

## 高信頼性組込みソフトウェア構築技術

北陸先端科学技術大学院大学 情報科学研究科 片山卓也

研究分担者 早稲田大学理工学部 コンピュータネットワーク工学科 中島達夫

研究分担者 京都大学 湯浅太一

研究分担者 国立情報学研究所 中島震

協力企業 NEC, NEC エレクトロニクス、松下電器産業、Nokia、オムロン、オムロンソフトウェア

### 1. プロジェクトの背景

組込みソフトウェアは、家電製品、自動車、携帯端末、制御機器などのあらゆる工業製品の心臓部に組み込まれ我々の生活を支えており、その機能や品質はこれら製品や機器の価値を決める最も重要な要素である。特に、ソフトウェアの不具合による経済損失などが報告される現在、高信頼性は組込みソフトウェアに対する最も重要かつ緊急な要求である。

従来、組込みソフトウェアはそのサイズが余り大きくなかったこと、また、比較的単純な機能の実現を行えば良かった事などもあって、その開発には最新のソフトウェアテクノロジーが用いられてこなかった。しかしながら、現在では、高度なユーザインタフェースや通信機能など製品に要求される機能が高度化すると同時に、利用可能なCPUやメモリなどのハードウェア資源に対する制約が緩和したことなどによって、組み込まれるソフトウェアが大規模化・複雑化し、これまでのソフトウェア開発方法論が十分に機能しなくなりつつある。最新のソフトウェア開発技術を組込みソフトウェアの開発に投入し、組込みソフトウェア開発を前進させることが強く求められている。

本プロジェクトでは、このような観点から高信頼組込みソフトウェア構築の問題を、(1) 構築環境、(2) 実行環境、(3) 実行基盤の3つの視点から総合的に解決することを目的とし、企業との密接な連携のもと、プロジェクト終了時に産業界への技術移転を目標に、以下の課題に関する研究開発を行っている。

#### (1) 組込み用オブジェクト指向分析設計技術

(北陸先端科学技術大学院大学)

#### (2) 組込みシステム向け基盤ソフトウェア (早稲田大学)

#### (3) 組込み用実時間 Java 技術 (京都大学)

### 2. 組込み用オブジェクト指向分析設計技術

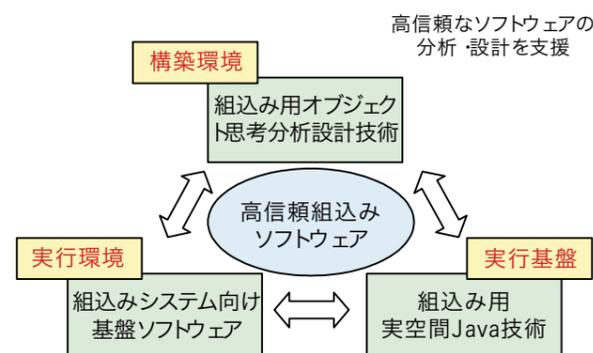
ソフトウェア開発上の問題の多くが分析・設計などの上流工程で作られ、それが全体の生産性や品質に支配的な影響を及ぼしていることはよく知られている。(1)では、組込みソフトウェアの特性を適切に扱うことが可能で、上流工程段階から正しさを確認・検証しながら組込みソフトウェアを構築できる方法論とそれを支える環境の実現に関する研究開発を行っている。具体的には、UMLによる記述、モデル検査や定理証明技術による検証、製品系列やアスペクト概念による設計などを可能にする環境の実現を行っている。

### 3. 組込みシステム向け基盤ソフトウェア

携帯電話やデジタルテレビ、あるいはこれから登場する先進情報アライアンスにより様々な新しいサービスが提供されることが予想される。これらの次世代の情報端末や情報アライアンスには、高度な通信や協調動作、セキュリティ、センシングなど多くの機能が要求されるが、(2)では、これらを実現する高信頼コンポーネントやミドルウェア、それらを動作させるオペレーティングシステムなどの基盤ソフトウェアの研究開発を行っている。

### 4. 組込み用実時間 Java 技術

組込みソフトウェアでは、一定の限られた時間内にレスポンスを返すこと、また、限られたメモリで動作することなど、動作時間や利用可能資源などに厳しい制約がついているのが普通である。従来はプログラマが細心の注意を払って、職人芸的技术によってこの問題を解決しており、開発コストおよび品質の両面で大きな問題となっていた。(3)では、これらの作業を最大限に自動化し、信頼性の高い組込みソフトウェアの生産性を高めるための研究開発を行っている。



## 組込み用オブジェクト指向技術

北陸先端科学技術大学院大学 情報科学研究科 岸知二、青木利晃、片山卓也

国立情報学研究所 中島震

協力企業 NEC, NEC エレクトロニクス

<http://kt-www.jasit.ac.jp/project/esociety>

### 1. プロジェクトの概要

ソフトウェアの問題の多くは上流工程で作られており、産業界でも設計品質に対する問題意識が高まっている。本プロジェクトでは高信頼な組み込みソフトウェア開発のための分析・設計の手法や環境を研究している。

研究にあたっては、ソフトウェア開発に関する最新の工学的、科学的成果を、実際の組込みソフトウェア開発に適用することを試みている。例えばオブジェクト指向開発、アスペクト指向開発、プロダクトライン開発といった工学的成果や、宇宙・航空・軍事といった分野でのみ利用されてきた高度だが高コストで利用の難しかった形式的手法やシミュレーションなどの科学的成果を、家電製品のような民需分野での組込みソフトウェア開発へ適用することを目指している。

具体的には、形式的手法のひとつであるモデル検査技術と呼ばれる技術を活用し、UMLを用いた設計の正しさを検証する手法について検討を進めており、今年度はUML設計を行うためのツールの実用性を高めるとともに、それを活用した事例研究、さらに中期的に必要な基本的な技術の開発という観点から研究を行った。

### 2. 2005年度の成果

以下に本年度の主要な成果について報告する。

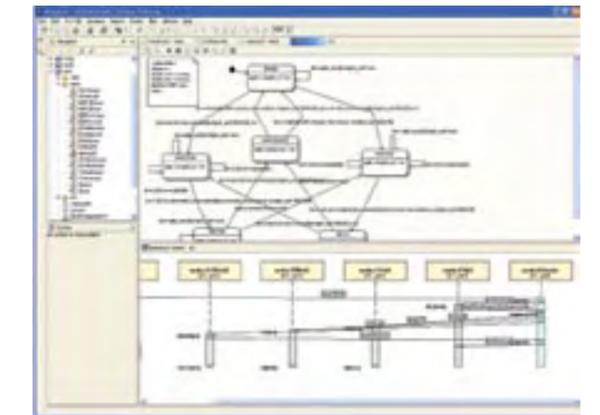
#### 2.1 UML 検証ツールの開発

昨年度までに開発した評価版とその評価に基づき、実際に使って評価をもらうためのベータバージョンを開発した。本バージョンでは、UMLで記述されたモデルを、モデル検査ツールのひとつであるSPINによって検証する機能を実際開発で利用しやすくするために、UMLモデルの解釈、SPINの入力言語への変換方法などを目的に応じてカスタマイズする機能や、ひとつのUMLモデルに対して行う複数の検証項目を設定し、それをバッチ的に検証する機能などを強化している。現在本ツールはいくつかの企業や大学において評価を行っている。

#### 2.2 事例研究

協力企業であるNECエレクトロニクス社から提供していただいたカーオーディオの設計を題材に、開発したツールを用いて設計検証を行った。設計モデルのサイズが大きいため、モデル検査可能な精密さで全体をモデル化することが実際的でないため、まず人間が読むためのUML設計モデルを作成した上で、

検証目的に応じて厳密な検証用UMLモデルを作成・検証するというアプローチをとった。この事例を通じて、各種の適用ノウハウや課題を整理することができた。



### 2.3 基本的技術の検討

設計検証に関わる、以下のような様々な基本的技術の研究開発を行った。

- ・アクション付きクラス図の検証法：クラス図に対して記述されたアクション言語の検証を、ステートチャート図を用いてメソッドの呼び出しをモデル化することにより行う手法を提案した。
- ・RTOS(Real-Time Operating System)に基づいたソフトウェアのためのモデル検査ライブラリ：RTOSを用いたプログラムの検証を容易にするために、μITRON4.0のサービスコールをシミュレーションするSPINのライブラリを開発し、それを基にした検証手法を提案した。
- ・UML/STD動作規則の拡張と形式化：組み込みソフトウェアの設計に適したステートダイアグラム(STD)の動作規則を提案するとともに、ダイアグラムからSPINのための入力記述を生成するソフトウェアを試作した。
- ・ストレス試験フレームワークへの応用：タスク間の依存関係を変化させても性質が成り立つかどうかを確認する試験の方法を、前項の動作規則に基づいて検討した。

## 組込みシステム向け基盤ソフトウェア

早稲田大学理工学部 コンピュータネットワーク工学科 中島達夫  
協力企業 松下電器産業, ノキアジャパン  
<http://www.dcl.info.waseda.ac.jp/>

### 1. 次世代組込み環境と信頼性

組込みシステムもユビキタス時代の到来により日常品にコンピュータを埋め込み情報を扱う様々な情報アプライアンスが出現すると思われる。情報アプライアンスは単体で機能するのではなく、組み合わせることにより様々な新しいサービスを作り出すことを可能とする。しかし、情報アプライアンスのソフトウェアは巨大なものとなり、ソフトウェアのバグによりシステム全体の信頼性を保証することが困難となる。また、情報アプライアンスはインターネットに接続されるため、様々なセキュリティに対する脅威からシステムを守る必要がある。組込みシステムは従来のエンタープライズシステムと異なり、リソースの制約を考慮したり、コンピュータの専門家以外の人でも用意に扱える必要があるため、従来の手法と異なる新しい技術の開発が必要となる。

### 2. 2005 年度プロジェクト概要

早稲田大学のグループでは、組込みシステムの信頼性を向上する 2 つのオペレーティングシステム開発の取り組みに関して取り組んでいる。ここでは、それぞれの OS における 2005 年度の成果概要を紹介する。

#### 組込み Linux 上のアカウントシステム

前年度開発した組み込み Linux 上の QOS サポートシステムを機能拡張してアカウントシステムとして実装をおこなった。アカウントシステムは、アカウントオブジェクトという抽象化の提供をおこなう。アカウントオブジェクトは周期と実行時間の 2 つのパラメータを持つ。アカウントオブジェクトをバインドされたプロセス群は、それらが各周期ごとに決められた実行時間のみを消費することを保証する。

アカウントシステムの応用としては、ダウンロードしたプログラムが規定した以上の CPU キャパシティを用いないようにすること、リアルタイムクラスと非リアルタイムクラスのプログラムを柔軟に調停すること、システム全体のオーバロードを検出すること、リアルタイムプログラムのための CPU リソース予約などが考えられる。

アカウントシステムは、極力 Linux カーネルを変更しなくても使用できるようにデザインされている。そのため、Linux カーネルは頻繁に変更されるにもかかわらず、アカウントシステムの維持を容易におこなうことが可能となる。

我々が現在開発したアカウントシステムは、日本エンベデッドリナックスコンソーシアムのリソースマネジメントワー

キンググループにおける標準化のリファレンス実装として用いられている。また、CE Linux フォーラムでも実際の家電機器で使用が適切に関する検討が進められている。

#### デバイスドライバの隔離化

Linux では、デバイスドライバはカーネルと同一のアドレススペースに存在するため、オペレーティングシステムをクラッシュさせる大きな原因となっている。

我々が開発した NEST システムは、ワシントン大学が開発した Nooks を拡張し、デバイスドライバと Linux カーネルを異なる空間で実行することを可能とする。NEST は Nooks と異なり、デバイスドライバを変更する必要がない。そのため、システムの移植性が大きく向上する。

