新しい情報組織化技術と
自然言語データの検索・編集への応用

佐藤理史

■研究のねらい

近年のインターネットの急速な普及は、今まさに、我々の社会を、情報を中心とした新しい社会へと導きつつある。インターネット上に構築されたワールドワイドウェブ (World Wide Web) は、ほとんどあらゆる種類の膨大な情報を持つに至っており、我々は、居ながらにして、これらの情報に自由にアクセスできるようになった。しかしながら、その一方で、あまりにも膨大な情報の中に有用な情報が埋もれてしまい、必要なときに必要な情報を見つけ出すことができないという問題が、深刻な問題として浮上してきた。また、長期的には、この膨大な情報をうまく整理して知識化し、後世に残していくことも考えていかなければならない。これらは、**膨大な情報をどうやって有効に活用できるようにするか**という問題であり、これからの情報化社会の発展のためには、この問題を解決することが必要不可欠である。

本研究は、「自動編集＝編集の自動化」という新しいアイディアに基づき、上記の問題の解決を目指すものである。ここで、編集とは、人間が情報を有効に使いこなすための知的作業全般を指す。編集には、次の２つの機能的な側面がある。

1. **情報をわかりやすくする**
   情報を利用目的に合った内容や形式に加工して提示すること。
2. **情報をつかいやすくする**
   情報を組織化し、多くの人々が利用しやすい形式にまとめあげること（情報の体系化と知識化）。

これらを機械的に実現することができれば、上記の問題に対する有効な解決策となる。本研究では、レジュメを自動編集するという課題を設定し、その実現に取り組んだ。

■研究成果

本研究の主要な成果は、**カテゴリにガイドされた自動編集**の提案と、その考えに基づく自動編集システムの実現である。これらにより、比較的簡単なレジュメが自動編集できることを実証した。

1. カテゴリにガイドされた自動編集

本研究で想定するレジュメの自動編集とは、次のような課題である。

> 「Xとは何か」という質問に対して、ウェブを調査し、それに対する答を1頁程度のレジュメとしてまとめること。

任意のXに対してレジュメを編集することは、短期間で実現できるような課題ではない。そこで、まず、Xのカテゴリに着目し、このカテゴリを限定した範囲で、レジュメの編集を実現する。

たとえば、「Xさんって誰？」という質問に対してレジュメを編集する場合を考えよう。この場合、編集対象のカテゴリは「人」である。このことにより、次のように編集方針が定まる。

- どんな情報源から情報を収集すればよいか — 有望な情報源は、人名録や紳士録、その人物のホームページなどである。

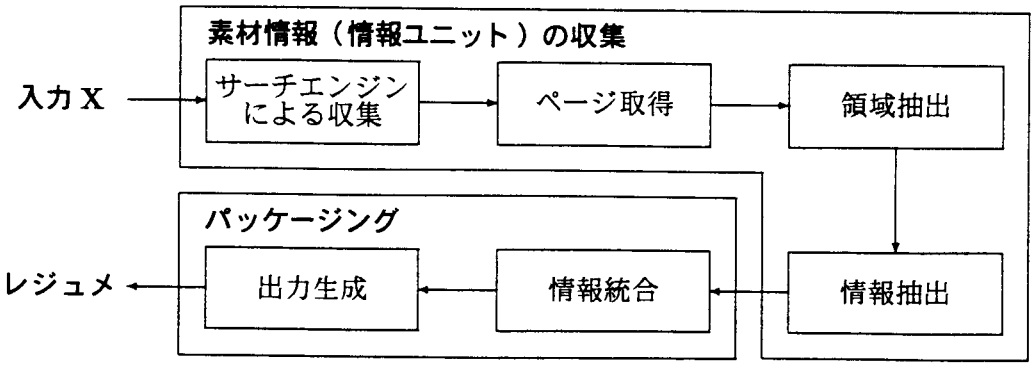

図 1: ハイパーテキストレジュメ自動生成システムの構成。作成した 3 つのシステムは、同じ基本構成をとる。

- レジュメにはどんな情報を含めるべきか — 本名、生年月日、職業、ホームページの URL などが主要情報となる。
- レジュメをどんな形式にまとめるべきか — 人物に関するレジュメの典型例は、書籍や論文における著者紹介（プロフィール）である。

