(\alpha\frac{\partial}{\partial\alpha}\ln\Xi\right)d\alpha$$

$$=-\frac{\partial}{\partial\beta}\ln\Xi\,d\beta-\beta\frac{\partial^2}{\partial\beta^2}\ln\Xi\,d\beta-\beta\frac{\partial}{\partial\alpha}\frac{\partial}{\partial\beta}\ln\Xi\,d\alpha-\alpha\frac{\partial}{\partial\beta}\frac{\partial}{\partial\alpha}\ln\Xi\,d\beta-\frac{\partial}{\partial\alpha}\ln\Xi\,d\alpha-\frac{\partial^2}{\partial\alpha^2}\ln\frac{\partial}{\partial\beta}d\alpha$$

$$\beta dN=\beta d\left(-\frac{\partial}{\partial\beta}\ln\Xi\right)=-\beta\frac{\partial^2}{\partial\beta^2}\ln\Xi d\beta-\beta\frac{\partial}{\partial\alpha}\frac{\partial}{\partial\beta}\ln\Xi d\alpha$$

となる。これらを式(24)の右辺に代入して整理すると、

$$\begin{aligned}&d\ln\Xi+d\left(-\beta\frac{\partial}{\partial\beta}\ln\Xi-\alpha\frac{\partial}{\partial\alpha}\ln\Xi\right)+\beta dN\\&=-\alpha\frac{\partial}{\partial\beta}\left(\frac{\partial}{\partial\alpha}\ln\Xi\right)d\beta-\alpha\frac{\partial^2}{\partial\alpha^2}\ln\Xi d\alpha\\&=\alpha d\left(-\frac{\partial}{\partial\alpha}\ln\Xi\right)\\&=\alpha dB\end{aligned}$$

となる。したがって、

$$\alpha dB=d\left(\ln\Xi-\beta\frac{\partial}{\partial\beta}\ln\Xi-\alpha\frac{\partial}{\partial\alpha}\ln\Xi\right)-\beta dN$$

が得られる。

## 付録B

式(10)　$\dfrac{X_j}{Y_j}=\dfrac{1}{1+e^{\beta+\alpha b_j}}$

より、

$$e^{\beta+\alpha b_j}=\frac{Y_j-X_j}{X_j}$$

が得られるため、式(20)は

$$\ln\Xi=-\sum_{j=1}^{m}Y_j\ln(1+e^{-\beta-\alpha b_j})=-\sum_{j=1}^{m}Y_j\ln(1+\frac{X_j}{Y_j-X_j})$$

となる。この式を代入すると、式(26)の右辺は

$$\begin{aligned}&\ln\Xi-\beta\frac{\partial}{\partial\beta}\ln\Xi-\alpha\frac{\partial}{\partial\alpha}\ln\Xi\\&=-\sum_{j=1}^{m}Y_j\ln\left(1+\frac{X_j}{Y_j-X_j}\right)+\sum_{j=1}^{m}\beta\frac{\partial}{\partial\beta}(Y_j\ln(1+e^{-\beta-\alpha b_j}))\\&\quad+\sum_{j=1}^{m}\alpha\frac{\partial}{\partial\alpha}(Y_j\ln(1+e^{-\beta-\alpha b_j}))\\&=-\sum_{j=1}^{m}Y_j\ln Y_j+\sum_{j=1}^{m}Y_j\ln(Y_j-X_j)+\sum_{j=1}^{m}\beta Y_j\frac{-e^{-\beta-\alpha b_j}}{1+e^{-\beta-\alpha b_j}}+\sum_{j=1}^{m}\alpha Y_j\frac{-b_je^{-\beta-\alpha b_j}}{1+e^{-\beta-\alpha b_j}}\\&=-\sum_{j=1}^{m}Y_j\ln Y_j+\sum_{j=1}^{m}Y_j\ln(Y_j-X_j)-\sum_{j=1}^{m}(\beta Y_j+\alpha Y_jb_j)\frac{1}{1+e^{\beta+\alpha b_j}}\\&=-\sum_{j=1}^{m}Y_j\ln Y_j+\sum_{j=1}^{m}Y_j\ln(Y_j-X_j)-\sum_{j=1}^{m}Y_j\left(\ln\frac{Y_j-X_j}{X_j}\right)\frac{X_j}{Y_j}\end{aligned}$$

$$= -\sum_{j=1}^{m} Y_j \ln Y_j + \sum_{j=1}^{m} Y_j \ln (Y_j - X_j) - \sum_{j=1}^{m} X_j \ln (Y_j - X_j) + \sum_{j=1}^{m} X_j \ln X_j$$

$$= -\sum_{j=1}^{m} Y_j \ln Y_j + \sum_{j=1}^{m} (Y_j - X_j) \ln (Y_j - X_j) + \sum_{j=1}^{m} X_j \ln X_j$$

となる。これを式(25)と比較すると、

$$H(P) = c\left(\ln \Xi - \beta \frac{\partial}{\partial \beta} \ln \Xi - \alpha \frac{\partial}{\partial \alpha} \ln \Xi\right)$$

であることがわかる。

学術情報センター紀要　第10号(1998)

研究論文

# ミドルウェアを用いた大規模なWebベースアンケート調査票の開発と回答者による利用の実態

# A development of large scale Web based questionnaire forms with a middleware and a profile of the answerers

学術情報センター　西澤　正己
Masaki NISHIZAWA
National Center for Science Information Systems

**要旨**

Webサーバ・アプリケーションツールキットであるWebObjectsを用いてWebベースのアンケート調査票フォームを開発し、郵送の回答様式と合わせて大学研究者を対象として調査をおこなった。調査票が大きく回答フォームとして8ページ、39項目あったがアプリケーションの工夫によって、附帯調査を除きページごとの回答漏れはほとんどなかった。しかし、調査票の規模が大きかったためWebベースの回答率は10.2%にとどまった。また、研究分野ごとに調べたところ、Webベースでの回答率の高い分野は部レベルでは理学(13.4%)、工学(12.8%)、全体としての回答者が50名以上の分科レベルでは情報科学(23.5%)、応用物理学・工学基礎(18.9%)であった。

**ABSTRACT**

A questionaire forms with WebObjects as a web server application toolkit were developed. The respondent can choose reply format from Web based form and ordinary postal mail, alternatively. The ratio of respondents choosing the Web based form is only 10.2% because of comparatively large questionaire form with 32 items in 8 submission pages. The ratio of answers including missing submission pages are about only 1% with contrived application format. The ratios of Web based respondents by research field are also presented.

[キーワード]　アンケート調査、研究者、WWW、ミドルウェア、研究環境
[keywords]　questionaire survey, researcher, WWW, middleware, research environment

## 1　はじめに

ここ数年ネットワーク接続が容易なパーソナルコンピュータの普及により、コンピュータネットワークを利用したアンケート調査が盛んに行われるようになってきた。コンピュータネットワークを利用したアンケート調査では、即時性やペーパーメディアでは難しかったユーザインターフェースの高機能化によるわかりやすい説明の提供等の利用者側のメリットの他、調査主体側ではデータ入力の不必要化、入力時の間違いの防止、処理時間の大幅な削減、コストの削減等大きなメリットが得られる。

現時点でネットワークを用いておこなわれているアンケート調査の大多数は、WWWを利用し、一方的に特定のURL上で掲示・募集したアンケートに対してそれを見て興味を持った人が回答するスタイルがとられている。これではある目的を持って特定のコミュニティの中から無作為抽出したサンプルを得ることはできない。回答項目が少ない場合は電子メールも利用できるが、回答フォーマットを固定しにくいので特に規模が大きくなった場合に、自動的なデータベース化やインタラクティブな間違い訂正が難しくなる。

ここで考えるアンケートの対象者は、主に大学等の研究者であり研究動向、研究基盤等の調査への応用、特に将来的には学術情報センターで実施している研究者ディレクトリ調査等への応用を目標としている。また、これまで全分野の大学研究者を対象としたネット

ワーク経由でのアンケート調査はほとんどなく、回答率がどの程度となるかの結果もない。

この報告書で述べるWebベースのアンケート調査票は「大学の研究者をとりまく研究環境に関する調査(科学研究費基盤研究A、代表: 太田和良幸)」[1]で使用するために開発した。この調査では、調査対象者を科学研究費補助金の「系・部・分科・細目表」の分科レベルに基づいて層化したうえで、サンプリングし、研究分野ごとの違いを明らかにすることを目的とした。このため調査対象が約8500人となり、結果のデータ入力、訂正等に困難が予想されたため、できるだけネットワークを通じての回答を得るためにWebベースの回答フォームを開発した。しかし、無作為抽出であるので、回答方法としては従来の方法である郵送との2種類の回答方法を用意した。このことによって、分野ごとのインターネットによる回答率も得られ、今後の調査の貴重な資料となった。

Webベースのアンケート調査票は混乱を避けるためできるだけ紙面での調査票と同様な書式とし、間違いを避けるためのプログラムも必要である。これまで、WebベースのアンケートCGIとshやPerl等のスクリプトを使って書かれるのが一般的であった。しかし、この方法では質問項目が多いアンケート等での変数の取り扱いや、多くのスクリプト起動による処理能力の低下、ファイルへの書き込み時のトランザクションの管理等で問題が出てくる。また、大きなアプリケーションになるごとにスクリプトの再利用性の低さから、アンケートごとにほぼ全体を書き換えなければならないという事態になり、アプリケーションの開発時間が長くかかってしまう。このような問題点を解決する手段として、ここ近年急速に進歩したWebサーバ・アプリケーションツールキットが挙げられる。これらの大部分は、自社製のDBMSへのアクセスを念頭に置いたものが多く、特定のDBMSにアクセスして動的なWebのページを表現するには適しているが、状態の管理やWebアプリケーションの異なるページ間での変数のやりとり等においてはやや弱いものが多い。この点ではWebObjects(NeXT Software)は優れており、今回のアプリケーション開発に採用することとした。

ここでは、この調査によって得られたデータをもとに、大学研究者を対象としたWebベースのアンケート調査の可能性と問題点について議論する。

## 2　調査対象

このアンケート調査の対象者は、国・公・私立大学(大学共同利用機関等を含む)の研究者8,455人である。これらの対象者は、学術情報センターが調査およびデータベース化をおこなっている「研究者ディレクトリ(平成7年度版)」に登録されている約13万人の研究者で国・公・私立大学および大学共同利用機関に所属する研究者中のから、研究分野を「科学研究費補助金系・部・分科・細目表」の分科レベル(一部例外あり)に基づく61分野に分けてサンプリングをおこなった。サンプリングに際しては、分野ごとのサンプリングに偏りがないように、(1)母集団の15%が200人未満となる場合は母集団の15%を無作為抽出、(2)母集団の15%が200人以上となる場合は200人を無作為抽出、としておこなった。以上のように分科レベルでの違いを明らかにするために、サンプリング数が8,455名とサンプリング調査としては大規模なものとなり、極めて質の高いデータが得られた。

回答様式は、紙面による郵送の様式とWebベースによる様式の2種類を用意した。また、分野ごとの母集団に対する回答者の一致をとるため、Webベースの様式用にはアクセスIDを付記し、アプリケーションから認証をおこなわせていただいた。

このアンケートに対しては、平成8年11月より平成9年の1月にかけて、回答をいただいた。最終的には4,994人から回答をいただき、回答率は59.1%であった。このうち、Webベースで回答をいただいたのは511名であり、全回答者の10.2%であった。

このアンケートの詳しい分野別、設置者別等の対象者数および集計結果の総括編は1997年6月付けで出されており、学術情報センター研究開発部のホームページ(http://www.rd.nacsis.ac.jp/index-j.html)からも閲覧できるようになっているので、そちらを参照していただきたい(平成9年12月から公開)。

## 3　アンケート調査アプリケーション

### 3.1　WebObjectsの利用

WebObjectsには、データ処理用にObjective-Cを基本とした、オブジェクト指向のWebScript言語が用意されている。また、WebScript言語によるコーディングでは困難なロジックや処理速度が必要な部分には、WebScript言語との双方向データアクセス手段(メソッド)が用意されているコンパイル済のObjective-Cコードが利用出来る(インターフェースさえ

テーブル 1　WebObjectsで利用できる変数の種類とその特徴 [2]

| | 変数を見ることができる場所 | 変数の寿命 |
|---|---|---|
| ローカル変数<br>トランザクション変数 | その変数が宣言されているメソッドの中<br>その変数が宣言されているスクリプトの中 | メソッドが存続している間<br>トランザクションの間。トランザクションは、「リクエストが入ってきてレスポンス(通常はHTMLページ)が出ていくこと」 |
| パーシステント変数(スコープはその変数が宣言されているページの中のみ) | WebScriptでは、その変数が宣言されているスクリプトのなか、Objective-Cでは、その変数が宣言されているクラスの中 | セッションが続いている間 |
| セッション変数 | WebScriptでは、その変数が宣言されているスクリプトの中、Objective-Cではメソッドを通してセッション情報を得る | セッションが続いている間 |
| グローバル変数 | アプリケーション・スクリプトの中 | アプリケーションが続いている間 |