Nest システムでは、デバイスドライバがメモリアクセス障害を起こしても、異なるアドレススペース内で稼働している Linux カーネルに影響を与えることはない。そのため、ドライバ内にバグが存在してもカーネル部に影響を与えることはない。また、デバイスドライバのバグが非決定的なバグである場合は、デバイスドライバを再起動するのみでシステム全体を再起動せずに実行を続行することが可能である。

実際に、フォールトインJECTIONテストをおこなったが、従来の Linux ではクラッシュしてしまったバグが生じても、NEST を用いた場合は、実行を続行することが可能である。また、システムのオーバヘッドについても実用的に問題とならないレベルであることが明らかになった。

## 組込みシステム用実時間 Java 技術

京都大学 大学院情報学研究科 湯浅太一  
<http://www.yuasa.kuis.kyoto-u.ac.jp/>

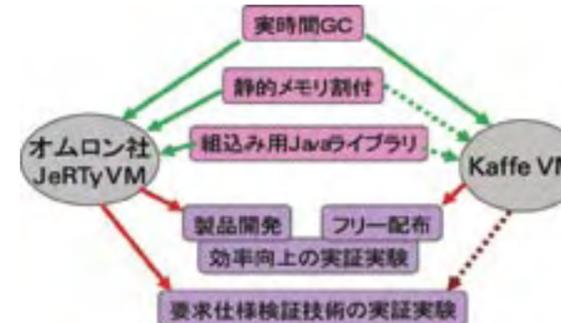
### 1. 概要

本プロジェクトは、オブジェクト指向言語 Java によって記述される組込み実時間アプリケーションの開発を効率化するための諸技術を開発するものであり、次のサブテーマから成る。

- ・実行基盤の開発
- ・実証実験
- ・要求仕様検証技術との統合

実行基盤の開発については、実時間組込みソフト固有の開発コストを軽減するための、自動化機能を備えた Java 処理系実装方式を開発し、試験的実装を行ってきた。具体的には、不要データの回収処理によるアプリケーション実行の中断を回避するための実時間 GC や、限られたメモリ空間を有効に利用するための静的メモリ割付、これらに対応できる基本的な組込み用ライブラリ群を、オムロン社の組込み用実時間 Java 処理系である JeRTy VM に試験実装するとともに、フリーの Java 処理系である Kaffe VM にも実装した。

平成 17 年度は、これらの成果に基づいて、実証実験を可能とするための作業および、要求仕様技術との統合に向けた作業を行った。以下にその概要を示す。



### 2. 実証実験に向けて

平成 16 年度までに試験的に実装した JeRTy VM 上の実時間 GC を、実用に耐えるレベルにまで改良し、従来版に較べて、アプリケーションの実行性能が平均して約 3 割向上させることに成功した。この処理系を、オムロン社が開発する組込み機器製品に実際に利用し、動作確認を行った。この製品をはじめ、いくつかの実用機器において、実時間性能などを検証していくことが今後の課題となる。

また、JeRTy VM 自体は、ライセンスの関係のために単体での一般公開はできないので、開発技術を試験利用したい潜在的ユーザーのために、開発技術を実装した Kaffe VM の一般公開を開始した。下記の Web ページからフリーソフトとしてダウンロード可能である。

URL: <http://kt.jaist.ac.jp/project/esociety/>

### 3. 要求仕様検証技術との統合に向けて

要求仕様検証技術を実際に利用するのはソフトウェア開発現場のエンジニアであるとの立場から、本プロジェクトでは、共同研究先であるオムロン社関係のエンジニアに、昨年度まで開発したライブラリを中心に、実際にいくつかのソフトウェアについてモデル検査を試験的に実施した。実施に先立って、北陸先端大や国立情報学研究所などの先生がたの講習会を開催あるいは受講するとともに、いつでも支援を行える体制を整えた。

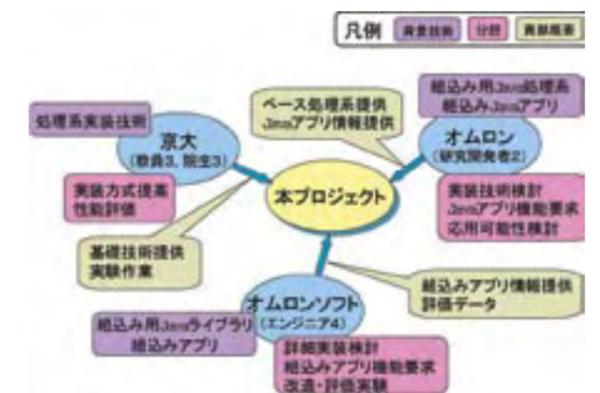
モデル検査の対象としたのは、並列性を含むために動作確認が容易でない以下の四つである。

- ・ JeRTy 簡易 Web サーバ機能
- ・ ブート処理サービス機能
- ・ Timer 機能
- ・ オムロン社開発中製品の通信プロトコル部分

最初の三つは、昨年度までに開発した組込み用ライブラリの一部である。モデル検査は初めてのエンジニアたちであったが、京都大学の支援のもとで、いくつかのバグを検出した。四つ目の製品プログラムについては、すでにバグが見つかったプログラムを使用し、実際にモデル検査によってそのバグが検出できることを確認した。

### 4. 実施体制

本プロジェクトは、オムロン (株) およびオムロンソフトウェア (株) と共同で実施している。その概要を次の図に示す。



## 次世代高性能コンピュータシステム上の高信頼ソフトウェアシステムの開発支援技術

東京大学 大学院情報理工学系研究科 コンピュータ科学専攻 石川 裕  
<http://www.il.is.s.u-tokyo.ac.jp/>

### 1. はじめに

複数のプロセッサや I/O 機器が動的に接続 / 拡張可能なハードウェアを有する次世代高性能アーキテクチャが出現した。これら機器は 10Gbps の性能を有する高速 LAN で接続されるようになる (図 1)。このようなアーキテクチャは従来のサーバコンピュータにとって変わることが可能となる。しかし、当該アーキテクチャ上で高信頼性を実現するためのシステムソフトウェアが未整備であり、現状、当該アーキテクチャの利用形態は限られたものとなっている。次世代高性能コンピュータシステム上で高信頼性を実現するためには、プロセッサ間通信機構、オペレーティングシステム、運用ツール等、多岐に渡りシステムソフトウェアを開発しなければならない。このようなソフトウェア開発のための支援技術および基盤ソフトウェアモジュールが求められている。

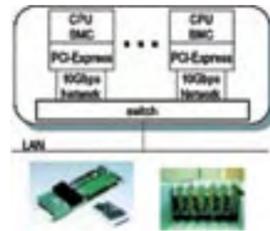


図 1 次世代高性能コンピュータシステム

本研究では、次世代高性能コンピュータシステム上で信頼性を実現するシステムソフトウェアを高信頼かつ効率的に開発することを可能とするため、計算機ハードウェアやネットワークの障害等を模擬するためのソフトウェア環境を構築した。また、保守・運用ツールなどの基盤ソフトウェア開発に必要な基本通信機構として、保守・監視系のための基本機構、OS 核間で使用する高速通信機構の開発を行った。

### 2. 管理保守ソフトウェア & 故障模擬ソフトウェアシステム

近年の PC サーバでは、IPMI と呼ばれる規格に基づいたハードウェア障害を監視するプロセッサ (BMC) がボード上に搭載されている。IPMI 機能を利用することにより、ハードウェア障害状況に応じた耐故障機能を組み込んだ高信頼システムソフトウェアが実現可能となる。このような高信頼システムソフトウェアが設計通りに実現されているかをテストするために、IPMI 規格に基づくハードウェアの故障模擬ソフトウェアシステムを開発した (図 2)。

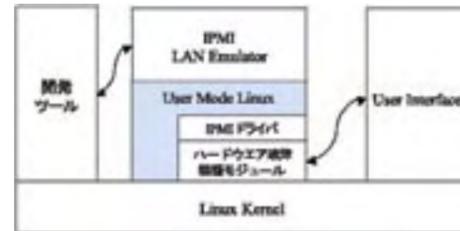


図 2 IPMI に基づく故障シミュレータ

本システムは以下の特徴を持つ。

- \* CLI および GUI による障害発生指示機能
  - \* 実機上の IPMI 規格に基づくハードウェアセンサ情報を取得し、故障シミュレータに設定する機能
- 高信頼システムを支援する保守監視系ツール Tenjin を開発した (図 3)。

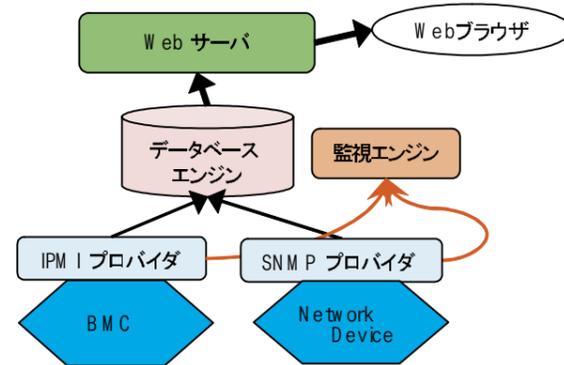


図 3 Tenjin ソフトウェアアーキテクチャ

Tenjin では、次のような機能が実現されている。

- \* WEB サーバ  
サーバサイドプログラミングにより、機器の表示系、システム運用ポリシーの設定などを行なう。ATML (A Temporal Markup Language) と呼ばれるタグ体系を開発し実装した。
- \* 監視エンジン  
機器監視を記述する Tenjin 監視スクリプト言語を開発した。Tenjin 監視スクリプト言語と ATML は一つの言語として設計されている。すなわち、ATML のタグ体系に対応して Tenjin 監視スクリプト言語が定義されている。
- \* データベースエンジン  
DMTF(Distributed Management Task Force) が規格化している CIM (Common Information Model) に基づいたデータベースエンジンを開発した。実装には MySQL サーバを用いている。
- \* プロバイダ  
プロバイダは、機器情報を取得し、データベースエンジン

に情報を登録、あるいは、機器からの Alert メッセージを受け取って、監視エンジンにインディケーションを発生させる機能を提供している。

さらに、平成 17 年度、高信頼システムを実現するための要素技術として Linux 上にプロセスマイグレーション機能を開発した。本機能は、TCP コネクションを維持したまま他のコンピュータへのマイグレーションを可能としている。

### 3. 高速通信機構 & 故障模擬・開発支援システム

10Gbps 級 LAN を用いて高信頼高性能通信を実現するためには、障害発生時の性能劣化を最小限にする高い信頼性を持つ新たなノード間通信プロトコルの開発が必要となる。このためには、障害発生を想定した動作検証が必要である。しかし、10Gbps 級のネットワーク環境における通信プロトコル開発において、動作検証のために様々な障害が発生するネットワーク環境を実際に生成し、動作検証することは困難である。このため、試行錯誤と経験則によるデバッグや最適化を行うこととなる。このような、原始的な開発手法では、十分な動作検証や最適化は難しい。

十分な動作検証や最適化のためには、障害が発生するネットワーク環境を模擬し、通信プロトコルの動作を定量的かつ視覚的に把握する必要がある。そこで、10Gbps 級のネットワーク環境において様々な障害が発生するネットワーク環境を生成し、その状況を検証するツール XGE-ProtoDevel を開発してきた。開発支援システムは、ソフトウェアだけでは実現できない。パケット喪失、CRC エラーなどを生成可能とするネットワークカード (Uzura と呼ぶ) を開発した (図 4)。Uzura は 10Gbps Ethernet 規格準拠のネットワークであり、PCI-X でホストコンピュータと接続される。さらに、本装置を利用して、ネットワークカード側でリモートメモリ間通信を実現する機能を開発してきた。