編集方針が定まれば、編集作業を自動化するプログラムを設計することが可能となる。このように、編集対象のカテゴリを基づいて自動編集を実現する方法論を、**カテゴリにガイドされた自動編集 (category-guided automated editing)** と呼ぶ。

2. ハイパーテキストレジュメの自動編集

上記の方法論に基づく自動編集をサーチエンジンの高度化に応用したものが、**カテゴリを限定したサーチエンジン**である。本研究では、このようなシステムの構成法を確立するとともに、実際に、住所探索、人物情報探索、用語説明探索の 3 つのシステムを作成した。これらのシステムは、(1) 知りたいことに対する簡潔な答え、(2) 答えを作成するために使用した情報源、(3) さらに詳しい情報を知るために有用なリファレンス、の 3 つの要素から構成される簡単なレジュメ（**ハイパーテキストレジュメ**）を出力する。

システムの構成を図 1 に示す。システムは、大きく、(1) 素材情報（情報ユニット）の収集と、(2) パッケージング、の 2 つの要素から構成される。

素材情報の収集では、与えられた入力に対して、まず、サーチエンジンを利用して、素材情報が掲載されている可能性が高いページの URL を収集し、それらのページを取得する。次に、取得したページを解析して、素材情報が記述されている部分を同定し（領域抽出）、その部分から素材情報を抽出する。これらのうち、技術的に難しいのは、領域抽出と情報抽出である。本研究では、HTML タグを利用した領域抽出法と、住所情報、人物情報、用語説明のそれぞれに対する情報抽出法を実装した。

パッケージングでは、収集された多数の素材情報を整理・統合して、それらをコンパクトな形にまとめることを行なう。これを実現するために、本研究では、属性の識別能力に基づく情報統合法という新しい方式を開発した。

作成した 3 つのシステムの概要は次の通りである。

1. 住所探索システム WIT/Doko
   与えられた名称に対する住所情報（住所、電話番号、URL）を探し出す。本システムの入力ページと出力例を図 2 に示す。この図に示すように、本システムは、同名の別対象を弁別する能力を持つ。

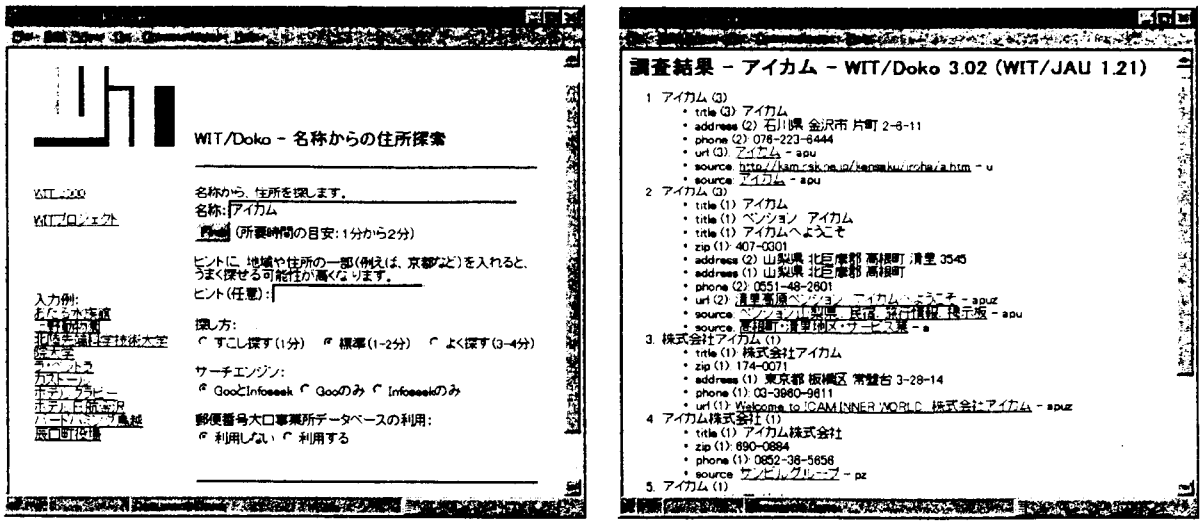

図 2: 住所探索システムの入力と出力。与えられた名称に対して、その住所情報をウェブから自動的に探し出す。同名の別対象を弁別する能力を持つ。この結果から、「株式会社アイカム」と「アイカム株式会社」が異なる会社であることがわかる。