Objective-CであればCやC++も利用可能)。WebObjectsにおいては、テーブル1に示すように、スコープの異なる変数が利用でき、状態の管理およびデータのページ間でのやりとり、変数の有効範囲を細かく制御できる。

これらのWebscriptやObjective-Cコードでは、OPENSTEP環境で利用できる強力なクラスライブラリが使用可能であり、文字処理やデータ検索用コードが簡単に作成できる。さらに、今回はデータ量があまり大きくならなかったので使用していないが、WebObjectsにはオブジェクト指向のDBMSアクセスのクラス及びメッソドが用意されており、Oracle、SybaseやODBCのインターフェイスを持ったDBMSに対してアクセスすることが出来るので、大きなデータセットの検索や既存のデータベースへのアクセスが、オブジェクト指向のクラスライブラリを使ってコーディングできる。また、97年にリリースされたVer.3では、これまでテキストエディタでコーディングしていたWebscript中の入力フォーム等がOPENSTEP環境(NEXTSTEP, OPENSTEP for mach, OPENSTEP Enterprise for NT and SOLARIS等)において、グラフィカルにコーディングできるようになっている。また、モジュールの再利用性も高くユーザグループ間では、ベーシックなモジュールのライブラリが構築されつつある。

### 3.2　アンケート調査アプリケーション

今回のアンケートにおいては、調査項目をできるだけ減らすため、研究者ディレクトリから得られる情報はそちらから得ることとした。このため、アンケート回答者の同定が必要となり、回答者にシリアル番号を付けさせていただいた。また、この番号をエンコードしたものを、Web版回答フォームのアクセスIDとして使用し、認証をおこなった。このアクセスIDの値はテーブル1のグローバル変数とし、複数の回答フォームで共通して使用したので、サーバアプリケーションへの接続時(セッション開始)に1度認証をすれば良いこととなる。また、回答者からのトランザクション(ページの要求、データの送信)が一定時間ない場合は、メモリ節約のため強制的にセッションを閉じるようにした。この場合、回答者がトランザクションの要求をした時点で、再び認証が必要となる。

調査項目を減らしたにもかかわらず、A4紙面で8章39項目12ページと回答項目が非常に多くなったので、Webの投函フォームにおいては関連項目あるいは章によってページを変えることとした(図1)。複数の回答ページに分けることによって、回答者が既にどのページに回答したのかが分からないと、回収漏れが出る可能性が高くなる。これを回避するために目次ページを設け、既に回答したページの目次には「回答済」の表示が併記されるようにした(図2)。また、投函が終わった際には次の回答ページへ、さらに最終ページを投函

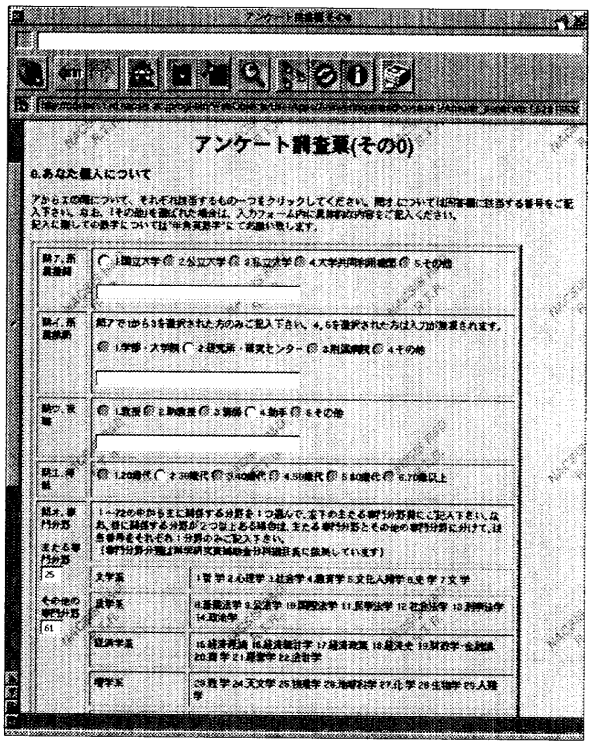



図　1

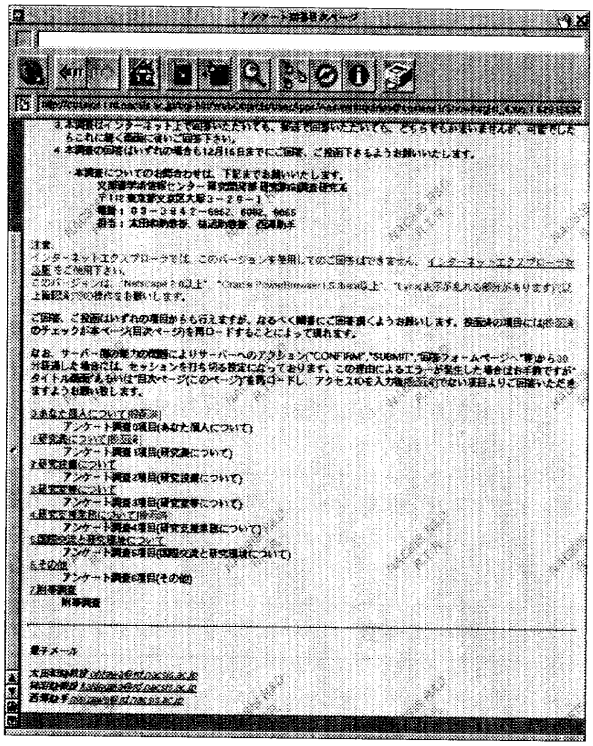


図　2

した際には投函確認ページへのリンクを設けている。

回答の間違いをなくすための工夫もWebScriptを用いると、容易にコーディングできる。スキップすべき質問に回答した場合や、多すぎる選択をした場合などには回答者に対して注意を促すことができ、さらに複雑な処理もObjective-Cを用いることによって、同様にコーディングできる。投函に関しては2ステップ方式とした。まず投函(submit)ボタンをクリックすると回答内容を表示するページが現れ、さらに確認ボタンを押した段階でサーバのファイルに書き込まれる。この段階で既に回答済のアクセスIDであれば、既に回答済である旨のエラーメッセージを出す仕様とした。回答者に対して書き換えを許す設定もDBMSやOPENSTEPの検索と文字処理のクラスライブラリ等を使って容易にコーディングできるが、今回は混乱の防止とセキュリティのため一度回答したページはロックされ書き換えできない設定で調査した(図3)。

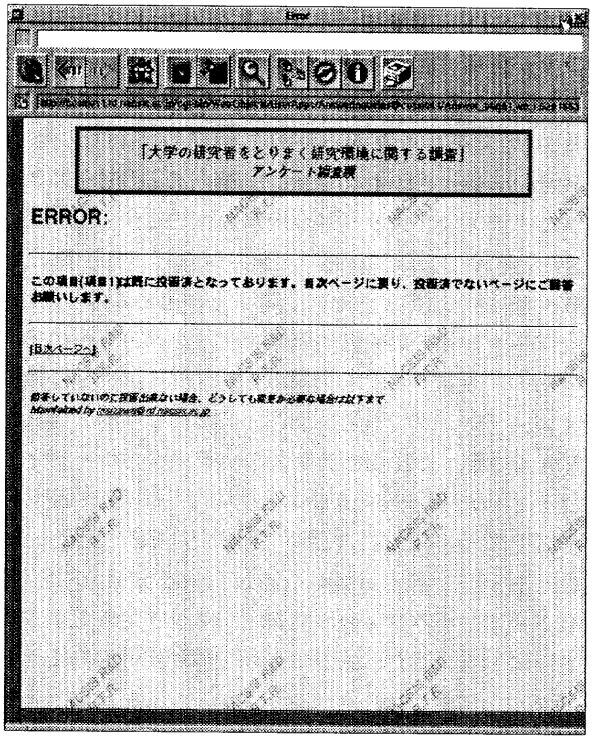



図　3

### 3.3　アンケート調査フォームの問題点と対応

(1) 1バイト文字と2バイト文字

特に数字の問題であるが、今回は開発時間が短かったために、一部(アクセスID等)で半角数字

を要求したので一部の解答者に対して認証時にトラブルが生じた。認証などの判断を伴う入力要求については、英数字の1バイトおよび2バイトの区別はソフトウェアで吸収すべきである。今回使用したWebObjectsでは内部表現にunicodeを使用しており、日本語文字コード判別はWebObjectsが自動的に行いunicodeに変換するので、簡単なプログラミングによって1バイト文字と2バイト文字の問題は回避できるであろう。

(2) 文字化けの問題

WebObjectsでは、クライアントに送信する日本語コードをSJIS, EUC, Unicode等から選択できるが、今回のWebサーバ用コンピュータのOSにはNEXTSTEP3.3J for Machを使用したので、送信する日本語コードを日本語EUCとした。この場合一部のSJISを日本語コードに採用するクライアントで文字化けが生じたようである。ラジオボタンやチェックボックスでは選択肢とユーザ選択の照合に選択肢の文字列全体を使っていた(日本語を含む)ので、一部に文字化けが生じた場合でも、チェックボックスやラジオボタンが上手く機能しなくなってしまった。また、この時点でのMicrosoft InternetExplorerではラジオボタンやチェックボックスのname文字列がEUCであっても選択文字列をSJISで返すという仕様であったので、サーバの文字コードがEUCである場合はラジオボタンやチェックボックスが上手く機能しなかった。

アンケートフォームの開発段階では原因が分かっていなかったので、チェックボックスやラジオボタンの部分を記入式に変更したバージョンを用意して対応した(図4)。そのため、回答様式が二種類になってしまった。しかし、後におこなった別のアンケートでは原因がほぼ特定できたので、ラジオボタンやチェックボックスでの選択肢と比較文字列を分離し、比較文字列にはASCII文字列を使用することとした。このことによりプログラミングは若干複雑となったが、InternetExplorerや少しの文字化けが起こった場合にも機能するシステムを構築できた。また、現時点では日本語コードの不統一、ネットワーク環境の問題等で文字化けを無くすことは難しい。少なくともサーバ側へ送信する文字列はできるだけ2バイトコードは避けるようにすべきであろう。

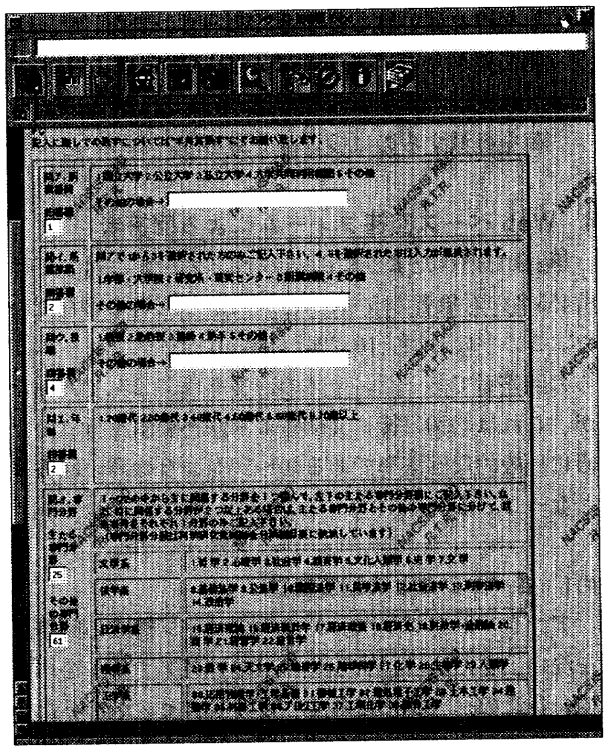


図　4

(3) サーバのパフォーマンス

今回は同時アクセスがそれほど多くなく、大きな問題にはならなかったが、これらのアプリケーションでは接続者ごとにセッションIDが割り振られ、メモリを消費するので同時アクセスが多くなるアプリケーションでは、メモリ消費とタイムアウト時間を短くしなければパフォーマンスが悪くなる。当初タイムアウトは15分に設定したが、回答者からタイムアウトが短すぎ、エラーが起こるとの指摘を受け30分に延長した。60分毎のユニークなホストからの最初のページへの書き込み数では調査票が回答者に届いた直後の月曜日に最大15ホストからの書き込みを記録している。今回は予想したほどのアクセスの集中はなく大きなパフォーマンスの低下はなかった。しかし、ページごとの回答済をチエックするのにすでに記録された回答データ中からIDのサーチをしたため回答者が増えるにしたがってパフォーマンスがかなり低下してきた。大規模なアプリケーションでは、データベースエンジンを使い、アプリケーションサーバ(Webサーバではなく、実際にアプリケーションを実行するマシン)を複数にして、負荷分散