平成 17 年度、本研究の応用範囲を増やすために、クラスタ上の科学技術計算アプリケーションで必要となる配列転置機能をネットワークカード側で実現する転置 RDMA 機能を実現した。配列の転置を伴うデータ転送は、FFT(Fast Fourier Transform) など多用される。Uzura ネットワークカードを搭載した 2 台の Intel Xeon 2.8GHz コンピュータ上で FFT ライブラリの一つである FFTW を用いて評価した。転置 RDMA 機能により、転置転送処理時間を 40.8% 削減、FFT にかかる処理時間全体を 34.9% 削減できた。



図 4 故障模擬装置付 10Gbps ネットワークカード

### 4. デモンストレーション

2005 年 11 月に開催された国際会議 SC05 においてシステムのデモンストレーションを行った。本デモンストレーションの概要を図 5 に示す。従来のシステムで実現できていなかった TCP/IP コネクションを保持したままプロセスマイグレーションする機能を示すために以下の手順でデモンストレーションを行った。

1. Compute ノードからクライアントに TCP/IP を用いて画像データを送信。
2. Compute Node 上のファンを故障させる。
3. 管理ノードが異常を検知し、Compute ノード上のプロセスを Standby ノードに移送し、Compute Node をシャットダウンする。

## プログラム自動解析に基づく高信頼ソフトウェアシステムの構築技術

### —高信頼プログラミング言語と高信頼プログラム開発環境—

東北大学 電気通信研究所 大堀 淳

協力企業 沖電気工業(株), 算譜工房(有)

<http://www.pllab.riec.tohoku.ac.jp/>

#### 1. プロジェクトの背景と目的

今実現しつつある高度情報化社会が、従来通りの信頼性と安全性を確保しながら発展していくためには、社会基盤としての高信頼ソフトウェアを効率よく構築する技術を確立する必要がある。特に以下の2点の実現が求められている。

1. プログラムの誤りを開発の初期段階で検出し高信頼プログラムの効率的生産を可能にする技術、
  2. C, Java 等の既存言語で書かれたプログラム部品を利用し、高信頼プログラムを効率よく開発する技術。
- これら機能の実現には、従来のソフトウェア開発技術を越えた新たな技術革新を必要とする。その最も有望な基盤が、型理論に基づくプログラムの自動解析技術である。この技術は、システムに基大な障害を引き起こす恐れのあるプログラムの不整合をコンパイルの段階で自動検出することを可能にする。

本プロジェクトは、型理論を基礎とし、上記2つの機能を実現の中核となる

1. 次世代高信頼プログラミング言語、及び
  2. 次世代高信頼プログラミング言語をCやJavaなどの既存言語とともに使いこなすためのプログラム開発ツール
- を実現し、さらにそれらを統合し、高機能高信頼システムの効率的な生産を可能にする高信頼ソフトウェア開発環境の実現を目指す(図1参照)。

これにより、ソフトウェア生産のボトルネックとなりつつあるテスト、デバッグ、保守コストの大幅な削減、さらにプログラムエラーに起因する壊滅的なシステム障害のリスクの大幅な低減が実現できると期待される。本プロジェクトでは、我々が保有する、世界的にみても最先端の型理論と型主導コンパイル技術を基礎としてこれら技術を実現する。

#### 2. プロジェクトの具体的目標

以上の目的を達成するために、本プロジェクトでは、以下のソフトウェアを開発する。

##### 1. 次世代高信頼プログラミング言語 SML#.

本プロジェクトの中心は、次世代高信頼プログラミング言語 SML# の開発である。SML# は、既存の高信頼言語である Standard ML を、我々の基礎研究の成果を取り入れて拡張した次世代言語であり、レコード多相性やランク1多相性などの新しい先端機能を実現するとともに、実用上重要なCなどの既存言語とのデータの相互運用性を関数型言語として初めて実現する。本プロジェクトでは、SML# の言語仕様、コンパイル

理論とコンパイル方式、コンパイラ、実行時処理系のすべてを開発する。

##### 2. 高信頼プログラム開発ツール

SML# を実用システム開発に応用するためには、既存言語とともに SML# を使いこなし、高信頼プログラム開発を効率よく行いことを可能にするプログラム開発環境が必要である。この実現のために、以下のツール群を開発する。

- (ア) SML# のプログラミングを支援する種々のプログラミング自動化ツール
- (イ) Java で書かれたプログラム部品を組み合わせて使用する場合の不整合を自動検出するツール
- (ウ) SML# を C, Java, およびデータベースシステムなどの既存システムとともに利用するための相互運用サポートライブラリ。

これらは、ソフトウェア生産の高信頼化に直接貢献するばかりでなく、高信頼ソフトウェアの研究の新たな展開のシーズとなると期待される。特に、SML# 言語は、レコード多相性や相互運用性などの先端機能を世界ではじめて装備した言語であり、高信頼ソフトウェア開発の生産性を大幅に高めるばかりでなく、日本発の最先端高信頼プログラミング言語のリリースによる世界への情報発信と、その波及効果として、我が国の高信頼技術の世界への普及等が期待できる。

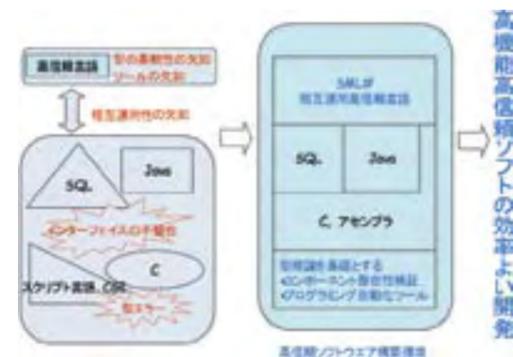


図1. プロジェクトの概要

##### 3. 実施体制

これら先端ソフトウェアを、システム開発現場にて活用できる実用性あるソフトウェア基盤として開発するためには、ソフトウェア開発を行っている企業との緊密な連携が不可欠である。本プロジェクトでは、種々のツールの有用性の現場に即した評価や分析およびプログラム開発を、沖電気工業(株)および算譜工房(有)と共同で行う体制を取る。

#### 4. SML# 言語

SML# は、以下の特徴をもつ静的に型付けられた関数型プログラミング言語である。

- A) 多相型レコード演算およびランク1多相性により、現在のML言語より強力な柔軟なプログラミングが可能。
- B) データベース, JAVA およびCとの高い相互運用性を実現
- C) 現在のML言語の標準仕様である Standard ML を包摂。特にA)とB)の機能は、国際学会レベルで広く認められた我々の基礎研究成果であり、その実用上の重要性から、プログラミングの国際的に著名な教科書でその実現が期待され、また研究プロジェクトなどでそれを越える機能の実現が目指されながら、いまだに実現していない最先端機能である。

本プロジェクトでは、以下各項目についてその基本方式およびその実現アルゴリズムを構築し、SML# を、高信頼実用ソフトウェア開発に耐える実用言語として構築し、世界にリリースする。

1. レジスタ抽象機械
2. 抽象機械コードへの型主導コンパイラ
3. 基本ライブラリ
4. 相互運用実行時ライブラリ
5. 最適化処理
6. ネイティブコードコンパイラ

この中で1から3までの項目が、平成17年度までに完成させリリース予定の項目、4から6が平成18年度から19年度までに開発し、プロジェクト終了時点でリリース予定の項目である。

平成17年度は

- ・モジュール言語のコンパイル方式とコンパイル処理
- ・外部関数呼び出し機構
- ・必須と指定されたすべての基本ライブラリ関数
- ・数千項目にもおよぶ系統的なテストの実施

の各項目の開発を行い、SML# 言語のフルセットの抽象機械コードへのコンパイラを完成させ、平成18年3月に、SML# 言語の解説等を含むwebページを以下のURLに構築し、<http://www.pllab.riec.tohoku.ac.jp/smlsharp/> コンパイラのアルファリリースを行った。

#### 5. プログラム開発ツール

SML# を、CやJavaデータベースシステムなどの既存のシステムと共に使用し、効率的な高信頼プログラム開発を支援する以下のツール群を開発する。

##### 1. SML# プログラミング自動化ツール群

新しい言語であるSML#を効率よく使用するための種々の自動化ツールを開発する。以下のものを含む。

- (ア) 清書プログラム自動生成ツール (SMLFormat)
- (イ) ドキュメント自動生成ツール (SMLDoc)
- (ウ) テスト支援ツール (SMLUnit)
- (エ) デバッガ

平成17年度はSMLUnitを開発し、平成16年度に開発したSMLFormatおよびSMLDocとともに、リリースを行った。デバッガは、平成18年度以降に構築を予定している。

##### 2. Javaでプログラム部品の不整合を自動検出ツール

現状のJavaの部品の結合に関する脆弱性を静的に取り除くことを目的としたツールであり、SML#などと共に、高信頼システムを構成する上で重要なツールである。平成17年度は、部品間の依存性解析関数およびGUIを開発し、予定通りツールを完成させた。

##### 3. SML# 相互運用サポートツール群

新しい言語であるSML#を実用言語として普及させるためには、CやJavaなどの既存言語やデータベースなどの既存システムとのスムーズな相互連携が鍵となる。そのためのサポートライブラリやツール群を開発する。

本項目は、プロジェクト後半の平成18年度以降に構築予定のものである。

##### 6. まとめと今後の展望

Cとの同一のメモリー構造に基づくシームレスな相互運用性の実現方式などの課題をすべて解決し、予定どおりSML#言語の抽象機械へのコンパイラの開発に成功した。これによって、本プロジェクトの目標の達成目処がたったと言える。

今後は、次の大きな課題であるSML#のネイティブコードへのコンパイラ、Javaおよびデータベースシステムとの相互運用性の実現、実用上有用な種々のプログラミングツールの開発等、プロジェクト後半で予定している課題の達成に取り組むとともに、それら開発と同時に、それら開発によってより高機能かつ堅牢になるSML#コンパイラを継続的にリリースし、研究成果の普及を図る予定である。

## 安全なシステム記述言語および高信頼 OS

東京大学大学院情報理工学系研究科コンピュータ科学専攻 米澤 明憲  
<http://www.yl.is.s.u-tokyo.ac.jp/e-society/>

### 1. 高信頼基盤ソフトウェアの必要性

高度情報化に伴い、コンピュータウィルスや不正アクセス、情報漏洩などの問題が、社会の安全性を脅かす深刻な事態となってきた。今後、さらなる情報化は不可避であり、コンピュータを利用した社会基盤の安全性を保証することが早急に必要である。

すでに我々の周りには数多くの応用ソフトウェアが稼働しており、その多くに安全性の問題があると考えられている。それら個々の問題を解決していくことは対症療法にすぎない。問題の根本は、これらの応用ソフトウェアが、安全性の問題が社会的に顕在化する以前に開発された基盤ソフトウェア（言語、OS）を利用している事にある。

本プロジェクトでは、型理論をはじめとするプログラムの解析技術によって、既存の基盤ソフトウェアの信頼性を強化することを通して、それらを使用する応用ソフトウェア全体の安全性を向上させることを目標とする。

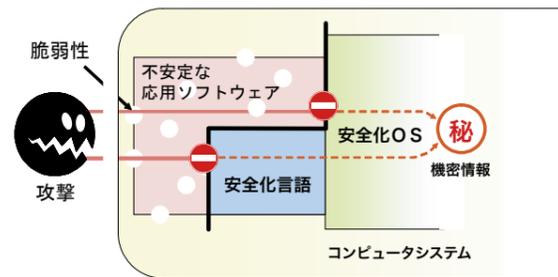


図1 言語とOSによる二重の防御

### 2. プロジェクトの概要

本プロジェクトでは、次の三つの方向性から上記の目標に取り組んでいる：

#### 高安全 C 言語コンパイラの開発

C 言語はその実行速度やハードウェアよりの記述能力から、Java などの新言語登場後の現在もシステム設計では最も一般的であり、多くのアプリケーションにも広く使われている。しかし、これらの C 言語で書かれたプログラムが誤動作したり、攻撃を受けることで、機密情報を盗まれる情報被害が多発している。これは、C 言語自体にはそもそも情報の機密性という概念がないこと、また、言語によるメモリ保護機構が欠くため、メモリ管理のバグによって容易にセキュリティホールが発生することに問題がある。