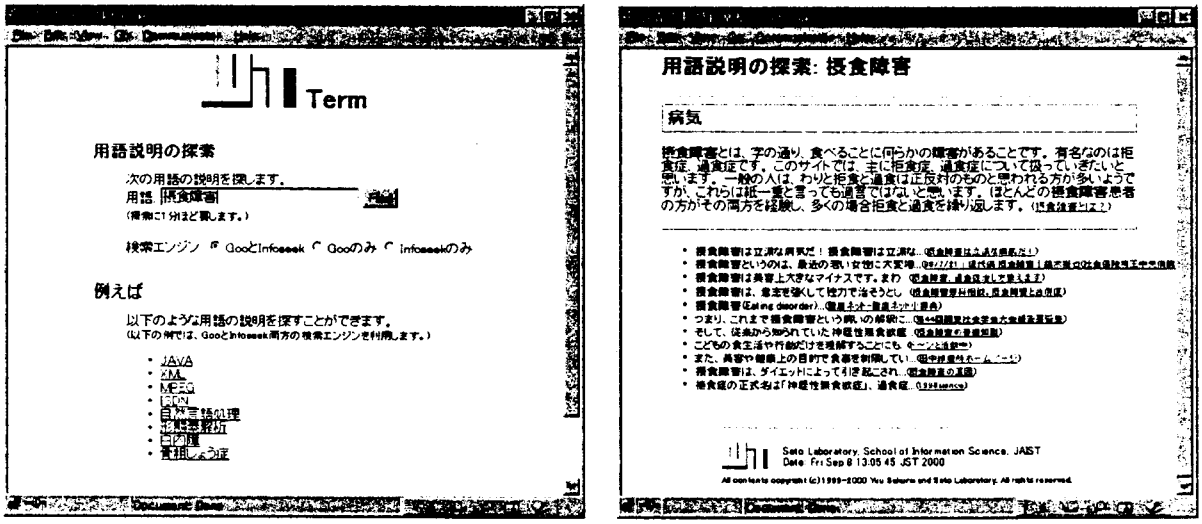

図 3: 用語説明探索システムの入力と出力。与えられた用語を説明する文章をウェブから自動的に探し出し、語義毎に整理して出力する。比較的新しい用語や通常の辞書に載っていないような専門用語の意味を調べることができる。

2. 人物情報探索システム WIT/Who
   与えられた氏名に対する人物プロフィールを探し出す。
3. 用語説明システム WIT/Term
   与えられた用語を説明する文章を探し出す（ウェブを仮想辞書化することを実現する）。入出力例を図 3 に示す。

これらのシステムは、典型的な情報要求（「住所を調べたい」、「ある人物について知りたい」、「ある用語の意味を知りたい」）に対して、その答えと、より詳しい情報へのポインタ（リファレンス）を提供することができる。