しなければいけなくなるであろう。また、サーバ側で処理しなくてもかまわない部分はJAVAによりクライアント側で処理するのも、サーバの負荷を軽減するのに役立つであろう。

## 4 Webベース回答フォームによる回答者

第2章で述べたように、このアンケートに対しては8,455名の対象者のうち、4,994人から回答をいただいた。この中の511名(10.2%)がWebベースフォームからの回答者であった。全体の回答率は59.1%であったが、これは約20日後に文部省より回答要請を行ったために一般的な回答率である40〜50%よりは高い回答率となっている。

### 4.1 各ページごとの回答者数

アンケートアプリケーションのページ構成の問題点を洗い出すために、Webサーバのログファイルを用いて、各々のページにどれぐらいのアクセスがあったかを調べた。ここでは、各ページの延べアクセス回数ではなく、特に断りがない限りアクセスがあったユニークなホストの数を調べている。また、ローカルなアクセスは除いてあり、ホームページは他のページからリンクされていないので、アドレスを通知した対象者以外のアクセスはほとんどなかったと思われる。

タイトルページへのアクセスがあったユニークなホスト数は614ホストであった。この中で回答フォームのページに進んだホスト数は599である。さらに認証をおこなったのは578ホストであり、正しく認証されたホスト数は572であった。このように認証を設けることによって、Webベースの回答フォームにアクセスした対象者のうち約3.5%のが先に進まなかったことになる。また、認証に失敗したホストもあるが、調査期間中の問い合わせなどから、全角(2バイト)文字でアクセスIDを入力したのがほとんどではないかと思われる。

3章でも述べたように、回答フォームは送信後確認ページがあらわれ、確認後ホストのファイルに書き込まれる2ステップ方式である。8ページある各回答フォームに記入し送信したホスト数および記入内容を確認して送信したホスト数を図5に示す。なお、アプリケーションの設計上の問題から、各回答フォームをロードしたホスト数はログファイルからはわからなかった。この図より、項目0において最初に送信した後、確認の送信を行わなかった例が多いが。これはMicrosoft InternetExplorer(IE)専用に別の回答フォームを用意したにもかかわらず、IEを使用した回答者がIE非対応の回答フォームに記入したため、記入内容と確認ページの内容が一致しなかったことが原因だと考えられる。これは図6に示したようにIE対応回答フォームとIE非対応フォームに分けた場合、IE対応フォームにおいてはこのようなことが起こっていないことから明らかである。また若干ではあるが、一部のブラウザでは文字化けが生じたようである。この場合も回答と確認ページでの内容が変わる場合が出てくる(選択文字列と異なった文字列が送信された場合デフォルトの回答が選択される)。これらの理由により約30名の潜在的Webベースの回答者を失うこととなった。項目1から項目6にかけては、確認の上送信した回答者は1%以内でほぼ一定である。項目0は回答者の属性についての質問であり、これだけに答えた回答者は若干多い。また項目7は付帯調査であり、これを除くと調査項目ごとの回答漏れはほとんどなかったと言える。ここでユニークなホスト数は約500と実際の回答者より少ないが、proxy経由等の回答者等の重複を含ま

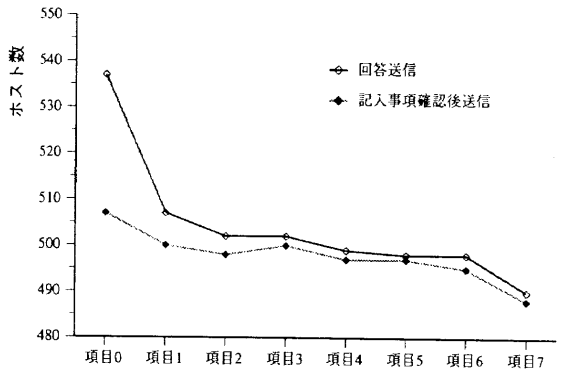



図 5

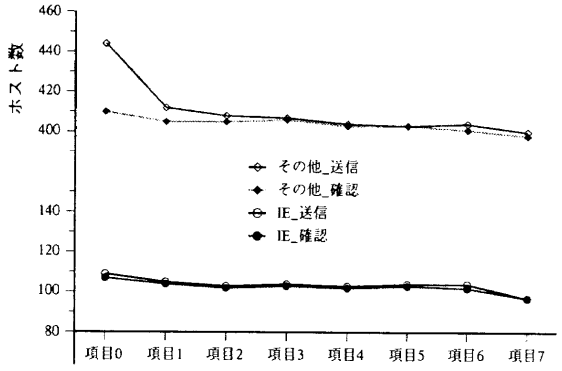



図 6

ないという理由による。

### 4.2 研究分野別 Web ベースの解答率

第2章の調査対象で述べたとおり、調査対象者の研究分野を61分野に分け、(1)母集団の15%が200人未満となる場合は母集団の15%を無作為抽出、(2)母集団の15%が200人以上となる場合は200人を無作為抽出としてサンプリングした。回答率が59.1%であるので、母集団が200人の分野では平均約120名の回答が得られている。この61分野中で回答者が25名以上あった52分野について、回答者中の Web ベースの割合を図7(a:人文・社会科学系, b: 理学・工学, c: 農学・医学, d: 複合領域・広領域)に示す。図中で、分野を( )で囲っているのは回答者が50名未満25名以上で比較的統計上の不確かさを多く含む分野である。

この中で、Web ベースの回答率が高い分野(括弧内は全回答者実数)は天文学26.9%(26名)、情報科学23.5%(115名)、プラズマ理工学21.4%(28名)、応用物理学・工学基礎18.9%(90名)などであった。また、少ないのは体育学2.1%(95名)、教育学2.6%(114名)、経営学3.0%(99名)、哲学3.8%(106名)である。人文社会系では全体に比べ約5ポイントほど低いが、その中では心理学9.6%(104名)や経済学6.9%(87名)など、統計調査等でコンピュータを使う機会が多い分野が高くなっていると思われる。自然科学系、理学においては数学7.5%(106名)を除いては平均より3ポイント以上高くなっている。なお、化学については作業上の都合から調査時期がずれ、Web ベースのフォームが利用できなかったので、省かせていただいたが、情報検索では利用が進んでいる分野なので、高い回答率が期待できたのではないかと思われる。工学の分野では応用物理学・工学基礎、土木工学で利用率が高かったが、平均として理学より0.6ポイント低かった。また、特に電気電子工学での利用率の低さが目立つが、ネットワーク経由でのレスポンスの低さ、フォームの大きさ等でメリットを感じなかった回答者が多かったのではないかと思われる。農学の分野では農芸化学が17.8%(136名)と特に高かったが化学分野との学際領域であり化学情報等で日常的な計算機利用が多いのであろう。また、農学全体では12.1%であった。医学分野は分科ごとの差が大きかった[3]。薬学、歯科の利用率が平均以上であったのを除くと他は10%以下の利用率であった。ま

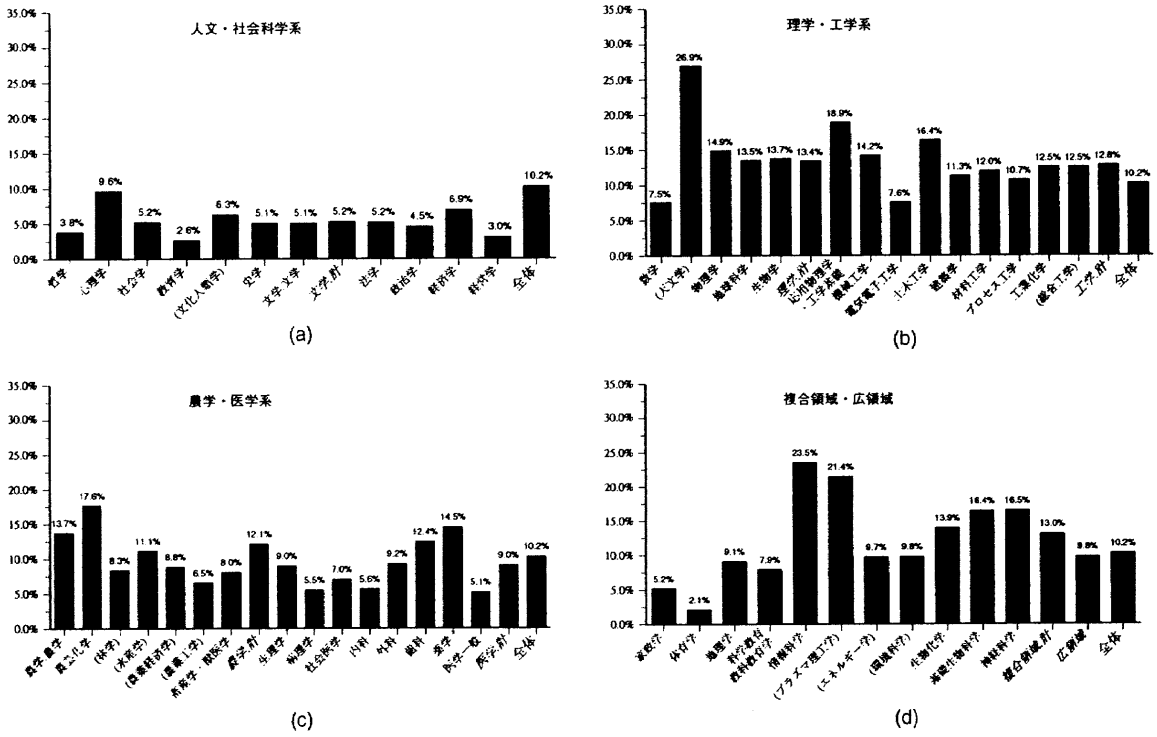



図　7

た、内科と外科の差が大きいのが目立った。複合領域も分科ごと関連が低いのでの差が大きい。50名以上の回答者があった分野では利用率が最も高い情報科学と最も低い体育学が含まれる。また、神経化学16.5%(85名)、基礎生物科学16.4%(128名)等の情報関連分野の利用率の高さが目立っている。その他、科学教育・教科教育学7.9%(114名)と人文社会科学系の教育学の違いも目立っている。広領域についてはほぼ平均に近い9.8%(82名)であった。

## 5　今後の課題

昨今のインターネットの普及からして20%以上のWebベースの回答を期待したが、結果的には約10%であった。少なくとも回答に時間を要する質問項目については、時間、場所を選ばない紙面のフォームが有利となる。特に今回の質問項目には、調査を要するものがかなり含まれており、全ての結果を揃えて回答しなければならないWebベースのフォームには不利な条件であった。Webベースの回答率が上がるためには、回答者のWebフォームへの慣れと、フォームに向かって10分程度で回答・送信できる程度の簡便さ、回答結果が回答者に残る仕掛けも必要であろう。

全般的に自然科学系の方が人文社会科学系よりWebベースの回答率が高く、その中でも平素情報検索等でネットワーク利用の多い分野が高くなる結果であった。しかし予想以上には人文社会系と自然科学系の差が大きくはなく、心理学ではほぼ平均に近い利用率である。これにコンピュータ保有率を考慮するとまた違った結果になったのではないかと思われる。また、フォームの大きさが、理学・工学を中心にかなり利用率低下に影響していると思われる。さらに、一部の大学では調査時期においてのWIDE-SINET間の回線の細さも影響しているはずである。WIDEあるいは商用ネット経由で接続されていた回答者は実質的にWebベースの回答は不可能であったと思われる。ストレスのないネットワークアクセス、回答項目の量の適正化は普及の最低条件であろう。

今回のWebフォームでは8ページの回答フォームを独立にし、大きなアンケートでも1回の拘束時間を短くしようと設計した。これによって懸念されるページごとの回答漏れはフォームの工夫によりほぼ防げたが、回答率が物語るかぎり回答者には大きなメリットがなかったものと思われる。この調査の約2か月後に科学研究費補助金審査員経験者や重点領域研究領域の代表者等に行った10項目の記述式中心のアンケートにおいては、郵送、Webベース、電子メールの3種の回答様式を用意した。この回答率は全体で約40%であり、個々の回答様式の比率は郵送が70%、電子メールが20%、Webが10%であった。この回答様式においては記述式中心であったので、ネットワーク経由に関してはWebより先に普及した電子メールの方が回答者には好まれたようである。また、電子メールに関しては書式が煩雑になるが、好みのメールソフト(インターフェース)が使用できること、タイムアウトの制限がなく、回答結果を保存しやすいなどのメリットがあったと思われる。