この問題を解決し、既存の C 言語で欠かれた基盤ソフトウェアをできるだけ自動的に安全化するため、安全なメモリ管理と機密情報の流れを監視する情報流解析機構を導入した、対攻撃耐

性を持ったコード生成コンパイラ VITC を開発する。

#### OS 用型付きアセンブリ言語の設計・実装

従来高級言語に使われていた型理論を適用した型付きアセンブリ言語をさらに拡張し、OS カーネルのスレッド切替機構、割込処理機構、デバイスドライバ等のアセンブリ言語で書かれた構成要素を再記述することで、OS の安全性を大幅に向上させる。また、この手法は、通常の OS だけにとどまらず様々なハードウェアにおいて、それに密着したソフトウェアの安全性保証に適用することもできる。

#### 定理証明器による OS 安全性の形式的検証

メモリ管理などの OS が提供する機能の安全性を定理証明器を用いて形式的に検証する技術を構築し、コンピュータシステム全体の信頼性を保証する。

### 3. 高安全 C コンパイラの開発

#### (1) 対攻撃耐性コード生成コンパイラ (VITC)

コンピュータウィルスをはじめとするセキュリティーホールへの攻撃からプログラムを防御し、機密漏洩を防止するためには、まずメモリ安全性を保証した C 言語コンパイラが前提となるが、それだけではサーバなどのソフトウェアの安全かつ継続的なサービス提供には十全ではない。なぜなら、メモリ安全性が保証されていたとしても、攻撃を受けた後のプログラムの実行継続は予期せぬ副作用を生じ、機密漏洩を引き起こす可能性があるからである。そこで本研究ではメモリ安全な C 言語コンパイラ技術に機密漏洩を防止する、型による情報流解析を組み合わせ、既存の C プログラムから、攻撃を受け不安定となっても機密漏洩を引き起こさないようなコードを生成するコンパイラ VITC の作成を目指す。

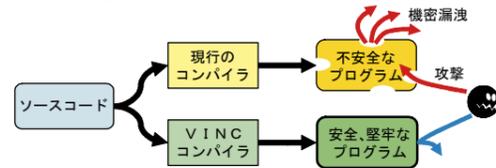


図2 VITC コンパイラ

型による情報流解析の研究は、Java や ML など、理論的に厳正な静的型システムを与えることができる言語系においてもっぱら行われ、情報流もほぼ静的に、プログラム実行前に解析することができた。それに対し、C 言語の型システムは非常に柔軟ではあるが正確性を欠く。そのため、既存研究と同様の静的情報流解析のみでは C 言語の表現力を保ちつつ機密漏洩を防ぐことは難しく、実行時の情報流の動的検査が不可欠である。そ

こで我々は昨年度までの VITC の型理論およびプロトタイプを進展させ、C 言語特有の機能であるキャストとポインタ周辺での動的検査の理論を構築、実装した。動的検査は実行効率を低下させるが、可能な限り動的情報流検査を省略することで、キャストのないプログラムではメモリ安全な C 言語コンパイラ単独と同等の実行効率が達成できた。

#### (2) 高い精度・効率・利便性を持つ情報流解析の研究開発

VITC の核となる情報流解析の精度・効率・利便性をさらに高めるために、静的型システムに加えてモデル検査器を導入した解析法の理論研究を行った。そして、理論の有効性を示すために、C 言語のサブセットを対象とした機密漏洩発見ツールのプロトタイプを開発した。特長として、漏洩箇所と漏洩条件の提示が可能であるために、ユーザの問題理解・修正の困難さが軽減されている。

### 4. OS 用型付きアセンブリ言語の設計・実装

#### (1) OS 用型付きアセンブリ言語 TALK と TOS カーネル

近年、静的プログラム解析技術、特に型理論が飛躍的に進歩し、多くのアプリケーションプログラムが「強く型付けされたプログラミング言語」（例：Java、C#、OCaml 等）を用いて作成されるようになった。これは、強く型付けされた言語で記述されたプログラムは、実行時に予期せぬエラーを生じない、ということが保証されるためである。

ところが、コンピュータを動作させる上で最も基礎的なプログラムである OS（オペレーティングシステム）は、未だ強く型付けされていない言語を用いて作成されている。このため、従来 OS の安全性を保証・検証することは非常に困難であった。実際、安全性が証明された OS は（機能が限定された非常に小さな OS を除けば）未だ存在していない。

そこで本研究では、強く型付けされた安全な言語を用いて OS を作成することにより、OS の安全性・信頼性を向上することを目指す。具体的には、OS カーネルの記述に適した強く型付けされた安全なアセンブリ言語を設計・実装し、これを用いて OS カーネルを記述することで、セキュリティ脆弱性の原因となるメモリエラー（バッファオーバーフロー等）が発生しないことが保証された、安全で高信頼な OS を実現する。

現在までのところ、OS カーネルの重要な機能であるメモリ管理機能（メモリ領域の確保と解放を行う機能）とスレッド管理機能（複数のプログラムコードを同時に実行する機能）を記述可能な強く型付けされたアセンブリ言語 TALK を設計・実装し、これを用いて実際に簡単な OS カーネル (TOS) を作成し動作を確認した。

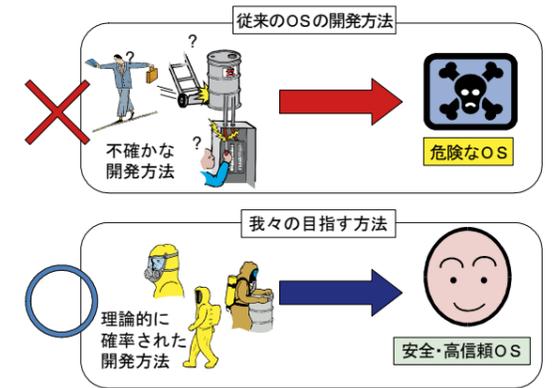


図3. 我々の方法と従来の方法の比較

#### (2) OS 向けの複雑な言語のための型システムの簡素化

従来の型理論は、数学的議論を容易にするため、機能が限定された理論的に簡潔なプログラミング言語を対象にしてきた。一方で、OS 等の現実的なソフトウェアを作成するためには、より実際の複雑な機能を持つプログラミング言語を用いる必要がある。このため、型理論の成果を OS に応用するためには、複雑なプログラミング言語を対象に数学的議論を繰り返す必要があり、これは冗長かつ煩雑であった。

そこで本研究では、このプログラミング言語の理論と現実のギャップを埋めるために、型理論の成果を複雑な言語に直接応用するための理論的枠組みを構築することを目指す。これにより型理論の成果を OS に直接応用することが容易となり、OS の安全化・高信頼化に大きく貢献する。

#### 5. 定理証明器による OS 安全性の形式的検証

通常の OS 上では複数のプログラムが同時に動作しており、メモリやデバイスなどの共通の資源を使用している。OS の仕事のひとつに、各プログラムが占有する資源の分離性の保証、例えば、タスク分離性、あるプログラムの使用しているメモリが他のプログラムによって勝手に読み書きされない性質などがある。OS がそれらの仕事を正しく行うかどうかはコンピュータシステム全体の安全性にかかわる重要な問題であり、間違った OS の上では、たとえ個々のプログラムが正しく書かれていたとしてもそれらの安全性はまったく保証できない。にもかかわらず、現在までこのような OS の安全性の検証はその複雑性からあまり研究されてこなかった。この方面での形式的検証方法を確立するのが我々の研究目的である。

我々は、この OS の安全性、特にタスク分離をターゲットに形式的検証方法の確立を目指した。タスク分離に必要なメモリ領域の分離性を表現できる分離論理 (separation logic) のためのライブラリ Seplog を定理証明器 Coq 上で実装し、それを実際に使って教育用 OS である Topsy のメモリ管理機構がタスク分離性を正しく実装していることを証明することに成功した。

## データ収集に基づくソフトウェア開発支援システム

奈良先端科学技術大学院大学 鳥居宏次

研究分担者 大阪大学 井上克郎

協力企業 日立公共システムエンジニアリング, 日立製作所,

NTT ソフトウェア, SRA 先端技術研究所

http://www.empirical.jp/

### 1. はじめに

本プロジェクトでは, エンタープライズ系, 組み込み系を問わず, 広くソフトウェア開発において, 他の科学や工学分野と同様に, 計測, 定量化と評価, そしてフィードバックによる改善という実証的手法 (エンピリカルアプローチ) の実践を目指す。以下では, 本年度の主な成果の概要を述べる。

### 2. EPM β版の企業プロジェクトへの適用

平成 16 年度に構築したデータ収集・分析システム EPM (Empirical Project Monitor) β版を, 独立行政法人情報処理推進機構ソフトウェア・エンジニアリング・センターとの連携の下, ソフトウェアエンジニアリング技術研究組合に属する 7 社 (トヨタ自動車 (株), (株) デンソー, 日本電気 (株), (株) 日立製作所, 富士通 (株), 松下電器産業 (株), (株) NTT データ (順不同)) によるマルチベンダー開発プロジェクトに導入し, データ収集・分析を行った。分析結果の一例を図 1 に示す。この図は, 開発工程で発生したバグを, 「発見工程」と「発見すべき工程」で分類し, 分類毎に平均修正コストを示したものである。直交欠陥分類法 (Orthogonal Defect Classification scheme. ODC 法) と呼ばれる分析法に基づく結果であるが, 「各ベンダーが基本設計工程で発見しておくべきバグをベンダー間での結合テストまで見逃してしまうと, その修正工数が突出して大きくなる」という従来の定性的な指摘を具体的な数値で示している。

産学連携に基づく特定プロジェクトでの EPM 導入だけでなく, より多くの企業プロジェクトの適用を実現するため, 2005 年 6 月からはプロジェクトホームページ (http://www.empirical.jp/) において EPM β版のアーカイブ方式と coLinux 方式をオープンソースとして公開した。また, 2005 年 8 月からは, coLinux 版英語バージョンを同じくオープンソースとして公開した。2006 年 3 月までのダウンロード数は表 1 のとおり。

| 言語  | 方式      | ダウンロード数 | 公開時期       |
|-----|---------|---------|------------|
| 日本語 | アーカイブ   | 232     | 2005 年 6 月 |
|     | coLinux | 285     |            |
| 英語  | coLinux | 85      | 2005 年 8 月 |

表 1 EPM β版ダウンロード件数

### 3. ESE 支援環境のプロトタイプ開発

定量的なソフトウェア開発データをソフトウェア開発支援に

利用する, いわゆる「エンピリカルソフトウェア工学支援環境 (ESE 支援環境)」の実現に向け, 「開発支援ニーズの明確化」, 「開発支援方式の開発」, 「プロトタイプシステムの設計, 実装」を行った。

開発ニーズの明確化においては, EPM で収集したデータなどから導出可能な多様なデータ分析結果とその傍証データをソフトウェア開発担当者に提示した。提示した分析結果の一例を図 2 に示す。この図は, 「(ある一定期間における) バグ残存率 = 残存バグ数の純増数 / バグ発生数」の時間変化を示している。バグ修正に割り当て可能な人的リソースが不足すると, バグ発生数に比べて修正バグ数は少なくなり残存バグ数は増加するため, バグ残存率は大きな値となる。連携企業におけるソフトウェア開発担当者へのインタビューの結果, 「(結合試験開始直後に) 重大なバグが発生し多くの人的リソースが必要になったために, 発生原因の特定などが済んでいる他のバグの修正が後回しになっていた期間」とバグ残存率が大きな値を示した期間が一致すること, また, バグ残存率がリソース不足の指標として有用であることが確かめられた。これとは別に得られた開発ニーズの例は次のとおりである。