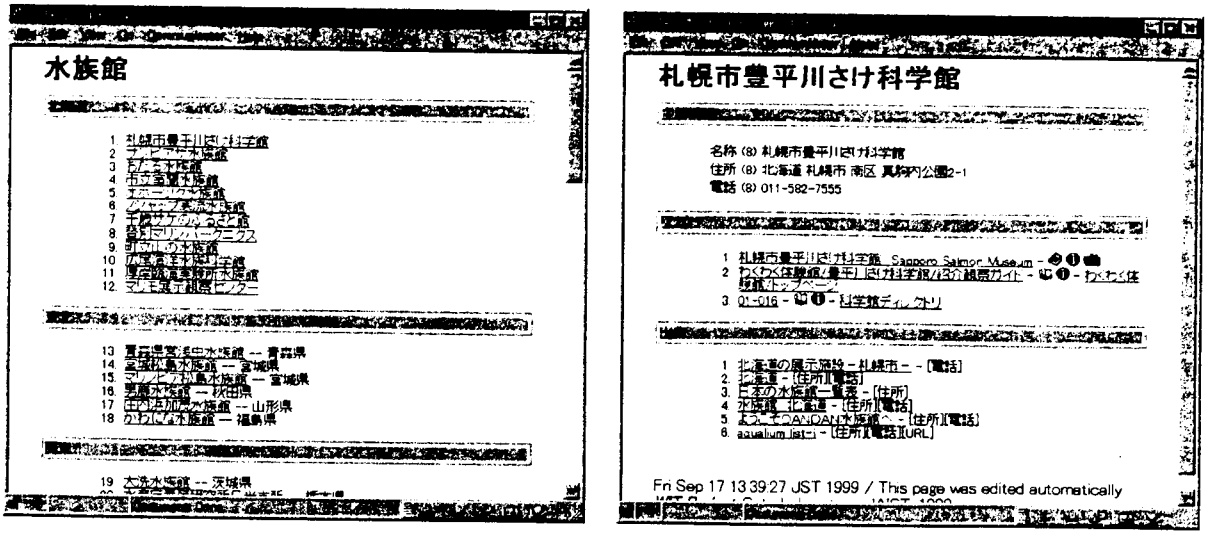

図 4: 自動編集された水族館ディレクトリの一部。左図は、全国の水族館を地域別に整理したページであり、それぞれの水族館をクリックすると右図のようなレジュメページが表示される。

3. ウェブディレクトリの自動編集

レジュメの編集をあるクラスに属する要素集合に対して適用し、それらを一つのパッケージとしてまとめれば、そのクラスに対するウェブディレクトリが構成できる。本研究では、次の 2 つのウェブディレクトリ自動編集システムを実現し、これまで人手で作られていたウェブディレクトリが自動的に作成できることを実証した。

1. 特定カテゴリ施設のウェブディレクトリの自動生成 (WIT/Links)
   「水族館」、「動物園」、「美術館」などのカテゴリ名を入力とし、そのカテゴリに対するウェブディレクトリ（リンク集）を自動生成する。まず、与えられたカテゴリ名（「水族館」）から、そのカテゴリに属する施設の名称（「おたる水族館」や「海遊館」）を収集する。次に、住所探索システムを用いて、それぞれの施設に対するレジュメページを作成し、最後に、これらを地域毎に整理して、目次ページを作成する。自動生成された水族館ディレクトリの目次ページと札幌市豊平川さけ科学館のレジュメページを図 4 に示す。
2. 地域情報ウェブディレクトリの自動生成 (WIT/Regional)
   日本の地域情報を 3427 の地方公共団体（市町村）毎に整理したディレクトリを自動生成する。このシステムは、上記のシステムとは異なり、まず、それぞれの地域の情報が掲載されているサイト（地域サイト）を発見・収集し、それらのサイト内のページを自動分類することにより、レジュメページを生成する。本システムは、2852 地域 (83.2%) に対して総計 4012 の地域サイトを発見することができた。これは、既存の地域情報ディレクトリに収録されているサイト数を凌駕する数である。自動編集されたディレクトリの一部を図 5 に示す。

4. 編集とは何か

本研究の隠されたテーマの一つは、「編集とは何か」ということを明らかにすることにあった。我々は、編集がおおよそどのようなものであるかを知ってはいるが、そこで行なわれていることを明快に説明する理論を持っていない。「編集とは何か」を説明すること自身、まだ解かれていない研究課題なのである。