ネットワークを通じての回答フォームで回答者にとって大きなメリットが出るのは選択的な1〜2ページ程度の回答フォームあるいは、単項目ていどの記述式の調査であろう。また、回答者が既に個人でデータベース化している発表論文リスト等の内容を記述する研究者ディレクトリ等の調査ではネットワーク経由の回答フォームに大きなメリットが生まれる。電子媒体による調査回答は集計側には大変大きなメリットをもたらす。現在文科系では、ネットワーク接続されたコンピュータの普及の過渡期にあると思われるが、数年後には全分野において定常的な利用に移って行くであろう。また、回答者にメリットが出るフォームを用意することによって、アンケート調査の主流はネットワーク経由に移っていくはずである。今後機会があるごとにより良い調査フォームで双方にメリットがある調査形態にしてゆきたい。

## 6　最後に

WebサーバアプリケーションツールキットであるWebObjectsを利用しての開発は、Objective-Cというマイナーな開発環境なので習熟に時間がかかった点を除いて、おおむね開発環境は良好であった。特にモジュール化が容易なので、コードの再利用性は非常に良好である。ほぼゼロの状態からはじめた今回のアンケート調査用アプリケーションの開発には約2カ月かかったが、その後行った別のアンケート調査用のアプリケーションでは約1週間で運用まで持ち込めた。最近になって数多くのベンダよりWebサーバアプリケーションツールキットが発売されており、DBMSと組み合わせてアンケート調査用アプリケーションのみならずかなり強力なWebベースのグループウェアや教育用アプリケーションの開発ツールとなるであろう。

4.2章で述べたように、人文・社会科学系を中心にはWebベースによる回答率が低い分野があるので、無作為抽出という面からもすべてWebベースにするのは難しいが、コスト削減や統計解析の迅速化、入力ミスの防止等の点で郵送による回答とWebや電子メールでの回答の併用はメリットが大きい。また、Webベースでの回答率を増やすためには、インフラの整備のみならず、ユーザインターフェースの改良や質問項目の量の適正化が重要である。

**謝辞**

貴重な時間を割いて丁寧な回答をして下さった、回答者の皆様に心からの感謝をいたします。また、「大学の研究者をとりまく研究環境に関する調査」の共同研究者であり、Webベースの回答フォームのユーザインターフェースへの助言、テストをして下さった太田和良幸教授(学術情報センター)、柿沼澄男助教授(学術情報センター)、孫媛助手(学術情報センター)、WebObjectsプログラミングやユーザインターフェースについて助言をしていただいた日比野欣也講師(神奈川大学工学部)に感謝いたします。

**参考文献**

[1] 科学研究費補助金報告書(課題番号08300010, 代表:太田和良幸), 平成9年6月
http://www.rd.nacsis.ac.jp/index-j.html

[2] WebObjectsに関するマニュアル:
http://www.next.com/WebObjectsまたは
http://www.nextjapan.co.jp/WebObjects

[3] 大学医局におけるコンピュータネットワークの整備・活用状況に関する検討など
http://www.dna.affrc.go.jp:10082/htdocs/JACS/JJCS/v3/n1/p41_45/index.html

学術情報センター紀要　第10号(1998)

講演

# Rating the Web

Boyd R. COLLINS
Information Technology Librarian
Alexander Library, Rutgers University
ラトガース大学アレクサンダー図書館情報技術図書館員　ボイド　R. コリンズ

**ABSTRACT**

Under the current rapid development of Internet, it is urgent need to establish criterion and methodology for evaluating (rating) the vast Web sites. Such criterion are different from that of the conventional print media: Authority of Web authors; Identification of authors; Interactive and directive confirmation with authors; Organizational identification of Web sites; Correctness, usefulness, comprehensiveness, uniqueness of Content; currency of content, Organization of Web pages (navigability), Analysis of the users' tasks. Rapid innovation of the Web technology as well as the conservative attitudes of users (in this paper, librarian) are two factors to note in evaluationg the Web. The Internet will never be able to achieve its true potential until those who have this understanding step forward to design access applications, to rate websites and start rating services, and provide leadership in Internet information policy.

**要旨**

インターネットの普及にともなって膨大なウェブ・サイトを評価する方法の確立が緊急に必要である。評価基準は、紙媒体の出版物に対する従来からの評価基準とは異なり、ウェブ発信者の権威、身元の明示、表示されたE-mailアドレスを活用した発信者との直接的な確認、ウェブ・サイトやE-mailアドレスの組織種別、内容の正確性、わかりやすさ、普及性、素材の独自性、更新性、Webページの構成と検索能力、利用者側の負担の分析などがある。さらにウェブ評価の際に配慮すべき要因として、急速に発展するWeb技術の革新があり、また利用者（本稿では図書館員）側の保守的な態度がある。インターネットを理解している者がアクセスアプリケーションを設計し、Webサイトを評価し、そのサービスを提供し、インターネット上の情報政策に対してリーダーシップを発揮していくことが必要なのである。

[Keywords]　Internet, Websites, Evaluation Methodology, Criterion
[キーワード]　インターネット、ウェブサイト、評価方法、評価基準

## Introduction

According to Epictetus, impressions must be carefully sifted before one lets oneself be influenced by them, or as he says more eloquently, "Be not swept off your feet by the vividness of the impression, but say, "impression, wait for me a little. Let me see what you are and what you represent. Let me try you."[2] On the web, the same principle applies, particularly in the current atmosphere of Internet hyperbole. Perhaps, the blossoming of extravagant predictions regarding the Internet has not yet hit Japan, but in America, it is in full bloom and constantly reaching new heights. The attitude of most librarians, however, has been far more skeptical. Librarians are on the front lines of patron need and patron anger when the sources we lead them to do not satisfy their need. In an academic library this is particularly critical since professors tend to direct a keen eye at the sources even for undergraduate papers. If librarians can be criticized for not detecting a popular periodical reference, how much more might professors look askance at the average

web citation.

The Internet has been a fertile source of mythical stories with absolutely no basis in fact, but something about the nature of the medium, the combination of the authority of the computer and the sense of community, albeit anonymous community, that it often fosters, for the truth of the rumors that abound. The Internet has no editorial board, yet the fact that we are conditioned to expect editorial controls of anything that is published causes us to take the information which we discover on the Net more seriously than it may actually warrant. When one considers that the Internet is not controlled or regulated by any governmental body, nor any consortium of companies, nor any nonprofit organization, one begins to look at the information available there much differently. A person with average computing familiarity and a continuous connection to the Internet can set up a web site with free software in about fifteen minutes. Therefore virtually anyone can be a web publisher. The bar for participation in the global information market has just been dramatically lowered. The demand for someone to sort through this morass has been just as dramatically raised.

As the Internet evolves, so will the criteria for evaluating all the myriad genres on it. The Infofilter Project attempted to provide a preliminary set of categories for differing types of Web sites, which included government-based pages, corporate brochures, product guides, personal web pages, etc. Since the demise of that project, the torch has passed to information and computer science to create a web taxonomy and attempts such as Ben Shneiderman's have promise. Differing types of web pages will have different sets of quality criteria applied to them. For instance, a financial news site will have a much more exacting measure for "currency" than a site devoted to classical studies. Nevertheless, the following set of criteria are general enough to apply with qualifications to the broad mass of Web-based material.

They evolved from many hours of E-mail discussion on the Infofilter listserv, as well as the criteria which were formulated with the editors of Library Journal during the production of the "WebWatch" column for that journal. In addition, they represent a summation of the many disparate efforts that the library and information science community has made. A good bibliography (or webliography) of these attempts can be found at http://refserver.lib.vt.edu/libinst/critTHINK.HTM.

In many ways, the web represents the ideal medium for developing critical thinking skills. In contrast to the print world, where the barriers to entry alone impose a certain degree of authority in the materials produced, on the web all of the burden of evaluation is on the individual user, especially since the library profession has failed to respond with adequate web-specific evaluative criteria. This being the case, librarians have fallen back on the criteria normally applied to reference material. With some major qualifications, many of these are fully applicable to the web and indeed applying them on the web is more critical than in the print world.

## Web Evaluation Standards

The print world has habituated us to a certain standard of trustworthiness in publications. This is due not only to the high cost of entry into the publishing world but to the fact that in order to compete effectively factual publications need to conform to exacting standards. While no sensible librarian would vouch for the accuracy of all the printed material in his or her library, most of that material which passes for factual has passed through a series of editors, reviewers, and selectors who have each made a positive evaluation of the material. What made the Time Cyberporn fiasco so shocking was that a magazine with as solid a reputation for factual reporting could be easily duped and on such a large scale (Time made it their cover story). Indeed the whole episode stands as an emblem for how different the standards of the print and online world's have become. The print world seems alternatively bemused, jealous, intimidated,

then fearful of the online world, but one lesson everyone seems to have learned from this episode is that the standards of the one world have not been applied in the other.

### Authority of Web Author

The one criteria that all agree is the most critical in evaluating Internet-based materials is the authority of the author or organization responsible for it, assuming this can be determined. This criteria alone demonstrates both how similar and how different the Internet is from previous publication mediums. Similar in that this is precisely the same criteria that determines the value of most printed materials, but different in that it is often much more difficult to determine the source of Internet material. Indeed, my experience of evaluating hundreds of web sites leads me to declare that one large and useful categorization of web sites can be made by separating those whose attribution can be determined from those which cannot.

To illustrate the importance and how central an issue this is on the Net, consider the typical situation in which one searches for a topic using one of the major search engines such as Infoseek or excite. After the search is executed, one usually has a page of "hits" of more or less relevance. Clicking on these "hits" will bring up pages, some of which will be top level home pages, but many will be pages deeply ensconced within the hierarchy of a website with no indication whatever of what website might have spawned the page, much less who the creator of the website was. Even if one is able to find the top page of the hierarchy of which one's "hit" is a member, determining the author is often nearly impossible.

When I first began writing the "WebWatch" column for Library Journal, my editor was so dubious of the value of most websites that she actually went to the extent of applying the same criteria to them as that applied to academic print resources. For instance, in investigating and reviewing a site, I would be required to determine the author's previous publication list in order to determine their qualifications for creating a site such as the one under review. While this may appear extreme for such an open medium as the web, it points in the necessary direction for applying critical standards to the material one finds in the hit lists.

Consider the case in which one finds a page that appears perfectly relevant to one's search, say an evaluation of a new website creation tool by Microsoft. If one were reading a print journal such as MacWeek and found the product evaluated negatively, one could easily compensate for the obvious bias of such a publication. The natural tactic in such a situation would be to consult a parallel publication such as PC Magazine and see if the product received a similar evaluation. If it did, then the case against it would be much strengthened. If not, then the opinion of MacWeek could more definitely be attributed to bias.

So how does one first determine, then evaluate authority on the Web? One must begin by making a firm effort to find the party responsible for the information at the website. This, by the way, is often not the person whose E-mail address is listed on the top-level page as "webmaster." The webmaster is often simply a programmer who was assigned the job of administering the web server and helping with page authoring from a technical viewpoint. The real source of the material could be a secretary in the organization who is providing product information obtained from the marketing department in the case of a company brochure website. In this case the authority of the site would be the trustworthiness of the company which is sponsoring the website. This assumes, of course, that one can affirm that the site in question is the creation of the company it claims to represent. The site might have been put up by a competitor to mislead potential customers, though such fraud would be easily detected on the Web and no cases of this kind have been known so far.

In the case of academic or other research materials, the same criteria that would be applied to the evaluation of a scholarly monograph should

be applied. Any material that has no attribution is useless from the academic viewpoint. No credence whatever can be placed in material whose author has not taken the trouble to identify him or herself. However, even when the author has been identified, no assent can be given until the credentials of the author have been reliably determined. There are two main methods of doing this. One is similar to the librarian method of assessing the reputation of the publisher. For instance, a book published by Harvard University Press is usually judged to be of higher authority than a book published by an unknown paperback publisher. On the web, this is usually equivalent to the authority of the organization that produced the website. The other and the one most often applied on the web is to write to the author via E-mail and decide on the basis of the author's reply what credentials he or she brings to the task. If the author does not provide an E-mail address or does not answer E-mail queries, then the authority of the site is put in grave question. This was the method most often used in the "WebWatch" column for Library Journal.

The determination of authority is an excellent illustration of how a problem raised by the new medium finds its solution in the interactive capabilities of the same medium. The presence of an E-mail address makes it straightforward to engage a website author in dialogue, which in the end can lead to a far more accurate determination of an author's expertise than the mere reputation of a publisher in the print world. E-mail is fast, easy-to-send, and frequently answered, making it a resource of tremendous value in evaluating information.

However, there are occasions when it might be inconvenient to E-mail an author. In these cases, a web search engine can often point one to the author's web page, assuming it has not been linked from the website in question. The personal web page will usually provide such information as institutional affiliation, current position, and degrees obtained, as well as major publications. In many cases, one can not only determine the author's publications but can find online reviews of those publications using a search engine. Newsgroup postings, which are also searchable, can be another revealing source of an author's credibility. Once again, this is a case of the medium producing technical solutions to problems arising from its content.