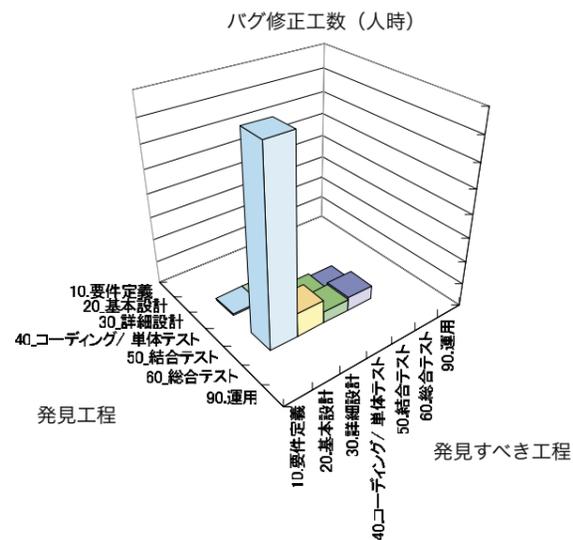


図 1 ODC 法によるバグ修正工数 (人時) の比較

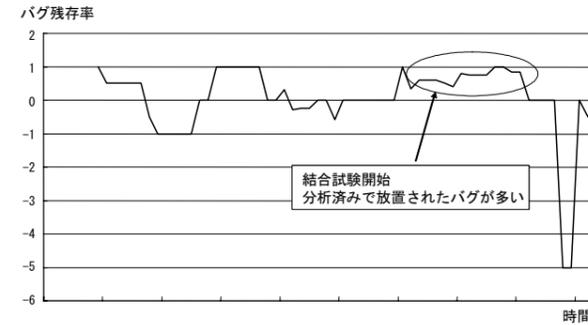


図 2 バグ残存率の時間変化

- ・要員の負荷を知る手がかりが必要である。
- ・コーディング工程の終盤で修正されたソフトウェア機能 / 部品を特定したい。
- ・データの収集・分析には, 既存の企業内システムとの連携が必要である。

開発支援方式の開発においては, データ分析と開発支援の関係を明確に定義可能とするため, JIS X 0141 の測定情報モデルに基づいたデータ定義を用いて, ソフトウェア開発プロジェクトの定量的管理指標を整理し, プロジェクトの特性に応じてカスタマイズする枠組みを提案した。また, CVS やバグ票だけでは判別できない, ソフトウェアの生産性や品質に影響する複雑な要因やソフトウェア開発のコンテキストを分析するための手段として, プロジェクトの模式的な再現により, プロジェクトの進行を直感的に観察する手法を提案した。

プロトタイプシステムの設計, 実装においては, 上記提案方式の設計, 実装に加え, 次のような EPM 機能拡張を行った。

- ・フィルタリング機能: 正規表現による分析対象データ指定。
- ・1 日単位のプロジェクトサマリ機能: ソースコード総行数, 総ファイル数, 総開発者数, 更新者数が n 以上のファイル数, 原因別 / 優先度別のバグ発生数・対処数・平均滞留時間など
- ・ロジカルカップリング検出機能
- ・パレート図出力機能

### 4. CVS 解析システムの構築

CVS データ (複数バージョンのソースコード) を利用して, コードクローン履歴という形で, コードクローンがどのような変遷を辿ったかを分析することによって, 複数バージョンに存在するコードクローン間の関連を抽出する手法を提案し, 提案手法に基づくプロトタイプシステムを開発した。

プロトタイプシステムでは, クローン量の変化をグラフ化することによって, 容易にコードクローン量を監視することが可能である (図 3 参照)。本システムは Eclipse プラグインとして実装されており, Eclipse を用いる多くの開発者は容易に利用することができる。また, クローン履歴の閲覧においても, Eclipse の持つ高度なコード閲覧支援機能を活用することができる。



図 3 CVS 解析システム

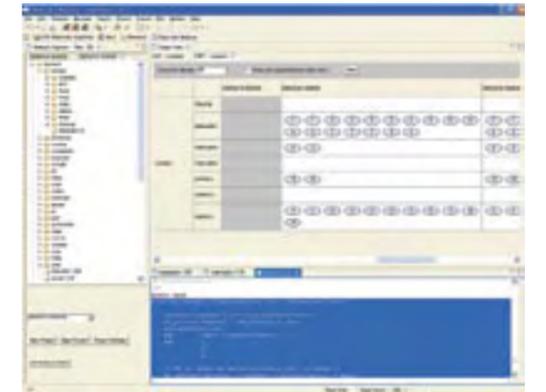


図 4 データ分析の高度化

### 5. おわりに

本年度は, EPM をマルチベンダー開発プロジェクトに適用すると共に, オープンソース化することで, EPM 適用事例を増やし, その有用性や妥当性を検討した。また, EPM 適用結果を開発プロジェクトメンバー (開発者) に提示することで, 定量データに基づく開発支援に対するニーズを明確にすると共に, それらニーズを満足する分析技術の開発に注力した。特に, EPM 適用対象としたマルチベンダー開発は, エンタープライズ系開発と組み込み系開発が混在するプロジェクトであり, その双方において EPM が有用であることが, 開発担当者へのインタビューにより確認できた。

研究開発の成果は, 学術論文 (プロジェクト開始からの 3 年で計 111 編), 雑誌記事 (同 14 件), 新聞報道 (同 6 件), 更には, エンピリカルソフトウェア工学研究会 (年 5 回開催, 参加者各回約 50 名) などを通じて広く発信した。また, 連携企業 11 社 13 名からなる EPM 産業部会では, 産業界ニーズの検討や EPM 支援システムを議論し, 外部有識者からなる評価委員会では, 成果の評価や今後の方針の検討を積極的に行った。来年度は, EPM 提供範囲の拡大やデータ分析の高度化 (EASE 分析法の確立, 図 4 参照) に向けた産学連携, 更には, 海外の研究者や企業との連携などを強化する予定である。

## 高信頼構造化文書変換技術

東京大学 情報理工学系研究科 武市正人

<http://www.psdlab.org/>

### 1. 高信頼構造化文書変換技術の概要

電子的な構造化文書情報の蓄積と効果的な情報利用技術は、インターネットを含む広範な情報の交換・流通にとってきわめて重要な位置を占めている。XML に代表されるこれらの技術は、発展の著しいウェブによる情報環境において既存の技術の延長線上で実務的に開発されたものであり、事実上の標準となつてはいるがその言語的な概念が十分に整理されているわけではない。このような体系的な処理技術の欠如に起因する問題が情報交換の発展を阻害している。すなわち、既存技術の使い回しによる姑息な対処や人手による個別対応が、信頼性および可搬性に欠ける文書情報を蓄積する一因となっているのであり、この問題を早急に解決することが重要な課題となっている。

本研究の中核となる Programmable Structured Document (PSD) は、プログラムの記述を含む文書を対象とすることで、構造化文書処理を効果的に実現しようとするものである。すなわち、PSD においては構造化文書をプログラミングにおける構造化データとみなし、プログラミング言語の理論的基盤を適用することによって、安全かつ信頼性の高い処理を実現する。構造化文書はプログラミングにおけるデータ構造と共通の性質を有しており、その処理を行うアルゴリズムの記述には関数型言語が適している。また、処理を実現するプログラムは対象文書に埋め込まれており、これによって文書の高い独立性と可搬性が実現される。

### 2. PSD の基盤技術

PSD 実現のために必要となる基盤技術としては、(i) 構造化文書に必要な準構造化データの概念を型として捉える形式的枠組の定義およびデータ型に基づく効率的変換手法、(ii) PSD のための計算機構を組み込んだ構造化文書の実現手法、(iii) 関数型言語におけるデータ型の理論の発展と準構造化データに適した型理論の構築、の三つから構成される。(iii) は、さらに代数的プログラム変換(演算)の成果を構造化文書に適用し、自己参照による変換戦略を文書自体に付随させるという演算随伴機構に関する理論を含む。

以上の基盤技術に立脚した PSD を操作するための PSD 処理システムは、文書処理ソフトウェア等による構造化文書の編集から発信にいたる処理を可能とし、構造化文書を変換するためのプログラムの自動生成を実現するものである。このような優れた特徴を持つ PSD 処理システムは、ウェブを含むオフィス環境における基本ソフトウェアとしても位置付けることができる。

### 3. PSD のもたらす効果

本プロジェクトでは、構造化文書に対する PSD 処理機構と変換規則を体系化し、ソフトウェアの信頼性確保の基礎となる言語仕様・設計仕様を設計し、構造化文書処理システムを実用開発する。その成果として、一般の文書作成者が利用できる高信頼 PSD 汎用ソフトウェアパッケージ(統合的構造化文書処理システム)と PSD 構造化文書を対象とした各種アプリケーション開発用のプログラミングシステム(構造化文書変換プログラム生成システム)を開発する。前者は構造化文書を効率よく作成し、それらを対象としたアプリケーションを半自動的に生成することを目的としている。また、後者は主としてアプリケーションソフトウェアの開発者向けのものである。

文書に価値を付与するためには作成時から文書を構造化しておく必要がある。統合的構造化文書処理システムは、それ単独または従来の文書処理ソフトウェア等と連携することで、付加価値の高い文書やコンテンツの作成を行うことが可能であり、情報発信のツールとして個人およびオフィスにおける利用が期待できる。特に、オフィス系ソフトウェアにおける XML 関連ソクくなってきており、構造化文書変換プログラム生成システムの利用は、このような文書の開発の効率を高め、生産性を向上させることになる。

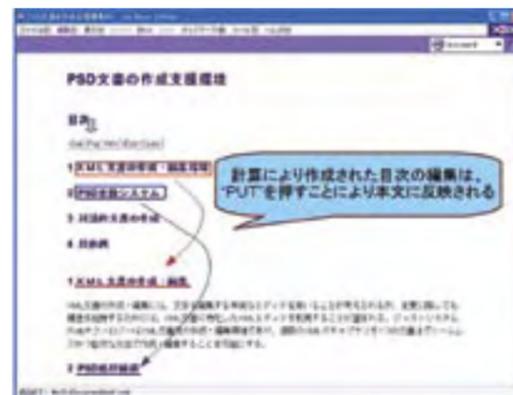
### 4. 本年度の成果

#### PSD 文書作成システムの開発

2005 年度には、これまでに開発した PSD 文書作成システムを改良し、文書中に埋め込むプログラムのコードを柔軟なものにした。その基本技術は 2004 年度に開発した双方向変換にある。

たとえば、文書の本文にあるセクション名を抽出して目次を付加した文書は、セクション名抽出と目次の体裁を整える処理を実現するためのコードを記述することによって得ることができる。これは、文書の内部に目次と本文の中のセクション名が一貫性を保つ必要のあるデータ間の依存関係を有する端的な例のひとつである。本文のセクション名を変更して再度計算を施せば目次も更新されることになるが、場合によってはこの計算によって得られた目次に変更を加え、その計算とは逆向きの計算を施すと本文がその目次の変更に沿って変更されるといったことも望まれよう。このような計算機構は、内部に依存関係のある文書を扱う際の編集作業の効率を大幅に高めることになる。このような順方向(目次の計算)の計算の定義と逆方向(目次の変更の本文への反映)の計算の定義を別個に与えることによってそのような計算機構を実現することは可能では