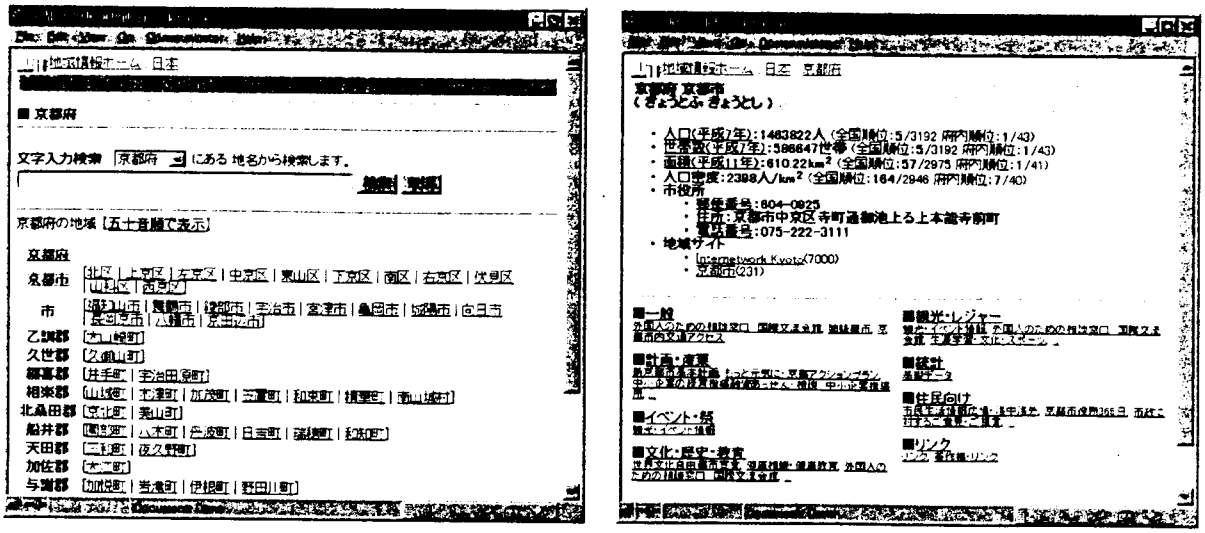

図 5: 自動編集された地域情報ディレクトリの一部。日本全国の地域情報が地方公共団体毎に整理されている。

本研究によって到達した認識は、次のようなものである。

「編集とは、デパッケージングとパッケージングである」

本研究を進めている過程で明確に意識するようになったのは、情報パッケージと情報ユニット（素材情報）の区別である。これらは、編集の製品 (product) と素材 (material) である。ここで重要なことは、それらは編集処理における相対的な概念であるということである。つまり、編集処理を一つ設定したとき、その処理における情報パッケージと情報ユニットが定まる。

これをもう一歩押し進めると、「情報ユニットの実体は、（より小さな）情報パッケージである」という考えに行き着く。たとえば、水族館ディレクトリの編集では、まず、各水族館に対してレジュメを作り、次に、それを束ねてディレクトリを作成する。ここで、各水族館に対するレジュメは、最初の編集処理（レジュメ編集）においては、情報パッケージである。こうして作られた情報パッケージは、次の編集処理（ディレクトリ編集）においては、最終的なディレクトリ（情報パッケージ）を構成するための素材（情報ユニット）として使われる。つまり、１つの情報パッケージは、より小さな情報パッケージ群から構成される。情報パッケージは、このような再帰的構造を内在する。その意味において、あらゆる情報は情報パッケージであると考えてさしつかえない。

レジュメの編集では、処理を情報ユニットの収集とパッケージングの２つのフェーズに分けた。もし、必要とする情報ユニットがそのままの形で入手できるならば、情報ユニットの収集は単なる収集である。しかし、ほとんどの場合、求める情報ユニットは、より大きな情報パッケージの一部として存在するため、そのパッケージを分解して、必要とする情報ユニットを取り出すという作業が必要となる。これを**デパッケージング (depackaging)** と呼ぼう。情報抽出や要約抽出といったものは、すべて、デパッケージングの技術に含まれる。

パッケージングとは、ひとまとまりの情報を作り出すことである。これに関しては、さらに理解を深める必要があるが、おおよそ、情報ユニット群に構造を持ち込むこと、あるいは、ある枠（フレーム）に情報ユニット群をはめ込むこと、と考えてよいと思われる。Mok の organization model や、Rosenfeld と Morville の organization systems (organization schemes and organization structures) が、これらをより明確に説明する足がかりとなろう。

デパッケージングの機械化が難しいのは、パッケージングによって持ち込まれた構造が、最終プロダクトに必ずしも明示的に記述されないという点に起因する。情報パッケージの提示におい

ては、我々の常識を利用した簡素化（たとえば、文脈からわかることは省略するなど）や意匠が凝らされており、それが機械的な認識の妨げとなるのである。

編集の品質は、コントロールされる情報パッケージのレベルの深さと関係している。たとえば、新聞の記事（情報ユニット）を集めてきて、それらを単に束ねただけであれば、編集の品質は低いと見なされる。一方、その記事をいったんデパッケージングして、見出しの長さ、記事の長さ、書き方などを統一化する処理（一段深いレベルでの情報パッケージング）を行なえば、編集の品質はより高いものと見なされる。

本研究で作成した自動編集システムがコントロールしているレベルは、せいぜい2段か3段であり、一言で言い切ってしまえば、「切り張り」の域を出ていない。より多くのレベルで編集処理を行うこと—すなわち、より詳細なレベルまでデパッケージングを行い、最終的なプロダクトの隅々まで、調和をもたらすこと—が、これからの自動編集の新たな目標となる。