Another technique that can be helpful in determining authority is to notice where the E-mail address or web page is hosted. For instance, the E-mail addresses of nonprofit organizations end with ".org." If someone with an E-mail address of isoc.org writes about the details of the new Internet addressing scheme, his or her words should be treated with some authority since isoc.org is the Internet Society, the organization that sets Internet standards. Similarly, the domain ".com" signifies a commercial domain which may well have an economic interest in promoting or disparaging a certain product.

## Content

The next and closely related criterion that we must examine is content. Is the material at this site useful, unique, accurate, or is it derivative, repetitious, or doubtful? Clearly, authority plays a major role here, but in determining quality the emphasis in this category is not so much accuracy as utility and relevance. For instance, there are thousands of pages describing how to use HTML to create web pages, but most of the information is derivative and repetitious. The authors would have done better service to the Internet community by simply pointing to one of the high quality tutorials such as "The Beginner's Guide to HTML" rather than afflicting the web with yet another mediocre description of the HTML language.

Whereas authority treats more of the credibility of the information, content is oriented more toward qualities such as comprehensiveness, currency, and uniqueness. Comprehensiveness denotes the ability of a website to cover its topic at the level of stated criteria for inclusion. In other words, if a site claims to be the most complete source for

information about classical Greece, does it live up to this promise? One of the most noticeable and most deplorable characteristics of the Internet as an information-seeking tool is its tendency to fragment information into potentially thousands of partially redundant locations. Sites that bring together Internet information on a particular topic in a comprehensive way are the gold nuggets of the veteran Infonaut. This quality is usually evaluated negatively, by working backward from the site's stated goals and asking, "Where are the glaring omissions? If any." By comparing the creator's intention and the result, one can get a fairly accurate measurement of this quality.

Another quality tied to comprehensiveness, but which is often overlooked is the uniqueness of the material. Does this site offer information that few or none other sites offer? If so, is it likely that this website provider has access to unique sources of intelligence that would allow him or her to certify the accuracy of the material? A prime example of this would be the ambitious American Memory Project, an attempt by the Library of Congress to digitize millions of items important to American history and put them online. Other sites without LC permission cannot legally offer this information. The Library of Congress owns all of the pieces involved which in most cases are unique items that have no copies elsewhere. This quality will probably become more and more important to the Internet as sources continue to proliferate and wise webmasters realize the need for specialization. The best sites are usually those that can capitalize on their unique strengths, which is usually access to special content that they alone are able to offer in a specifically value-added form.

### Currency

Currency is another aspect of authority and content, but needs to be carefully evaluated according to type of website under consideration. In terms of design, the site should have clear and obvious pointers to new content, such as a "What's New" page. The site should also be updated on a regular basis as appropriate to the kind of material being presented.

Dating web pages is an art that can be aided by helpful browsers. For instance, later versions of Netscape allow one to view page information (Page Info on the "View" menu) including the date the page was last modified. This is often critical information in evaluating a web site. Courteous web authors are conscientious about including the date the page was first created, as well as the date last modified. However, if they overlook this courtesy, with modern web tools it is a relatively simple matter to check the date of a page using the "page info" function in Netscape or the page properties in Internet Explorer. If these two methods fail, then an E-mail to the author of the page is often called for, since date information is often critical to the evaluation of a site. At the very least, it indicates how dedicated a web author is to the quality of information at the site. Research has discovered that the frequency of updates is one of the major attractions that cause visitors to return. For sites such as technology news, weekly or in many cases daily updates are essential to maintaining high levels of interaction. In virtually all cases, validity of information is directly related to currency.

An aspect of currency that is often overlooked is the creator's dedication to keeping his or her links up to date. Sites with an abundance of dead links are particularly ungratifying at a time when having automatic link checkers are widely and freely available on the Net. Almost equally crucial to effective web design is the ability to keep up with current technology, not for the sake of technology, but for the sake of information. Authors who create their webs in a rush of enthusiasm and then let them molder untouched for months are extremely common these days, highlighting the virtues of those who are constantly revising and upgrading their pages according to feedback from users. This cycle is even more compelling when it is combined with appropriate use of technology to enhance usability rather than merely to impress.

### Web Page Organization

Organization is the criterion that charts the most new evaluative territory. On the web, this term is often interchangeable with navigability. Is the material easy to get through and to and are the links from one section of the site to another intuitive and well organized. Are there hidden layers in the site that are difficult to discover? Is the arrangement of links uncluttered? Is the information broken down into logical and digestible parts? According to Shneiderman, "Grid layouts and consistent structure help provide familiar landmarks during navigation on the first visit and return visits. Indexes and shortcuts give frequent users paths for rapid traversal."[3] These are some of the key elements of effective design. Though navigation for the web is in the early stages, some basic principles are beginning to emerge.

Searchability is another factor that distinguishes the web from previous information access utilities such as paper and ink. It is primarily an extension of organization in that it refers to the ability of the user to successfully find the information he or she may be seeking. Some of the key elements that contribute to good navigability are:

1) is the information divided into logical units;
2) does the site design display a clear hierarchy from the general to the more specific;
3) has the hierarchy been used to structure the relationship between the more general and more specific information; and
4) does the site structure allow for effective means of expansion.

Well-organized sites are sites that match user's expectations in ways that expand the user's insight. Or as Patrick Lynch has put it, "When confronted with a new and complex information system users begin to build mental models, and then use these models to assess relationships among topics, and to make guesses about where to find things they haven't seen before. The success of your Web site as an organization of information will largely be determined by how well your actual organization system matches your user's expectations. A logical site organization allows users to make successful predictions about where to find things."[5]

### Analysis of the User Task

Critically important for the effective implementation of such design is the analysis of the user task. One recent breakthrough in organizational design has been the development of the Object-Action interface. This design method begins with decomposing a complex information problem into its constituent parts according the elements of a specific task, then correlating the metaphorical instantiation of the site with the tasks analyzed. According to Shneiderman, "For example, a music library might be presented as a set of objects such as collections, which have shelves, and then songs. User may perform actions such as entering a collection, search the index to a shelf, and reading the score for a song. The interface for the music library could have hierarchies of menus or metaphorical graphic objects accompanied by graphic representations of the actions, such as a magnifying glass for a search."[4]

One of the major purposes of performing this task analysis is to create the most effective means for users to navigate a site. Some aids to this process are the navigation bars and site maps are two common means for providing this guidance, though more are beginning to emerge. As a sidenote, the major search engine sites are often the places to watch for innovation on the Web, particularly for navigational innovations.

An aspect of searchability that is often overlooked is a clearly written guide or set of pointers to how to use the search capabilities of a site. Search engines are not appropriate for every site and often they can be a mere dodge by a lazy webmaster, who produces a search engine rather than spend the design time needed to make the site more transparently navigable. The point at which a search engine is necessary is often estimated at about 200 pages. Once a search engine is implemented, it should be fast, return accurate results in relevance-ranked order. It should also provide field-specific search-

ing and allow access to some kind of glossary or thesaurus so as to familiarize users with appropriate terminology used at the site.

Compared to considerations such as the above, graphic design is relatively minor consideration, except to the extent that it can support navigation by the use of the many tools of graphic manipulation. The literature of page layout has been found to be one of the most effective resources for the web. Some of the obvious sins have been described at length in any number of guides to web design. The most prominent of these sins is the use of overlarge graphics, particularly on the first pages, forcing users to wait inordinate amounts of time just to determine whether the site is at all worthwhile for their purposes. On a deeper level, one needs to ask whether the graphics truly enhance the informational content or whether they distract or cover the emptiness of a site's content or lack thereof. Finally, it is necessary to ask whether the graphical aspect is truly needed or whether the material could be more appropriately rendered simply by the use of text.

## Innovation of the Web Technology

One element that may have seemed of minor importance in the early days of the Web, but has since been raised to higher and higher prominance year by year has been the innovative use of web technology. Here I try to emphasize not so much the use of animated graphics (unless they are of more informational than entertainment value), as the enhanced capabilities for interactivity and evencommunity building, for this is the central innovation of the web - the ability to form new kinds of communities. What is most of all to be avoided is the "shovelware" syndrome - the now pervasive tendency in advanced societies to transform what was created in one medium into another medium simply because the new medium has become fashionable. For instance, during the early 90's media companies jumped onto the CD-ROM bandwagon by reproducing their content in digital format. They would even add a few graphics and soundclips and call it multimedia. The fact that most of the text-based content was completely inappropriate to the new medium was totally overlooked. To add insult to injury, these shovelware CD-ROMs often had only the most rudimentary navigation aids. In essence they consisted of digital reproductions of materials that were well-suited to the paper medium where they originated, but which were transformed to the new medium without a thought for what we now refer to in the Web world as "information architecture", the intelligent design of a product's information structure.

On its lowest level, innovative use of web technology involves the avoidance of the sin described above. As we ascend to higher levels of technological design, we begin to ask questions such as how have the creators of this site integrated the technology into its navigational substance so that it enhances rather than distracts from the content presented. For instance, a technology that has potential to enliven sites is known as animated gifs. These are images that simulate motion by showing several different images in rapid succession, similar to flip books used to display cartoon character antics. However, in most cases these animations serve more to irritate and distract users than to provide any enhancement to content. They have therefore earned the endearing sobriquet of "dancing baloney." A whole host of related technologies such as Shockwave, Flash, and most Java usage, seem equally solutions in search of a problem. Lacking basic understanding of information design, the computer programmers who have been hired to design websites have in many cases added flashy technology to sites simply because it was possible to do and it serves to impress management with one's skills.

In regard to this issue, Jakob Nielson, one of the premier writers on web-interface design makes a key distinction in describing the applets that adorn websites. Effective applets can be divided into two broad categories: functionality applets and content applets. Incontrast to technology for its own sake described above, these categories of

applets have precise functions within the web composition. Functionality applets are "Independent mini-applications in their own right with state transitions and multiple views (e.g., a tabbed dialog). Functionality applets often manipulate "real-world" data that exists separately from the webpage (e.g., allowing customers to manage their checking accounts, inventory control, server administration)."[6] On the other hand, content applets are "Tightly integrated with the content of a Web page. Examples include site navigation controls (e.g., an active sitemap, outline flippers to expand and contract a hierarchical listing), active content (e.g., a model of an engine that can be rotated, animated, and otherwise manipulated in place), and minor functions (e.g., a currency converter). Typically, running a content applet has no results other than changing the appearance of the current webpage." Both applets have a direct relationship to the content of the site, and while I have no study at hand, anecdotal evidence suggests that content rather than flashy graphics and in-your-face technology is still the best guarantor of return visits to a site. If not for the wave of hype that now engulfs all things Internet, common sense would have suggested as much some time ago.[7]

Beyond applets, many "traditional" web technologies such as CGI scripts can enhance a site by adding interactive capabilities such as threaded discussion groups - in effect community bulletin boards. Done well, that is with an understanding of the nature of Internet communities, such interactivity is the soul of what makes the Internet different from previous mass mediums and what has fueled its fantastic growth.

## Attitude of Librarian

Librarians too often think of the medium as a kind of oversized online database. Thinking in this way does an injustice to the Internet and to online databases. The Internet is simply a massive network of computers capable of connecting with each other. The fact that the content on some of these computers might be of interest to library patrons was certainly not the major aim of creating the Internet. Online databases were created by commercial entities with specific purposes and audiences in mind. The two are simply incomparable because they are entirely different information structures. In one sense, it's like comparing a phone conversation with telephone lines. One enables the other, but the two are not comparable. In the same way, the Internet enables many technologies, among which are an abundance of online databases, but the Internet itself is not a database in any useful sense of the word. So I have to suppress a chuckle when I see librarians typing search terms into an Internet search engine such as excite and expecting to find useful results. This is clearly a case where experience only misleads.

Though there is an abundance of useful, accurate, and timely information on the Internet, its use as an information tool will, I believe, ultimately be seen as secondary to its primary role. Since the Internet is not and will almost certainly never be a database, it will probably always take trained information specialists to mine the Net. The primary role of the Net and the one that is widely recognized as such now in the advanced parts of the network is as a community. The Internet is, quite simply, the greatest tool for bringing communities of interest together that has ever been invented. Far from the traditional image of computers as giant calculators, the primary purpose for these machines has become communication devices. And this fact does not make them any less useful from the point of view of the librarian community. In fact, librarians have been at forefront of using the medium to exchange information. Community forums, in the form of listservs, chat groups, Usenet newsgroups, and threaded Web-based discussions are at the heart of what makes the Net such a uniquely valuable medium.