あるが、それらの計算が相互に逆の変換になっていることを保証する必要もある。昨年度に開発した双方向変換技術は、順方向の計算の定義によって、逆方向の計算が自動的に得られるというものである。これによって、依存関係の存在するデータの一貫性の保証が得られるとともに、順方向・逆方向双方の処理コードを記述するという手間がなくなることになる。本年度はこのような双方向変換による計算機構をインタラクティブ文書 iDocument に組み込み、(株)ジャストシステム社の xfy システムのプラグインとして実現した。



#### PSD 文書変換システムの開発

本プロジェクトにおいては、2004 年度に開発した双方向変換技術を高信頼文書変換技術の中核とすることとして、本年度よりその本格的な取組を開始した。



2005 年度には、文書作成システムで実現したような、依存関係を保証する双方向変換機構を含むより一般的な構造化文書に対する双方向変換のための Java ライブラリ BiXJ を開発した。この BiXJ の表現能力は高く、XML 処理によく利用されている XQuery や XSLT による処理の多くを表現することができる。加えて BiXJ は双方向性を持つため、XQuery や XSLT の処理系では実現できない逆方向への変換も同時に実現することができ、文書変換プログラムの開発に高信頼性ととも高生産性を保証することができる。文書作成システム iDocument の双方向性の実現においても、この BiXJ のライブラリを用いている。

構造化文書変換の高信頼性を確保するためには、双方向変換技術のさらなる課題解決が必要である。XML 文書の変換プログラムを XSLT で記述する際には、言語機能に基づく正当性の証明は困難であり、結果として正しくないコードが蓄積される懸念があるが、本研究では文書変換のコードの自動生成を追究し、文書作成時の編集操作によってコードを生成する XEditor のプロトタイプを開発した。また、双方向変換技術による基本的な動作確認のためのシステムとして、データの依存関係をもつ文書内で、データの変更を双方向性により伝播させる機構を検証するための環境 XDoc を開発した。



#### PSD 文書処理効率化システムの開発

構造化文書の変換処理にあつては、双方向変換技術以外にも解決すべき課題が存在する。2005 年度には昨年度に引き続き、大規模な構造化文書変換のための技術開発を行った。大規模文書を変換する際には、文書全体をメモリ内に置くことができない。ある種の変換処理はストリームとして効果的に実現することができるが、本年度には、木構造処理の典型である再帰に基づく木構造変換モデル内の計算を対応するストリーム処理に置き換えるアルゴリズムを開発し、それに基づいて得られたストリーム処理の高速化法についてもその道筋を示した。

### 5. おわりに

2005 年度は、PSD 文書処理システムの開発を行った。特にそれらの中核となる双方向変換技術の研究を進め、2006 年度以降の高信頼構造化文書変換技術開発の基盤を構築した。

## 高信頼 WebWare の生成技術

名古屋大学 大学院情報科学研究科 阿草 清滋

研究分担者 和歌山大学 システム工学部 鯉坂 恒夫

研究分担者 愛知県立大学 情報科学部 山本 晋一郎

協力企業 富士通研究所, 富士通ソフトウェアテクノロジーズ, 野村総合研究所

<http://www.agusa.i.is.nagoya-u.ac.jp/research/webware/index.xhtml>

### 1. はじめに

本研究の目的は, WebWare の信頼性と安全性の向上のためにエンジニアが行う作業の支援と, デザイナーとエンジニアの協調作業の支援が可能な統合的 WebWare 開発環境を構築することである.

今年度は, WebWare のテスト支援, 解析技術, 作成支援システムにおいて以下の成果を得た.

### 2. WebWare のテスト支援

WebWare は要求の追加・変更が頻繁に発生するソフトウェアである. 特にページを生成するプログラムである JSP には見映えの変更から使い易さの向上まで様々な変更要求が絶えず発生する. そこで, 昨年度は JSP 単体テストの工数削減を目指し, JSP 単体テスト支援ツールを開発した. 本年度は実際のプロジェクトに対して本ツールを適用し, その有用性を評価した.

JSP 単体テスト支援ツールは WebWare の仕様からテストケースを生成し, テストの自動実行を行うツールである. ツール利用者はエディタを利用して生成されたテストを編集したり, 実行するテストを選択したりすることができる. 実際のプロジェクトへ本ツールを適用するにあたり, JSP 単体テストを行う開発者との意見交換を経て, 実運用に必要な機能として (1) テスト実行エラーから回復可能なテスト実行機能と (2) 機械処理可能な証跡保存機能を実現した.

ツール改善の後, 実際のプロジェクト 3 件に対して本ツールを適用し, その評価を行った. ツール利用の有無によるテスト工数の変化を表 1 に示す. 生成されるテストケース数は (A)4,992 件, (B)314 件, (C)912 件である.

|       | A    | B    | C    |
|-------|------|------|------|
| ツール有り | 30.6 | 1.9  | 3.9  |
| ツール無し | 76.2 | 5.77 | 11.6 |

表 1 JSP 単体テスト工数の変化 (単位は人/日)

本ツールを利用することにより適用プロジェクトにおける JSP 単体テストの工数が約 3 分の 1 に削減された. ツールの導入コストに関して, 本ツールは WebWare の仕様からテストを生成するため約 1 人/日と低いコストで導入可能である. また証跡として画面イメージだけでなく画面に表示されている HTML も保存するように機能を改善したことにより, JSP 単体テスト結果で利用した証跡を回帰テストのテストオラクルとし

て利用することが可能となった. これらより, JSP 単体テスト支援ツールの有用性を確認した.

### 3. WebWare の解析技術

WebWare は種々のコンポーネントから構成され, コンポーネント間のデータの受け渡しによって機能を実現する. このため, あるコンポーネントで不具合が発生し, その原因が他のコンポーネントにあった場合, その位置の特定は難しい. これまで, 各コンポーネントそれぞれに閉じたデータ依存解析は確立されているが, WebWare を構成する異種のコンポーネントを跨るデータ依存解析は提案されてこなかった. そこで, WebWare を構成するコンポーネント間のデータ依存関係を定義し, WebWare 全体のデータ依存グラフを提案した. 昨年度の成果である WebWare 解析ツールに対して, WebWare のデータ依存グラフを構築する機能およびデータ依存グラフに基づいてデータフロー解析を行う機能を追加し, WebWare におけるデータフローの把握支援環境を実現した.

WebWare のデータ依存関係は, WebWare を構成する個々のコンポーネントのデータ依存関係とコンポーネント間のデータ依存関係からなる. コンポーネント間のデータ依存関係は, 遷移関係のあるコンポーネントにおける入出力データの対応によって定義される. WebWare のデータ依存グラフは, 各コンポーネントのデータ依存グラフをコンポーネント間のデータ依存関係を表す辺によって接続して構築される (図 1).

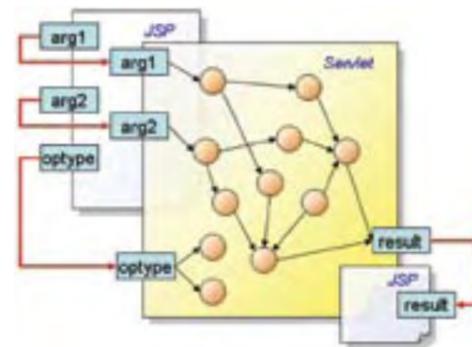


図 1 WebWare のデータ依存グラフ

WebWare におけるデータフローの把握支援として WebWare 解析ツールに対して (1) コンポーネントの入出力データ項目の表示機能 (2) 同一データ項目の表示機能 (3) 遷移発生箇所を表示機能 (4) 依存するデータ項目の追跡機能を追加した (図

2). WebWare のデータフローを把握することにより, 不具合の発生位置と原因の発見が容易になった.

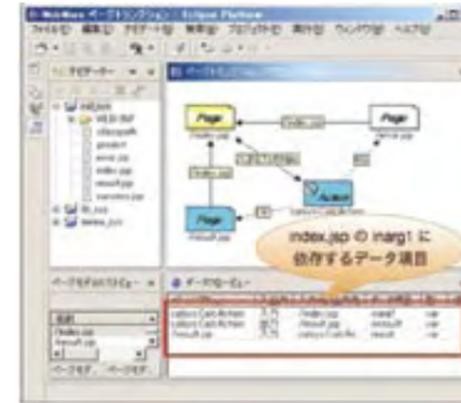


図 2 依存するデータ項目の追跡機能

### 4. WebWare の作成支援

WebWare のクライアントサイドで表示される情報の多くは, サーバサイドで管理されているデータを起源としている. 一方, 表示レイアウトに関する情報は一般にスタイルシートと呼ばれ, 標準的なスタイルシート記述として CSS が用いられる. このため, WebWare の作成支援にはデータベースやスタイルシートとページ記述との連携に対する支援が必要である. そこで, データベースとの連携に対する作成支援としてデータベーススキーマの変更に伴う WebWare の保守支援ツールを開発し, スタイルシートとの連携に対する作成支援として HTML 記述と CSS の相互参照関係解析ツールを開発した.

データベーススキーマの変更に伴う WebWare の保守支援ツールの開発では, データベースを利用する WebWare として PHP で記述されたものを対象に, データベースを変更した際に必要となる修正作業を支援するツールを, Web アプリケーション開発環境の一つである DreamWeaver のプラグインとして作成した. このプラグインでは, データベーススキーマの変更に関する情報を与えると, PHP 中の SQL 文などを解析し, 修正の必要な場所を提示するとともに, 可能であれば PHP 記述の変更を行う.

HTML 記述と CSS の相互参照関係解析ツールの開発では, HTML 要素と CSS セレクタの対応を, プロパティの重ね合せや適用優先順位を考慮して出力するツールを作成した. CSS の利用は複数のページ間におけるレイアウトの統一などの利点があるが, HTML 要素と CSS セレクタの対応関係が複雑になり WebWare の保守が困難になる. 本ツールを用いることにより, ある HTML 要素のスタイルを宣言している CSS セレクタ群の取得や, CSS セレクタの変更による HTML 記述への影響範囲を求めることができる (図 3).



図 3 HTML-CSS 対応解析ツールの出力例

### 5. おわりに

本研究では, 中間目標として以下を設定していた.

\* テスト支援:

WebWare リポジトリを利用した WebWare テストツールを作成し, WebWare テストに必要な工数をこれまでの 1/5 に削減する.

\* 解析技術:

WebWare を構成する各プログラムに対する解析器とリポジトリを完成する.

\* 作成支援:

WebWare リポジトリに格納された開発リソースからエンドユーザ情報を識別し, これらの配置および構造化を支援するためのツール群を作成する. また, 表現構成主導の構造化文書編集ツールを開発する.

今年度までに, テスト支援として (1) WebWare の整合性検査ツール, (2) WebWare 解析ツール, (3) JSP 単体テスト支援ツールを開発した. 整合性検査ツールによりバグの検出がテスト実行前に可能になりテスト総数を削減できる. 解析ツールによりバグの原因の特定が容易になりデバッグコストを削減できる. 単体テスト支援ツールによりテストの設計, 実行にかかるコストを削減し, 回帰テストも容易に行える. これらにより WebWare テストに必要な工数を 1/5 に削減できた. また, 各プログラムに対する解析技術として (4) HTML 解析器, (5) JSP 解析器 (6) JavaScript 解析器を開発し, WebWare 全体を表すモデルとして (7) ページ遷移モデル, (8) データフローモデルを開発した. さらに, 作成支援として (9) JavaScript に対応した Web サイト解析ツールと (10) HTML 記述と CSS の相互参照関係解析ツールにより配置および構造化の支援を行い, 構造化文書編集ツールとして (11) スタイル変換技術に基づく WebWare フレームワークと (12) WebWare 保守支援ツールを開発した. これらにより, 中間目標を達成することができた. 本研究の成果の一部は本プロジェクトの Web サイトおよび Sapid プロジェクトの Web サイト (<http://www.sapid.org>) から入手可能である. また, 協力企業の製品にも組み込まれている.