■今後の展開

この3年間の研究は、紆余曲折があったものの、「情報の自動編集」というアイディアの具体化とその実現手法の検討、および、デモンストレーションシステムの実装などを行なうことができた。これらにより、自動編集という新しい研究領域の大枠を示せたのではないかと考えている。その意味において、本研究は、十分な成果をあげることができた。

しかしながら、本研究の途上で設定した例題—レジュメの自動編集—は、完全な形で実現できたわけではない。これまでに実現できたことは、簡単かつ典型的なレジュメを「切り貼り」によって生成することであり、それは、人間が行なっている編集のほんの一部分に過ぎない。今後は、「切り張り」を越えて、テキストの加工や複数情報の融合を含む、より高度な自動編集の実現に向けて研究を進める必要がある。特に、テキスト加工の切札となるべき、テキストの言い換え（パラフレーズ）の自動化は、本研究の中でも試みたが、十分な成果をあげることができなかった。この問題に再びチャレンジする必要がある。

本研究で作成した自動編集システムは、単なるデモンストレーションのためのシステムではなく、いずれも、現在のニーズを考慮し、実際的に使えるシステムを指向したものである。さらに改良してソフトウエアシステムとしての完成度を高めれば、実用システムのレベルに達することができよう。このような実用化の方向も今後模索していきたいと考えている。

■成果リスト

- Satoshi Sato and Madoka Sato. Rewriting Saves Extracted Summaries. *Intelligent Text Summarization*, Technical Report, SS-98-06, American Association for Artificial Intelligence, pp76–83, 1998.3.
- Satoshi Sato and Madoka Sato. Rewriting in Automated Editing of QA-Pack. *JSPS-HITACHI Workshop on New Challenges in Natural Language Processing and Its Application*, Central Research Laboratory, Hitachi, Ltd., pp59-64, 1998.5.
- 佐藤理史. 論文表題を言い換える. 情報処理学会研究報告, Vol.98, No.82, 98–NL–127, pp187–194, 1998.9.
- 佐藤理史, 奥村学. 電脳文章要約術 —計算機はいかにしてテキストを要約するか—. 情報処理, Vol. 40, No. 2, pp. 157–161, 1999.2.
- 佐藤理史. インターネットのためのテキスト処理. 言語処理学会第5回年次大会チュートリアル資料, pp33–42, 1999.3.

- 佐藤理史. 論文表題を言い換える. 情報処理学会論文誌, Vol.40, No.7, pp2937–2945, 1999.7.
- Satoshi Sato and Madoka Sato. Automatic Generation of Web Directories for Specific Categories. Research Report, IS-RR-99-24I, School of Information Science, Japan Advanced Institute of Science and Technology, 1999.7. (Presented in *AAAI Workshop on Intelligent Information Systems, Orlando, July, 18-19, 1999*)
- Satoshi Sato and Madoka Sato. Toward Automatic Generation of Web Directories. *Proc. of International Symposium on Digital Libraries 1999 (ISDL'99)*, pp127–134, Tsukuba, September 28-29, 1999.
- 佐藤理史. ワールドワイドウェブを利用した情報探索. 言語処理学会第6回年次大会, pp447–450, 2000.3.
- 桜井裕, 佐藤理史. ワールドワードウェブを利用した用語検索. 情報処理学会研究報告, Vol.2000, No.53, 2000–NL–137–4, pp23–29, 2000.6.
- Satoshi Sato, Automated Editing of Hypertext Résumé from the World Wide Web, The 2001 Sumposium on Applications and the Internet (SAINT-2001), in press.

知的情報処理技術の新展開

橋田　浩一

人工知能や自然言語処理などの知的情報処理技術において、1980 年代は、意味や推論や文脈などのスローガンを高らかに掲げ、自律的な知的主体を工学的に実現するという夢に燃えた時代だった。しかし、そこで暗黙の内に前提されていた、「孤立した思惟」としての知の捉え方には、状況に埋め込まれた行為や相互作用などを強調する論者によって 1980 年代の初期から批判が加えられ、知的情報処理の研究はこれを受けて創発や状況依存性の考え方を取り入れつつ新たな方向へと発展する。