So when evaluating innovative use of web technology, my focus tends toward uses that enhance this community-building aspect of the Net. Sites such as firefly are particularly compelling. Firefly describes itself as follows: "Firefly is the leading

provider of products and services for relationship management and advanced personalization." What this means is that a customer's relationship with a company's products can be managed by gathering and leveraging "this valuable profile information to offer highly personalized experiences" and thus "offer personalized recommendations based on an individual's profile." These personalized recommendations flow from an agent-based analysis of a database of user preferences which are compared with the preferences of the user in question. Using these techniques groups of users with similar tastes can be brought together for purposes of relationship building or information sharing, such as similar tastes in movies or music. They are often modeled according to the recommendation, "If you liked Total Recall, you will probably also like True Lies, since 84% of those in your preference category expressed this taste." The applications of this type of webpage to library service should be apparent to those who have struggled with online catalogs.

## In Conclusion

I would like to make a plea for librarians to get involved in the Web and start to make the kind of filtering judgments that only human beings can make. Librarians naturally feel overwhelmed by the sudden rise of the Internet and tend to flounder about with false comparisons between the Net and what they are familiar with, such as online databases. After all, they are supposed to be the informationexperts. Yet they have been largely caught by surprise by the growth and popularity of the Internet, ironic in a profession that prides itself on the promotion of information access. Yet it is time for us to recover from our surprise and begin to make more substantive contributions to the Net than mere lists of URLs. Librarians fill the essential human role of mediating between technology and people, a role never so critical as today. They have a unique ability and one not notable in computer scientists to understand information access from the user's viewpoint, not the engineer's or the designer's. The Internet will never be able to achieve its true potential until those who have this understanding step forward to design access applications, to rate websites and start rating services, and provide leadership in Internet information policy. If librarians fail in this challenge, they will be failing their profession and their users, who are crying out for this guidance. And in the end, if the passive response continues, others will step forward to fill the void.

## Notes and References

[1] All firefly quotes are from the firely webpage at www.firefly.com

[2] Epictetus, Discourses. Shneiderman, p.577.

[3] Shneiderman, Ben. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Reading, Mass.: Addison-Wesley, 1998. P.565.

[4] Patrick Lynch,Yale C/AIM Web Style Guide, 1997, <http://info.med.yale.edu/caim/manual/contents.html>

[5] Ben Shneiderman, Designing the User Interface, Addison-Wesley, 1998, p.567.

[6] Jakob Nielson, "Applet Usability: Stepping Outside the Page", Alertbox, Oct. 15, 1997, <http://www.useit.com/alertbox/9710b.html>

[7] Nielson, <http://www.useit.com/alertbox/9710b.html>

学術情報センター紀要　第10号(1998)

講演

# ウェブの評価

# Rating the Web

ラトガース大学アレクサンダー図書館情報技術図書館員　ボイド　R. コリンズ
Boyd R. Collins
Information Technology Librarian, Alexander Library, Rutgers University
訳　立命館大学総合情報センター情報管理課　石井　奈穂子
Nahoko ISHII
Division of Information Management, University Library, Ritsumeikan University

## 訳序

本稿は文部省科学研究費補助金国際共同研究「海外における日本情報の需要と供給に関する研究」(課題番号07044017)（研究代表者：井上　如）の一環として、1997年12月8日より16日にかけて招へいした米国ニュージャージー州立ラトガース大学アレグザンダー図書館情報技術及びレファレンス図書館員であるMr. Boyd R. Collinsの講演"Rating the Web"の日本語訳である。

コリンズ氏は米国図書館協会（ALA）の機関誌のひとつであるLibrary JournalにWebWatchと題して連載してきたコラムに対して、1996年度カーナーズ優秀編集賞（ベスト・レギュラー・コラム）を受賞しており、優秀なウェブ・サイトを発見し評価するために図書館員で構成する全国的な共同チームであるインフォフィルター（Infofilter）プロジェクトの創設メンバーでもあり、米国図書館協会のネットワーク情報資源評価委員会の創設メンバーである。

本稿は1997年12月12日午後に学術情報センター別館会議室で開催された報告会で"Internet Flotsam and Jetsam"と題して英語で講演されたもとになった論文であり、インターネットのホームページを評価することの意義、方法、基準などについて検討し、利用者側における評価のための手がかりと、発信者側が配慮すべき事項について述べたものである。

国際共同研究「海外における日本情報の需要と供給に関する研究」では、日本情報発信をめぐる問題点を検討しているが、本稿は米国における最近の問題点を明らかにしつつも、日本情報発信にとって共通する課題を提起しているものと位置づけている（内藤衛亮）。

## はじめに

エピクテトス(訳注：ギリシアのストア派の哲学者)が言うには、印象というものは、人がそれらから影響を受ける前に選別されなければならない。また、彼がさらに雄弁に語るには「印象の鮮やかさに惑わされるな。印象だって？ちょっと待ってくれ。君が何者なのか、何を表現するのかを見せてくれ。まず君を試させてくれ。」(2)ということだ。Web上においても、同じ原理が、特に近年の過剰なインターネット市場において当てはまる。おそらく、インターネットに関する途方もない量の予測が実現することは、まだ日本では起こっていないだろう。しかし、アメリカでは、インターネットは今や満開であり、さらに高い到達点に至ろうとしている。しかしながら、ほとんどの図書館員は、インターネットに関してはるかに懐疑的である。図書館員は、利用者のニーズに直面し、彼らの欲求を満たさない資料を提示した際には叱責を受けることもあるのだ。学術図書館においては、教員はたとえ学部生の論文であっても、その出所には厳しい目を向けているため、特に重要である。もし図書館員がポピュラーな参考文献を見つけることができずに非難されうるのならば、教授たちはWebページの普通の引用をどれだけ疑いの目で見ていることになるだろうか。

インターネットは、まったく事実に基づかない創造的物語の源である。しかし、このメディアの持つ本質を少し述べるならば、コンピュータの権威と社会観念（匿名の社会であり、時としてうわさが氾濫するという事実を助長する社会ではあるが）との組み合わせを持っていることである。われわれが、出版されたものならどんなものでも、編集者の規制を期待してしまうことは、われわれがネット上で発見する情報を実物以

上に深刻に受け止めてしまうことにつながる。それにもかかわらず、インターネット上には、編集用掲示板が存在しない。インターネットは、いかなる政府、企業の共同体、および非営利団体からも、制御・規制されていないとみなす場合、人はその情報を違った角度から検証し始める。コンピュータに関する平均的な知識と、インターネットへ常時接続が可能な者は、ものの15分程で、無料のソフトを用いることによって、Webサイトを立ち上げることができるのである。そのため、実際に誰もがWebパブリッシャーになることが可能なのである。グローバルな情報市場へ参加するハードルは劇的に低くなった。情報を発信したいと考える人も劇的に増えた。

インターネットが発展するのに伴い、無数のジャンルを評価していく基準も発達するであろう。Infofilterプロジェクトは、政府関係のWebページから、企業のパンフレット、商品ガイド、個人的なWebページなどなどを含む異なるタイプのWebサイトに至るまで、カテゴリー分けを提供しようと試みた。このプロジェクトが活動を停止したことによって、情報科学分野の人間に、Web分類をしていく業務やBen Shneidermanが確約したような試みを引き渡すことになった。異なったタイプのWebページには異なった評価基準が必要となる。例えば、金融のニュースサイトには、古典文学研究のページよりも、「通貨」に関する正確な尺度が必要となる。しかしながら、以下に述べる基準は、Webベースの資料の品質を吟味するのに十分一般的である。

それらの基準は、Library Journal誌の編集者が、"Web Watch"コラムを作成する際に評価基準を組織立てたのと同様、Infofilterメーリングリスト上でE-mailによって長い時間をかけた議論の末に、発展してきた。さらにそれらは、図書館と情報科学コミュニティとが作り上げた多種の努力を集積したものでもあるのだ。このような試みによって評価された適切な書誌（あるいはWebの書誌）は

http://refserver.lib.vt.edu/libinst/crtTHINK.HTM

に見ることができる。

多くの場合、Webは批評能力を育てる理想的なメディアである。出版物に対してある程度の権威をもっていることを課している印刷物の世界とは対照的に、Webサイトを評価していく一切の義務は、利用者に負わされることとなった。特に、図書館員が適切なWeb特有の評価基準を提供することに失敗したためである。そのため、図書館員は、通常、参考資料に適用される基準に頼ることになった。いくつかの主要な制限とともに、それらの多くはWebに適用することが完全に可能であり、Web上でそれらを適用することは印刷物の世界で以上に重要である。

## ウェブの評価基準

印刷物の世界では、われわれは出版物における信頼性の基準に慣れてしまっている。これは印刷物の世界へ足を踏み入れることが高くつくということだけによるのではなく、他の出版物と有効に競争するためには、厳しい標準に一致する必要があるという事実にもよる。思慮深い図書館員は、彼ら自身の図書館の蔵書がすべて的確であると保証することはないとしても、その蔵書のほとんどは編集者、書評家そして選書係による肯定的な評価を受けているのである。タイム誌のサイバースペース上のポルノ事件があれほどショッキングだったのは、あれほど確かなレポートで堅実な世評を得ていた誌が、いとも簡単にあれだけの規模でだまされたことにある（タイム誌はかの記事をカバーストーリーとして掲載した）。この一連のエピソードは、印刷物世界の標準と、オンライン世界の基準とがこんなにも異なっているということの象徴として位置づけられる。印刷物の世界は、どちらかというとオンライン世界に対して心を奪われ、嫉妬し、威圧され、そして恐れを抱いているかのように思える。しかし誰もがこのエピソードから学んだことは、一つの世界の基準は別の世界では適用されないということであった。

## ウェブ発信者の権威

インターネットベースの資料を評価していく中で、だれもがもっとも重要だと認識する基準の一つに、そのサイトの作者や組織の責任者の権威がある。この基準は、インターネットのどこが以前の出版メディアと類似していて、どこが異なっているかを実証している。類似点としては多くの出版物の価値を決定づけている基準であるという点である。相違点は、インターネット資料の出所を決定づけるのが、しばしば困難だということである。確かに、何百というWebサイトを評価してきた私の経験から言うと、一つの巨大で有益なWebのカテゴリーは、その属性を決定しにくいページを無視することで、作成することが可能なのである。

この問題がネット上でどれほど重要で、どれほど中心的な問題かを示すために、Infoseekやexciteと

いった主要な検索エンジンを使って、あるトピックについて検索するという典型的な例を考えてみよう。検索を実行した後に、たいてい妥当かどうかは別にして、「ヒットしました」との結果を得る。この"hits"というボタンをクリックして当該ページに飛んでいくと、そのうちのいくつかは高いレベルのホームページであろう、しかしながら、その他の多くは、そのWebサイトに何があるのか何の表示もしていないWebサイトや、作者不明のWebサイトであったりするのである。万が一、この階層構造の一番上のページにたどり着くことができたとしても、その作者を決定することはほとんど不可能に近い。

私がLibrary Journalに"Web Watch"の連載を開始した当初、編集者は、多くのWebサイトの価値に関して大きな疑問をいだいていた。そのため彼女は、学術書に引用されたのと同じ基準を適用して評価しようとした。例えば、Webサイトを調査し評価するために、私はそのサイトの作者の過去の出版物のリストを提示するよう求められた。サイトを作成するのに必要な能力を決定つけるためにである。Webのようなオープンなメディアでは、これは極端に思えるかもしれない。一方、ヒットリストにあがってくるような資料に対して重要な基準を適用することは必要である。

検索した結果、完璧に適合するページを見つけることができた例を考えてみよう。それがマイクロソフト社製の新しいWebページ作成ツールの評価であったとする。もし例えばMacWeekのような印刷されたジャーナルを読んでいて、その製品が否定的に評価されていたならば、人はすぐにそのような出版物が明らかに偏見に満ちていることに気づくであろう。そのような状況を自然に調査するならば、PC Magazineのような類似した雑誌を調べてみて、同じような評価を受けているかどうかを確かめるとよい。もしそうだったならば、その商品に関する反論はがぜん強くなる。同じような評価を下していなければ、MacWeekの見解は偏向がかかっていることになる。

Web上の権威をどのようにして決定付け、そして評価していけばよいのだろうか？まず最初に、堅実に努力してWebサイトの情報に対して責任を負っている人物を見つけることから始めるべきである。ところで、この人物はしばしばトップページに"Webmaster"としてE-mailアドレスの記載してある人ではないことがある。Webmasterはしばしばサーバを管理し、技術的な見解からWebページをオーソライズする単なるプログラマーであることが多い。企業のWebサイトの場合、実際の情報源は、マーケテイング部から得られる製品情報を提供している、その会社の秘書であったりすることもある。この場合、このサイトの権威は、Webサイトを運用している会社の信頼性となるであろう。もちろん、このことは、問題のサイトがその会社が作り出したものであると仮定した場合である。そのサイトは、消費者を潜在的に誤解させるものとして競争相手によって糾弾されるかもしれない。そのような不正手段は、そう知られている訳ではないのだが、Web上で簡単に発見される。