## インターネット上の知識集約を可能にするプラットフォーム構築

早稲田大学 理工学術院 村岡洋一

http://www.muraoka.info.waseda.ac.jp/~katayamail/e-society/

### 1. はじめに

インターネット上の Web サーバから発信される情報量は、本プロジェクトにより収集した 120 億の Web ページをもとに推定すると、2006 年 3 月時点で 7618 万台の Web サーバから合計で 360 億ページに上ると推測される。これは、全世界の人口が約 65 億人であることから、一人当たり平均 5.5 ページのデータを発信していることに相当する。

このような膨大な Web 上には、人間が一生かかっても学ぶことのできない情報、知識、ノウハウが凝縮されていると言っても過言ではない。これまで、我々は、このような膨大な情報を利用するために、Google や MSN、そして Yahoo! に代表される検索エンジンを利用してきた。各種情報の検索はもちろんのこと、辞書の代わりとしての利用、適切な翻訳語を探すための利用など、その用途は広範囲に渡る。

こうした検索エンジンに対して、最近注目されているのが、分析エンジン (Analysis Engine) である。分析エンジンとは、Web ページを含む膨大な情報の中から有用な情報を見つけ出すことを目的としたエンジンである。このような分析エンジンとしては、1999 年から研究プロジェクトとして実施されている IBM アルマデン研究所の WebFountain が有名であるが、対象とする Web データの規模は 20 億ページに留まる。これに対し本プロジェクトでは、世界中の Web ページから 120 億ページ、データ容量にして約 240TB のデータを対象とすることにより、規模において世界一を達成すると共に、こうした膨大な情報からの知の創出を目指している。

### 2. 研究課題と進捗状況

本プロジェクトでは、商用・研究用を通じて世界最大となる 120 億の Web ページを対象に、(1) 平均して 1 ヶ月以内の新しいデータに更新することを可能とする WWW クローラーを開発すると共に、(2) 利用者の検索目的に応じて必要となる情報を抽出する知識フィルタリング技術の開発を目指している。3 年目である 2005 年度は、国内 5 箇所に設置した合計 70 台 (80CPU) の WWW クローラーを用い、2006 年 2 月までに 120 億 URL の Web ページ収集を行った。また、2004 年度までに収集した 50 億 URL 規模でのデータ解析を開始した。

以下に 2005 年度の研究課題と成果概要を示す。

#### 2.1. 世界最大の 120 億 URL の Web データ収集

2005 年度は WWW クローラーに使用する PC を 30 台 (30CPU) 追加し合計 70 台 (80CPU) とし、国内 5 箇所 (早稲田大学内 2 箇所、インターネットデータセンター、NTT 未来ねっと研究所、国立情報学研究所) から合計で 120 億 URL の

Web ページ収集を行った (図 1)。収集のための起点として設定した WWW サーバ数は約 4,000 万であり、これらの内、2,545 万台の WWW サーバの収集を完了した。表 1 に主要ドメイン別の収集ページ数を示す。

全世界の Web サーバ数は、英国 Netcraft 社の統計によると 2006 年 3 月時点で 7,618 万台である。本プロジェクトで収集した Web ページ数は、2,545 万台の Web サーバから 120 億ページであり、比例により全世界の Web ページを推計すると約 360 億ページとなる。

全世界の Web ページ総数については、NEC 北米研究所の主任研究員であった Lawrence らによる推計が有名である。1998 年に Science 誌に掲載された論文では WWW サーバ当たりの Web ページ数が 190 ページ、1999 年に Nature 誌に掲載された同氏の論文<sup>2</sup>では、186 ページとなっている。さらに、総務省情報通信研究所が 2004 年 2 月に実施した WWW コンテンツ統計調査においても 1 サーバ当たりの平均 Web ページ数は 202 ページとなっていることから、これまで WWW サーバから平均 200 ページが発信されているとされてきた (図 1)。

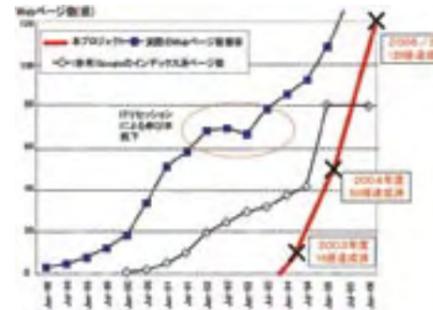


図 1 Web ページ収集経過

| ドメイン | 収集 Web ページ数    | 割合   |
|------|----------------|------|
| com  | 6,348,073,486  | 52%  |
| de   | 684,883,968    | 6%   |
| org  | 659,050,268    | 5%   |
| net  | 628,534,193    | 5%   |
| jp   | 490,344,079    | 4%   |
| uk   | 316,119,335    | 3%   |
| その他  | 2,876,677,991  | 25%  |
| 合計   | 12,003,683,520 | 100% |

表 1 収集 Web ページの内訳 (2006/2/22 時点)

しかし、本プロジェクトによる 120 億ページ収集によって、「近年では 1 サーバ当たりの Web ページ数が大きく増大し 472 ページ程度になっている」ことが判明した。この主な原因は、CGI 等によって動的に生成される「動的ページ」の急増である。

#### 2.2. 50 億のデータを対象とした Web マイニング

2004 年度までに収集済みの 50 億 URL の Web データを対象に、Web データの解析を行った。図 2 は、Web ページのリンク構造を模式化したものである。CORE は最大の強連結部分、すなわちあるページから別のあるページを経由して元のページに戻ることのできる「相互にリンクしているページ群」を示す。一方、IN は CORE に対するリンクのみを持つページ群、OUT は CORE からリンクを持つページ群を示す。1998 年の世界的な調査では、CORE が 28%、IN と OUT が各々 21% となっていたが、本プロジェクトでの調査により、近年、「相互にリンクし合う CORE 部分が大きくなり世界の Web 構造が密になっている」ことが判明した。

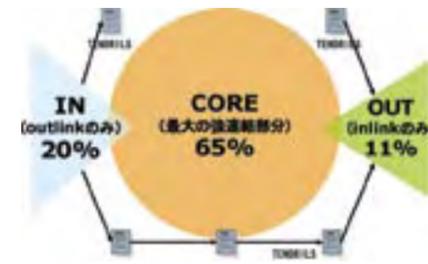


図 2 Web ページのリンク構造

さらに、本プロジェクトでは、収集した Web データからの有用な情報を得るための試みとして、国内最大となる 9,600 万件のブログの検索を実現する「ブログ検索システム」(図 3) や Web 上に掲載された製品採用情報を自動解析し現在どのような製品がどこで採用されているかを表示する「日本全国 IT 導入事例分析可視化システム」(図 4) を試作した。

また、要素技術として、Web ページを自動的に分類する「Web コミュニティ自動抽出の精度向上」(表 2) や Web ページ中の感情表現を自動抽出するための新しい手法として「アイテムの距離と属性を制約としたシーケンシャルパターンマイニング手法」(表 3) の提案を行った。

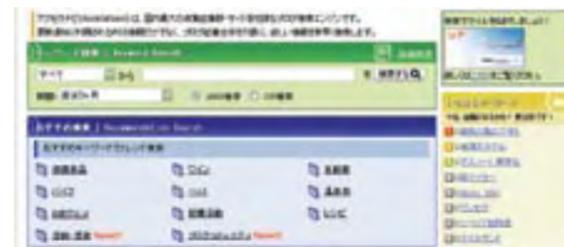


図 3 ブログ検索システム (アクセラテクノロジーとの連携)



図 4 日本全国 IT 導入事例分析可視化システム

| シードページ        | 抽出 URL             |
|---------------|--------------------|
| bbs.tip.ne.jp | www.uni.or.jp      |
|               | www.c-fm.co.jp     |
|               | www.sanwa-co.co.jp |
|               | www.mm-inoue.co.jp |
|               | 以下省略               |

表 2 コミュニティ自動抽出例 (山口県下関市周辺の企業)

| 頻度  | 名詞  | 形容詞 / 形容動詞 | 評価 | 頻度  | 名詞    | 形容詞 / 形容動詞 | 評価 |
|-----|-----|------------|----|-----|-------|------------|----|
| 241 | 内容  | 濃い         | +  | 117 | 絵     | 綺麗         | +  |
| 217 | 本   | 良い         | +  | 99  | レベル   | 高い         | +  |
| 206 | 奥   | 深い         | +  | 92  | 情報    | 多い         | +  |
| 186 | 本   | 好き         | +  | 91  | カッコ   | いい         | +  |
| 170 | 本   | 面白い        | +  | 90  | 値段    | 安い         | +  |
| 125 | テンポ | 良い         | +  | 85  | ストーリー | 面白い        | +  |
| 119 | 内容  | 深い         | +  | 83  | 絵     | かわいい       | +  |

表 3 プラスイメージで利用される単語ペア

### 3. 実施体制

本プロジェクトは、プロジェクトリーダー村岡洋一 (早大・理工)、サブリーダー山名早人 (早大・理工) のもと、富士通株式会社、アクセラテクノロジー株式会社、NTT 未来ねっと研究所、国立情報学研究所等と共同で実施し、技術移転を進めている。

2006 年度以降は、企業との連携をさらに密にすると共に、これまでの試作アプリケーションや要素技術の実用化を目指し研究開発を推進する。

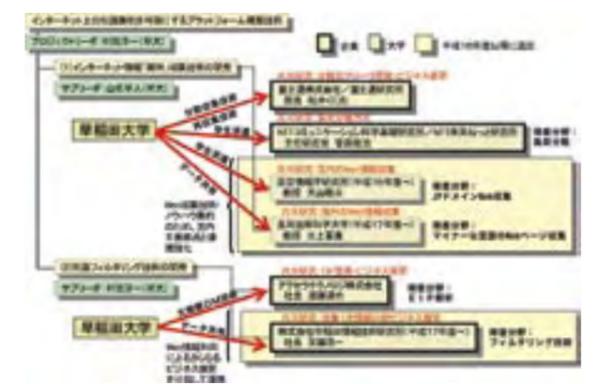


図 5 実施体制

## 先進的なストレージ技術

東京大学 生産技術研究所 戦略情報融合国際研究センター 喜連川 優  
 協力企業 日立製作所, NTT プラットフォーム研究所  
<http://www.tkl.iis.u-tokyo.ac.jp/project/e-society/>

### 1. はじめに

情報通信技術の革新が進む中、情報システムにおけるデータ格納庫であるストレージの役割はサーバを凌駕するに至りつつあり、ストレージには一層の高信頼化、高性能化、管理容易化が求められている。とりわけ9.11テロ以降、災害時にもデータを失うことなく業務を継続させることができるディザスタリカバリ (DR) 技術が注目を集めており、また、爆発的に増加し続けるデータに対する超高速アクセス技術、管理容易化技術の実現は必須の課題である。

本研究プロジェクトにおいては、IT社会を支える安心・快適・便利なストレージの実現のため、ストレージをサーバ上のアプリケーションと融合させる次世代ストレージ技術：ストレージフュージョン (Storage Fusion) の確立を目指す。具体的には、図1に示すように高度ディザスタリカバリ機構、ストレージ超高速アクセス機構、ストレージ管理コスト低減機構の開発に焦点を絞って研究を進めている。

高度ディザスタリカバリ機構では、災害時のデータ欠損により人手による整合性確認処理が必要となることから迅速な復旧が困難となっている問題を解消すべく、ストレージ自身が常に一貫性を維持する機構を具備するトランザクショナルストレージを提案する。即ち、本技術により、整合性確認処理を自動化し、