しかしいずれにせよ、自律的知性の設計はきわめて困難なテーマであり、短期的な成果が挙がるはずはない。こうして、実世界のデータと統計的な手法を用いた、定量的な性能評価や統計的な学習が可能なテーマに関する研究が、1980 年代の後半から広く行なわれるようになった。自然言語処理においても、古く 1960 年代に構想された統計的手法が復活し、現在に至っている。言うまだもなく、1990 年代半ばからのネットワーク環境の普及による大量の電子データの流通はこの流れを加速した主要な要因のひとつである。この意味で 1990 年代はデータの時代だったと言えよう。

ところが、統計的手法により大量のデータをして語らしめるデータの時代も今や過ぎ去ろうとしている。統計的手法はすでに練れていて急速な発展は望みにくいというのがそのひとつの理由だろう。しかし、より重要な理由は、多数の事物の全般的傾向を扱う方法としての統計的手法の効力が、社会の情報化につれて減衰しているということである。

電子的ネットワークの普及によって大量の情報が流通するようになったため、大量のデータの処理に対する社会的要請が生ずるとともに研究用のデータが入手しやすくなり、こうして上述のようにデータの時代が加速された。だが、社会の情報化がさらに進展して産業構造を変えるまでになると、事情はそれほど単純ではなくなる。しばしば指摘されるように、情報化に伴って生産者と消費者とが直結され、取引が一対一になれば、規格品や画一的なサービスの大量生産と大量消費を特徴とする産業構造が崩壊する。こうして、統計的手法に代表される大量で定型的な情報処理の効力はもはや絶大なものではなくなり、逆に、個別的なサービスのための個別的な情報処理の必要性が高まっている。各情報ユニットの意味的な構造に依存し、個々の顧客の個別的な情報ニーズに対応するための情報処理が求められているのである。

たとえば情報検索では、統計的な方法に基づいて候補を順位付けたり、絞り込みのために追加すべきキーワードを提案したりしている、全般的な傾向を捉えるのが本来の目的である統計的な方法によって個別的な検索要求に対応するのは原理的に不可能である。所望の情報に到達するには、いわゆる意味ネットワークのような個別的な意味構造を手掛かり

とする研削の方法が必要と考えられる。

情報の編集は、研削、抽出、変換、提示などを、きわめて重要な総合的技術である。大量の情報へのアクセスはとどまるところ編集に尽きると言っても過言ではない。当然ながらそこでもまた、単なる統計処理ではなく、個別的な意味構造を扱う情報処理が主要な役割を演ずる。上記のような検索の側面だけでなく、抽出や変換も、意味構造に依存して行なう必要がある。佐藤の言う情報パッケージ、情報ユニット、およびデパッケージング、パッケージングなどはまさにこのことを指摘した概念と言えるだろう。

1990年代がデータ（表層的な情報）時代だったのに対し、2000年代は、以上のような意味において、コンテンツ（意味内容）の時代だと言える。ただし、ここで言う意味内容とは、1980年代以前の「孤立した思惟」における「意味」や「文脈」とは異なり、社会的に共有され、さらに広い共有と再利用を意図するものである。いわゆるＩＴ革命とは実はコミュニケーション革命であり、その本質は情報の効率的な共有と再利用にある。既存の情報の編集は情報の共有と再利用のための強力な手段であり、コンテンツの時代に欠かせない技術である。

このように考えてくると、知的情報処理技術に関する研究の今後の展開においては、2つの流れが絡み合いながら進むものと予想できる。ひとつはさまざまな意味でのインタラクション、もうひとつは情報世界と物理世界との融合である。

情報の共有と再利用が重要なのは、良質のコンテンツを生産するために相応のコストがかかるからであり、それは、その生産が主として機械ではなく人間によって行なわれるからである。したがって、人間による知的生産を人間と機械のインタラクションによって共有・再利用が可能な形に定着させる必要がある。情報の自動編集や半自動的な意味構造化（タグ付け）などの技術はそのためのツールになる。

情報の共有と再利用とは、情報世界の意味構造化を社会的に共有するということである。一方、われわれの身体は物理的時空にあり、本来の生活がアナログな世界で営まれている。したがって、情報世界と物理世界を融合し、意味構造化を物理世界にも拡大することにより、人間の生活全般を知的に支援する体系的なサービスを提供することが、知的情報処理技術の次の目標であろう。物理的環境を含む状況に埋め込まれた行為というテーゼが、こうして社会的要請と直結したリアリティを帯びることになる。自律的な知的行為者の設計という、知的情報処理技術の究極の目標が、そこから垣間見えてくる。