学術的資料やその他の研究資料に対しては、学術研究論文の評価に適用されるのと同じ基準が、適用されるべきである。帰属性を持たない資料は、学術的な見解からは何ら役にはたたない。著者自身の身元を明らかにしていない論文は、信頼がおけないのである。しかし、著者が認識された場合でも、著者の信頼性が実証されるまでは、賛同することはできない。それを認識するためには、二通りの方法がある。一つは、図書館員が出版社を評価する際に使用する方法である。例えば、ハーバード大学出版局で出版された図書は、通常、未知のペーパーバック出版社から出版されたものよりも、高い権威を持つと判断される。Web上では、Webサイトを提供している組織の権威に依拠する。もう一方の、そしてWeb上でもっともよく使用される方法に、著者に直接E-mailを出し、著者が自分の業務にどのような証明をもっているかを見た上で、判断するというものがある。もし著者がE-mailアドレスを公表していなかったり、あるいはE-mailでの質問に返答しなかったような場合には、そのサイトの権威には大きな問題があるということになる。この方法は、Library Journalに連載されている"WebWatch"で使用している方法である。

権威を決定することは、新しいメディアによって生じてきた問題を、この対話可能な同一メディアでどのように解決するかということの優れた一つの例である。E-mailアドレスの存在は、利用者からWebサイトの作者との対話の線をまっすぐに結びつける。これは最終的には印刷界における単なる出版社の評価以上に、正確に著者の専門的知識を決定づける。E-mailは反応が早く、送信も簡単で、すばやく返答することができる。そのため、情報を評価する場合に計り知れない価値をもった資源であるといえる。

しかしながら、著者にE-mailを送信することが不

便な場合がある。そういった場合、問題のWebサイトからリンクがはられていないと仮定した上で、検索エンジンが著者のWebページを指示していることがある。個人のWebページは、通常主な出版物で得られるのと同等の情報、例えば、所属団体、現在の地位、そして取得学位などを提供している。多くの場合、検索エンジンを用いて、著者の出版物についての評価を見た上で、その妥当性を決定することが可能になる。ニュースグループへの投稿もまた、検索可能であるが、それもまた著者の妥当性を計る資料となりうる。繰り返しになるが、これはその内容から生じてくる問題の技術的な解決法を示すことのできるメディアの場合なのである。

権威を決定づけるのに役立つその他の手段として、E-mailアドレスやWebサイトがどこの所属のものなのかを知ることが挙げられる。例えば、非営利団体組織のE-mailアドレスは、最後が".org"で終わる。もし、"isoc.org"のE-mailアドレスをもつ誰かが、インターネットアドレスの割り当てに関する計画について詳細に書いたとするならば、その発言は、ある種の権威を持つものとして取り扱われるに違いない。なぜならば、"isoc.org"て終わるアドレスは、インターネットの基準を決定しているInternet Societyのものだからである。同様に、".com"のドメインは、商用のサイトであることを意味し、ある商品の販売を促進しようとしたり、誹謗しようとする営利目的を持っているかもしれない。

権威を決定づけるのに役立つその他の手段として、E-mailアドレスやWebサイトがどこの所属のものなのかを知ることが挙げられる。例えば、非営利団体組織のE-mailアドレスは、最後が".org"で終わる。もし、"isoc.org"のE-mailアドレスをもつ誰かが、インターネットアドレスの割り当てに関する計画について詳細に書いたとするならば、その発言は、ある種の権威を持つものとして取り扱われるに違いない。なぜならば、"isoc.org"て終わるアドレスは、インターネットの基準を決定しているInternet Societyのものだからである。同様に、".com"のドメインは、商用のサイトであることを意味し、ある商品の販売を促進しようとしたり、誹謗しようとする営利目的を持っているかもしれない。

## 内容

さて、次に、すぐにわれわれが調査しなければならないことは、Webの内容である。そのサイトの資料は有効でユニークで正確か？それとも二次的で、くどく、信用のおけないものなのか？はっきりいって、権威はここで重要な役割を果たすのである。しかし、その品質を吟味する際に、カテゴリー分けは、ユーティリティや関連付けほど重要ではない。例えば、Webページを作成するためのHTML文法の使用法に関して何百ものページがあるが、そのうちの大部分は二次的でくどいものである。著者は、インターネットのよりよいサービスのためには、「初心者のためのHTMLガイド」といった高品質のページを指摘するべきである。HTML言語に関するあまり質のよくないページでWeb社会を悩ませることはないのである。

権威は、その情報の信憑性を問いかけるが、Webの内容は、わかりやすさ、普及性、独自性といった特質を問うものである。理解しやすさは、ある一定のレベルの基準でもってそのトピックをカバーするWebサイトの能力を示している。言い換えるならば、もしあるサイトがギリシャ古典に関する情報についてはもっとも完全なソースであると宣言するならば、どうやってそれを証明するのか？情報検索ツールとしてのインターネットのもっとも顕著でありもっとも嘆かわしい特性の一つに、情報を何千という部分に分けてしまうことがある。ある特定のトピックに関するインターネット情報を分かりやすい方法で説明してくれるサイトは、ベテランのInfonaut（情報航海士）にとって宝の山である。Webの品質は、よく否定的に評価される。サイトからサイトへと後ろ向きの姿勢でみてまわり、次のように問いかけるのだ「手抜かりはどこにあるのだ？もしあるとしたらだが」。制作者の意図と結果を比較することによって、品質の正確な尺度といったものを得ることができる。

わかりやすさと密接にかかわっているが、しばしば見落としてしまいがちな点に、素材の独自性があげられる。このサイトは、他のサイトにない情報を提供しているだろうか？もしそうなら、そのWebサイトは、素材の正確さを証明する独自の知的情報源へのアクセスを提供しているだろうか？そのもっとも良い例が、アメリカン・メモリー・プロジェクト計画である。これは、米国議会図書館がアメリカの歴史にとって重要な項目について電子化し、オンラインにのせていこうという試みである。議会図書館の許諾なしに、他のサイトがこの計画についての情報を提供することは法律上不可能である。米国議会図書館は、必要となるすべ

ての資料について自館で所有しており、それらの多くは他では手に入らない貴重なものである。情報源が激増し続け、賢明なWebmasterたちが専門化の必要性を認識しだしたように、品質の高さは、インターネット上でますます重要になってきている。最良のWebサイトというのは、独自性を発揮しているサイトであり、かつそのサイトだけが提供することができる付加価値の付いた内容へとアクセスできるものである。

**更新性**

権威や内容の別の側面であるが、情報が更新されているかもまた、Webサイトの種類によって注意深く評価することが必要である。デザイン的なことを言えば、Webサイトは新しい内容に関して、例えば「新着情報」ページのような分かりやすく明白なポインターをもっているべきである、サイトは、現状にあわせた形に、適切に定期的に更新されるべきである。

Webページの日付けを示すことは、有益なブラウザーの助けによってなされるものである。例えば、ネットスケープの最新版では、そのページがいつ更新されたかの情報も含んでいるページ情報（"view"メニューの中のPage Info参照）をみることが可能となった。Webサイトを評価する上で、そのページがいつ作成されたのかも重要な情報となる。丁寧なWebページの製作者は、誠実にも更新日のみならず制作日を入れるだろう。しかし、彼らが親切を見過ごしても（日付を入れ忘れても）、現代のWebツールは、簡単にそのページの作成日付をチェックすることができるのである。ネットスケープの場合は"page info"機能であるし、インターネットエクスプローラの場合は"page property"から見ることができる。もしその二つの方法でうまくいかなかったら、作者にE-mailで問い合わせをすればいいのである、なぜならば日付の情報は、しばしばサイトの評価にとって重要であるからである。少なくとも、この件は、作者がどれだけWebページの品質保持を手がけているかを明らかにするのである。絶えず更新することによって、利用者は再度そのページへと立ち寄るということが、調査からも立証された。技術分野ニュースのサイトは、その高いレベルを保持するには、1週間に1度、あるいは日々の更新が必須である。多くの場合、その情報が有効か否かは、更新性と密接にかかわりがある。

しばしば見落とされる更新性の一面として、献身的に作者がそのページのリンクを最新にしておくことが挙げられる。多くのデッドリンク先を持つサイトは、自動的にリンク先をチェックしてくれるプログラムがネット上で入手可能である時代においては不愉快である。同様に効果的なWebの設計にとって重要なのは、現在の技術革新についていくことの出来る能力である。それは技術のためでなく情報のためにである。急いで熱中してWebを立ち上げた作者は、何か月も朽ちるにまかせていることがあるが、今日ではこういったことが急激に増えている。利用者からのフィードバックによってそのページを絶えず改訂したり、更新したりする人々を目立たせる結果となっている。この循環は技術の適切な使用法と結びついて、単に印象づけるだけでなく、利便性を高めるような場合においては賞賛的でさえある。

**Webページの構成**

Webページの構成は、もっとも新しい評価しうる分野である。Web上では、この言葉はしばしば「ナビゲーション機能」と言い換えられる。その素材は取り出しやすいか、うまくリンク付けや構成立てが出来ているかどうか。サイトの中に、隠れた層があり、行きつくのに手間がかかるか。リンク構成がバラバラになっていないか。情報を論理的な部分と要約した部分とに分けることはできるか。Shneidermanによると「グリッドのレイアウト及び一貫した構造は、様々なページを行ったり来たりする中で目印を提供してくれている。インデックスやショートカットは、よくそのページを訪れる利用者に、迅速な縦断を可能にする未知を提供してくれる」(3)。これらは、ページの有効な設計において重要な要素である。Webのナビゲーション機能は未だ初期段階にあるが、いくつかの基礎的な法則は固まりつつある。

検索能力は、紙やインク媒体といった従来の情報と、Webとを区別するもう一つの要因である。これは、捜している情報を首尾よく見つけることができるかに依っているという点で、基本的には構造の拡張であると言える。いくつか、よいナビゲーターの重要要素を述べよう。

1) その情報は、論理的（ロジカル）な単位に分割できるか、
2) そのサイトは、一般的な内容からさらに特定の分野まで、明白なピラミッド型構成にデザインされているか、
3) その階層構造は、より一般的な内容とより特

定分野の内容との関係を構築しているか、

4）その構造は次への展開を効果的にしているか。

うまく組織立てされたサイトは、利用者の知識を増やす方法で、その目的をかなえることができる。あるいは、Patrick Lynch が次のように述べている。「新しく複雑な情報システムに直面した時、利用者はあるメンタルモデルをつくる。そして、トピックの関係を評価する時、かつてたどりついたことのないホームページに関して推測するのにそのモデルを使用する。あなたの Web サイトが情報組織として成功しているかどうかは、実際のシステム構造が利用者の要求に沿っているかで決定できる。論理的に構造化されたサイトとは、利用者にとってどこに何があるのかがはっきりわかるサイトである」[4]。

## 利用者側の負担の分析

そのような設計を効果的に実行するのに重要なものは、利用者側の負担の分析である。構造の設計における最近の革新的な出来事の一つに Object-Action インターフェースの発展がある。この設計方式は、複雑な情報を、特定の作業要因に沿った形の構成要素に分解し、サイトの比喩的な具体例と分析された作業との相互関係を示すことから始められる。Shneiderman によれば「例えば、音楽図書館であれば、収集対象は、所蔵物そして歌であろう。利用者はそのコレクションにまず入り、インデックスを検索し、そして、歌のスコアを読みにいく、という一連の行動を起こすであろう。音楽図書館のインターフェースとして、メニューの階層構造と、検索のための拡大ツールのような、アクションをグラフィカルに表現するものに伴われた比喩的なグラフィック・オブジェクトを有するだろう。」[5]

この業務分析を行う主要な目的の一つは、サイトをナビゲートする有効な手段をつくりだすことである。この過程にとっては、ナビゲーションバーやサイトマップが役立つし、これらはこのガイダンスを提供する上での共通の主要な手段となる。付け加えるならば、主な検索エンジンのサイトは、Web 上の新機能、特にナビゲーションの新機能が配置されている。

しばしば見落とされるのだが、検索能力についての一つの基準は、明確に書かれたガイドやサイトの検索能力をどのように使うか示す一連の指針であったりする。検索エンジンはすべてのサイトに適切なのではなく、怠惰な Webmaster によって単なるごまかしとなったりもするのだ。彼らはサイトをナビゲートしやすくするのに時間を費やすよりも、むしろ検索エンジンを作ったりしてしまうのである。検索エンジンが必要になるのは、およそ200ページを越えるサイトである。検索エンジンを備えるのならば、反応は速くなるべきで、適切な順序立てとともに正確な結果がかえってくるべきである。検索エンジンは特定の分野に関する検索をも提供すべきであり、また利用者がネット上で使用される専門用語に習熟できるよう、用語集やシソーラスへもアクセスできるようにするべきである。

上記のような考え方と比較して、多くのグラフィック操作ツールの使用によって、ナビゲーションを支援することができるという部分を除いて、グラフィック・デザインはあまり考慮すべきことではない。ページレイアウトに関する文献は、Web のためのもっとも有効な資源のうちの一つであると分かった。明白な違反のうちのいくつかは、どの Web 設計ガイドにも詳細に記述されている。これらの違反の中でもっとも顕著なものは、トップページに大部なグラフィックをのせることである。これは、そのサイトが利用者にとって有効なのかどうか判断するためだけに、利用者に過度の時間の負担を強いている。さらには、人はグラフィックスが本当に情報内容を高めているのかどうか、ただ混乱するだけなのか、サイトの内容の空虚さを隠しているのか、内容自体が欠如しているのかを見極めなければならない。最終的には、グラフィックが本当に必要なのか、資料が、シンプルにテキストだけで適切に表現することができるのかどうかを判断することが必要である。

## Web 技術の革新

Web の初期段階ではあまり重要視されていなかったが、年々その重要度が増してきた要素の一つに、Web 技術の革新がある。アニメーション画像の多用について強調しようとしているのではなく（娯楽的な要素より情報的価値が高いのでなければ）、双方向性能力を高めることや、さらにはコミュニティの設立を強調しようとしているのである。新しいタイプのコミュニティをつくる能力は、Web の中心的革新性なのであるから。だれもが避けるべきものに、“shovelware”症候群があげられる。高度化した社会では、あるメディアでつくられたものが、新しいメディアがはやっているからという理由だけで移行するのである。例えば、90年代初頭に、メディア会社は、資料の内容を電子化し、CD-ROM の流行に飛び乗ったのである。彼らはいく

つかの画像や、音声を加え、マルチメディアと称した。さらにひどいことに、これらのshovelwareCD-ROMはしばしば粗末なナビゲーション機能しか持ち合わせていなかった。要するに、それらは元々は紙のメディアで作成されたものであり、それに適した資料を電子化したもので成り立っている。しかし、われわれが今Web社会に望んでいる「情報構造」(情報構造の知的な設計）に言及することなしに、新しいメディアに変形したのである。

初期レベルにおいて、Web技術を革新的に使うということは、上述の違反を避けることを意味する。われわれは、技術設計のより高いレベルに至るにつれ、そのページの作者がどのように技術をナビゲーション機能に結びつけていってくれるのかを気にし始める。そのことによって、コンテンツを台無しにせずに良いものにしているのである。例えば、サイトを活気付ける可能性のある技術に、アニメーションgifがある。これは、迅速にいくつかの絵を見せることによって、動きがあるようにみせる技術である。漫画のキャラクターのおどけたしぐさを表示するのと同様の技術である。しかしながら、ほとんどの場合、こういった動画は利用者にとって内容をよりよいものにするというよりも、利用者をいらいらさせ、混乱させる。そのため、それらは「踊るばからしいこと」というあだ名を得た。Shockwave、Flash、Java使用法といった、関連技術のホスト全体は、そのような問題への答のように思える。情報設計についての基礎的な理解を欠いているため、Webサイトを設計するために雇用されたプログラマーが、彼ら自身に能力があるということや能力を認識させる機会であるという理由によって、多くの場合、技術を誇示することになってしまうのである。

この論点に関して、Webインターフェース設計の第一人者である、Jakob NielsonはWebサイトを区別する際、Webサイトを飾るアプレットを重要なキーとしている。有効なアプレットは二つの大きなカテゴリーに区分することができる。機能的なアプレットと、内容のアプレットである。上で述べたような技術とは対照的に、これらのアプレットのカテゴリーは、Web構成の明確な要素を持っている。機能的なアプレットとは、「状態の遷移と多様な見方について自身の権限を持っているミニアプリケーションである（例：タブ付けされた対話)」[6]。機能的なアプレットは、しばしばWebページからは分離して存在している「実社会」のデータを操る（例えば顧客が銀行預金確認や預金残高管理ができること、サーバ管理等)。他方、内容のアプレットは「Webページの内容と密接にかかわっている。例えば、サイトナビゲーション機能(例：アクティブなサイトマップ、階層的なリストを拡張・対照するアウトラインヴューア）や、アクティブな内容（例えば回転され動かされ、適当に操作されるエンジンのモデル)、そして小さな機能をもつもの(例：通貨変換プログラム）である。典型的に内容的アプレットの実行は、現在のWebページの外観を変更する以外に何の結果ももたらさない」。どちらのアプレットもWebサイトの内容に直接関わりを持っている。私が研究したわけでないのだが、未発表の調査結果では、利用者がリピーターとして何度もそのサイトを訪れるのは、派手なグラフィックスやこれでもかとの技術ではなく、内容が充実しているからである。そうでなければ、現状のインターネットをすべて飲み込む誇大広告の波について、それ以前に常識で予想できたであろう[7]。

アプレットの他にCGIスクリプトのような多くの「伝統的な」Web技術がある。それはディスカッショングループ（効果的な電子掲示板コミュニティー）に会話型の能力を付け加えることによって、Webの質を高めることができる。インターネットコミュニティの原理を理解した上でうまく実行されたなら、そのような双方向性はインターネットが過去のマスメディアとは異なっており、そのすばらしい成長を育てる源となる。

## 図書館員の態度

図書館員は、メディアを特大のオンラインデータベースの一種であると考えがちである。このような考え方は、インターネットおよびオンラインデータベース両方に対して正しい判断をしていない。インターネットは、簡単に言えば、互いに接続することができる大規模なコンピュータネットワークであるといえる。こういったコンピュータの中身が、図書館利用者にとって興味深いかもしれないという事実は、確かにインターネットを創造していく主な目的ではなかった。オンラインデータベースは、特定の目的と対象者を視野に入れ、企業によって作成された。これらは、まったく異なる情報構造を持っているので、単純には比較できない。私は、図書館員が、exciteのような検索エンジンに、検索したい言葉を入力して、よい結果が帰ってくるのを待っているのをみるにつれ、くすくす笑いをおさえなくてはならないのである。これは、明らかに過去の経験が誤解させている例である。

インターネット上には有益、正確で、タイムリーな情報が豊富であるが、その情報ツールとしての機能は、私が思うに、二次的な役割として認識されるだろう。インターネットは決してデータベースとはなり得ないし、だからこそネットから情報を掘り起こす情報専門家が必要となるのである。初期のネット、およびネットワークの高度な部分で広く認識されているものは、ネット上のコミュニティである。インターネットは、簡単に言えば、関心を持つものが集まってコミュニティを造っていくのに最適なツールである。伝統的な大型計算機としてのコンピュータのイメージとは違い、これらマシンの当初の目的はコミュニケーション装置となることであった。そしてこの事実は、図書館コミュニティの視点からみて、使いづらくなるという訳ではない。実際図書館員は情報を交換するために、メディアを使用する最前線にいる。メーリングリスト形式のコミュニティフォーラムや、チャットグループ、ネット上のニュースグループ、そしてWebベースの議論はネットをユニークで価値あるメディアにしていく核心なのである。

したがって、Web技術の革新的な使用方法についての評価を行う時、私の焦点はネット上にこのようなコミュニティを形成する要素を高める方法に向きがちである。「Firefly（蛍）」のようなサイトは人を動かさずにはおかない。Fireflyは、自身のことを次のように述べている。「Fireflyは、関係管理や高度な個別化のための商品やサービスの主要な供給者である」。これが何を意味しているのかというと、消費者と商品との関係は、「高度に個別化した経験を提供するために」、そして「個人のプロフィールを元にした個別化した推薦を提供するために、この価値ある情報を」集め動かしていくことによってなされうるのである。こういった個別化した推薦は、ユーザの嗜好データベースを代理店側からの視点で分析することから生じる。その嗜好は問題となっているユーザーの嗜好と比較される。こういった技術を用いることによって、類似の課題を持つ利用者のグループは、例えば映画や音楽といった共通の趣味に関する情報を収集し共有していくことが可能となるのである。それらはしばしば推薦によってモデルにされる。「もし君が『トータルリコール』を気に入ったならば、君はおそらく"True lies"も気に入るだろう。なぜならば君の好みのカテゴリーは84%の確率でこれをはじきだしてきているのだから」。こういったタイプのWebページを図書館サービスに応用することの可能性は、オンラインカタログで苦労している人には明白であろう。

## おわりに

結論として、私は図書館員に、Webに取り組みはじめ、人間だけが作り上げることの出来る情報を精査するシステムをつくりだしてほしいと望みたい。図書館員は、当然、インターネットの急速な普及に閉口し、ネット上の資源と、それまで慣れ親しんできた資源（オンラインデータベースのような）とを意味なく比較しようとしがちである。結局、図書館員は情報のエキスパートであると考えられている。彼らはインターネットの成長と普及を驚きでもって受け止めている。皮肉なことに、その職業が情報アクセスを促進するという役割を担っているにもかかわらずである。既にわれわれにとって、驚きから回復し、ただのURLリストの羅列ではない、ネット上に本質的なものをつくりだす時がきているのである。図書館員は、技術と人との間を取り持つ大切な役割を果たしていくべきであり、その役割は今日ほど重要になったことはかつてなかったのである。彼らはコンピュータ科学者にはないユニークな能力を持っており、技術屋や設計者の視点からではなく利用者の視点から情報アクセスを理解しているのである。インターネットは、次のようなことが実行されるまで、その真の潜在能力を決して発揮しないであろう。インターネットを理解している者がアクセスアプリケーションを設計し、Webサイトを評価し、そのサービスを提供し、インターネット上の情報政策に対してリーダーシップを発揮していくことが必要なのである。もし図書館員がこの試みに失敗するならば、彼らはその職と、そしてこういったガイダンスを求めている利用者とを失うことになるだろう。そして、消極的な反応が続く限り、他の者がその空間を埋めようと躍起になるであろう。

## 注および参考文献

[1] fireflyに関する引用はfirefly webpage (www.firefly.com)から得た。
[2] Epictetus, Discourses. Shneiderman, p.577.
[3] Shneiderman, Ben. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Reading, Mass.: Addison-Wesley, 1998, P.565.
[4] Patrick Lynch, Yale C/AIM Web Style

Guide, 1997, <http://info.med.yale.edu/caim/manual/contents.html>

[5] Ben Shneiderman, Designing the User Interface, Addison-Wesley, 1998, p.567.

[6] Jakob Nielson, "Applet Usability: Stepping Outside the Page", Alertbox, Oct. 15, 1997, <http://www.useit.com/alertbox/9710b.html>

[7] Nielson, <http://www.useit.com/alertbox/9710b.html>

# 「学術情報センター紀要」投稿規定

## 投稿資格

1．学術情報センターの職員、各種委員会委員、セミナー等の講師及び研修員
2．学術情報センターを利用する大学及び研究所等の職員
3．他の機関において学術情報センターと関連する研究、開発及び運用等に従事する者
4．その他編集委員会が適当と認める者

## 投稿原稿

投稿原稿は、学術情報センターの事業に関連する次のいずれかで、刊行時において未発表の原著であること。

1．研究論文
2．事例／調査報告
3．レビュー／展望
4．資料

ただし、編集委員会が特に指定した原稿はこの限りではない。

## 執筆要領

1．原稿は、原則としてフロッピー・ディスク及びプリントアウトにより提出する。
フロッピー・ディスク及びファイルの形成は以下の通りとする。
- サイズ　3.5インチ2HD
- フォーマット　MS-DOS(1.44MB)フォーマット
- ファイルの種類
  - 文字　MS-DOSテキストファイル
    ASCIIコードまたはSHIFT-JISコード
  - 図表等　Post Scriptファイル
- 分量　図表等を含め、刷り上がり(2段組)約50枚以内とする。

2．原稿の構成要素には次を含むこと。
邦文標題及び英文標題
執筆者名(ローマ字による読み表記とも)及び所属(邦文及び英文)
要旨(邦文及び英文)
キーワード(邦文及び英文)
本文(原則として邦文又は英文)
注記／引用文献
なお、英文論文に関しては必ずしも邦文を必須とはしない。

3．図・表はそのまま印刷できるように別紙に正確に作成し、図・表の番号・題名を欄外に記入し、かつ本文に挿入箇所を指示すること。

## 原稿の採否

投稿原稿の採否は、専門家(原稿テーマによっては、編集委員会以外の者に依頼)による査読の後、編集委員会において決定する。

## 著者校正

校正は、著者に依頼する。

NACSIS