無データ欠損を補償する新しいディザスタリカバリ技術を開発し、災害時の迅速な業務再開を目指す。また、多大なコスト負担を要するDRシステムの適用範囲拡大のために、低コスト化技術の確立も目指す。

ストレージ超高速アクセス機構では、データベース管理システム (DBMS) が作成する問合せ処理実行計画に相当するクエリプラン等、サーバで実行される処理に関するアプリケーションの情報と、ストレージ内部の動作状況の情報を融合する (フュージョン) ことにより、ストレージキャッシュへのデータの先読み等、ストレージシステムの内部制御を最適化し、アプリケーションの性能を飛躍的に向上させる新しいストレージ基盤技術を確立する。

又、ストレージシステム管理費に係わる人件費の割合はストレージ全コストの7割を上廻るとされ、特に有能な管理者に限られている現状では、ストレージの管理は一層深刻な問題となりつつある。ストレージ管理コスト低減機構に関する研究においては、セルフオーガニゼーションやセルフモニタリング等のデータベースシステムの自立管理技術を確立するとともに、ストレージ管理を大幅に容易化し、運用管理コストの低減を目指す。

### 2. 高度ディザスタリカバリ機構

災害時にもデータ欠損がないDR方式として、ハイブリッド転送方式とログのみ同期転送方式を創案、実装し、トランザクショナルストレージ技術を確立した。

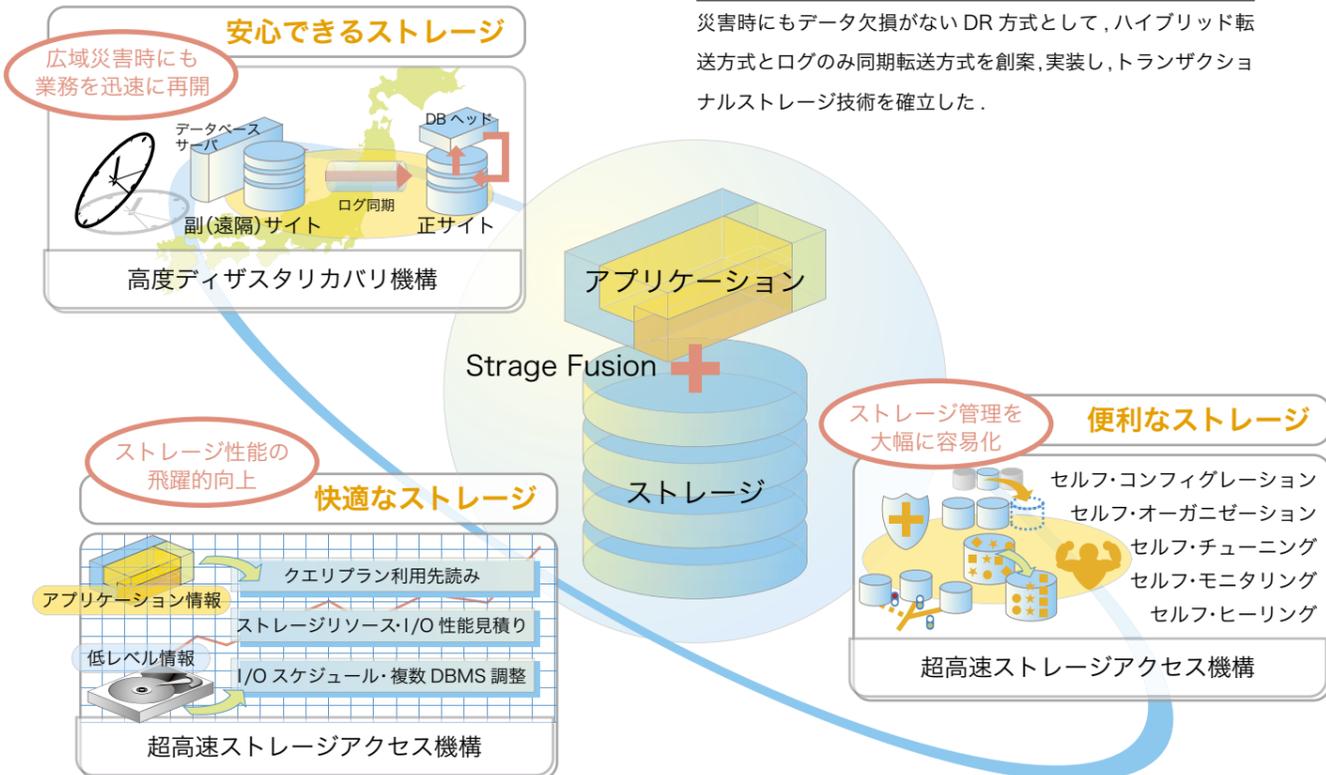


図1 ストレージフュージョン：IT社会を支える安心・快適・便利なストレージ

### (1) ハイブリッド転送方式 (製品化)

ログを同期リモートコピーにより、DBを非同期リモートコピーにより転送するハイブリッド転送方式を創案し、回線シミュレータを用いた評価の結果、広域災害を考慮した構成においてオンライン性能確保が可能であることを確認した。

### (2) ログのみ同期転送方式

平常時の回線コストを削減すべく、DB部の転送を不要とするログのみ同期転送方式 (図2左) を創案し、同様に広域災害を想定した環境下で評価を行い、非同期方式に比べてわずかな性能劣化に抑えられることを確認した (図2右)。また、当該方式に関して、安価なDBヘッドでログ適用を可能とするためのログ適用高度化方式を考案した。

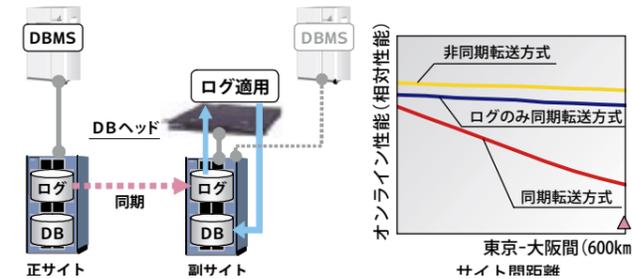


図2: システム構成 (左) と正サイトオンライン性能評価結果 (右)

### 3. ストレージ超高速アクセス機構

ストレージ超高速アクセス機構に関する研究では、アプリケーションの情報を基にしたストレージ内部動作の最適化によるストレージ性能の向上を目指し、クエリプラン利用先読み技術を開発した。当該技術は、DBMSが作成するクエリプランをストレージに融合し、アクセスされるデータを予測・先読みすることにより従来では得られない高いI/O性能を得るものである。本技術のプロトタイプを、商用ストレージを用いて構築し (図3)、TPC-H Q8ベンチマークを用いた評価により、キャッシュヒット率が向上し、大幅に高速化出来る事を確認した (図4)。

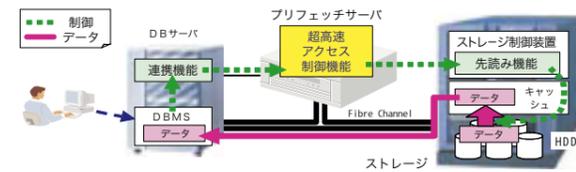


図3 クエリプラン利用先読み技術プロトタイプ構成概略

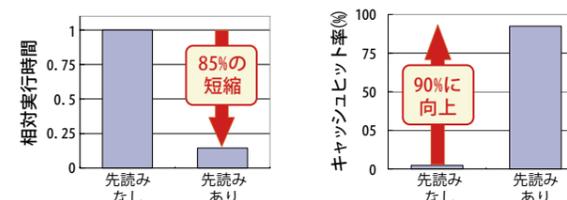


図4 プロトタイプの評価結果 (TPC-H Q8 Benchmark) 相対処理時間 (左) とキャッシュヒット率 (右)

### 4. ストレージ管理コスト低減機構

ストレージ管理コスト低減機構に関する研究では、ストレージの管理を大幅に容易化すべく、以下の技術を開発した。

#### (1) 自立再編成ストレージ

自立的に構造劣化を検出し、人手を介さず再編成を高速に実施する自立再編成ストレージを開発した。(図5)

#### (2) システムモニタリング技術

個別に運用管理されてきたストレージとDBMSの統合的な管理を目的として、ストレージとDBMSの性能情報の統合的表示技術を開発し、その一部製品化し、ストレージ資源コンシャスなDBMSの内部状態可視化を実現した (図6左)。

#### (3) ボトルネック検出技術

ストレージとDBMSの性能チューニングにおいて最も難しい課題はロック競合である場合が多い。当該問題の診断コストの低減を目的として、DBMSの排他待ち診断支援システムプロトタイプを開発した。(図6右)。

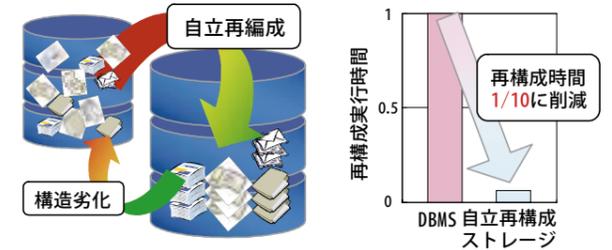


図5 自立再編成ストレージ



図6 ストレージリソースのDBMSによる使用内訳 (左) とDBMSの排他待ちの診断 (右)

### 5. まとめ

ストレージとデータベースアプリケーションを融合 (ストレージフュージョン) することにより高信頼かつ高性能なデータ格納プラットフォームを実現する為の新技術を開発した。

高度DR機構においては、トランザクショナルストレージという基本的アイデアを基に、データ保護とオンライン性能確保を両立するログのみ同期転送技術を開発した。ストレージ超高速アクセス機構においては、クエリプラン利用先読み技術プロトタイプを開発するとともに実機を用いて評価し、大幅な性能向上を確認した。ストレージ管理コスト低減機構においては、再編成機構をストレージに融合させた自立再編成ストレージを開発した。更に、DBMSとストレージの性能情報の統合的表示ならびに、データベースシステムの排他待ち診断支援に関する新技術を開発した。

今後はストレージフュージョンの更なる深化を目指す予定である。

## 先進的な Web 解析技術

東京大学 生産技術研究所 戦略情報融合国際研究センター 喜連川 優、協力企業 三菱電機  
<http://www.tkl.iis.u-tokyo.ac.jp/project/e-society/>

### 1. 目的

Web 上では企業や省庁、個人による情報発信が刻々と行われており、近年では実世界の様々な事象が網羅的かつ即時的に Web に反映されるようになってきている。実世界と Web の間には一種の転写構造 (図 1) が形成されつつあり、サイバー社会の構造を把握し、その変化を追跡することは、実社会に起こる事象の背景や兆兆を探る上で極めて有効と考えられる。本プロジェクトでは、Web 上の社会知の高効率で高度な利用

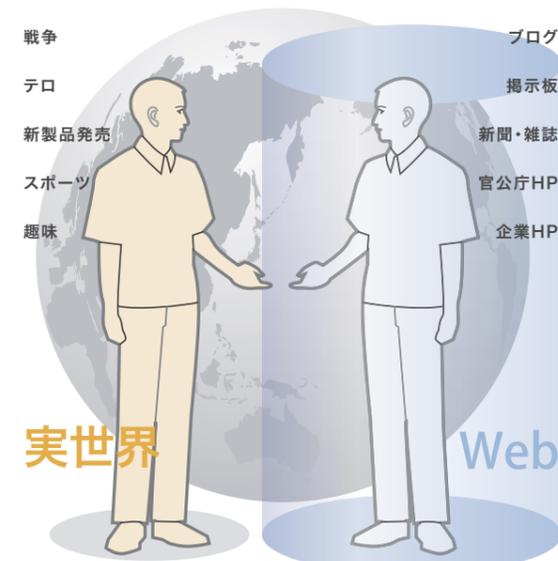


図 1 実世界と Web の転写構造

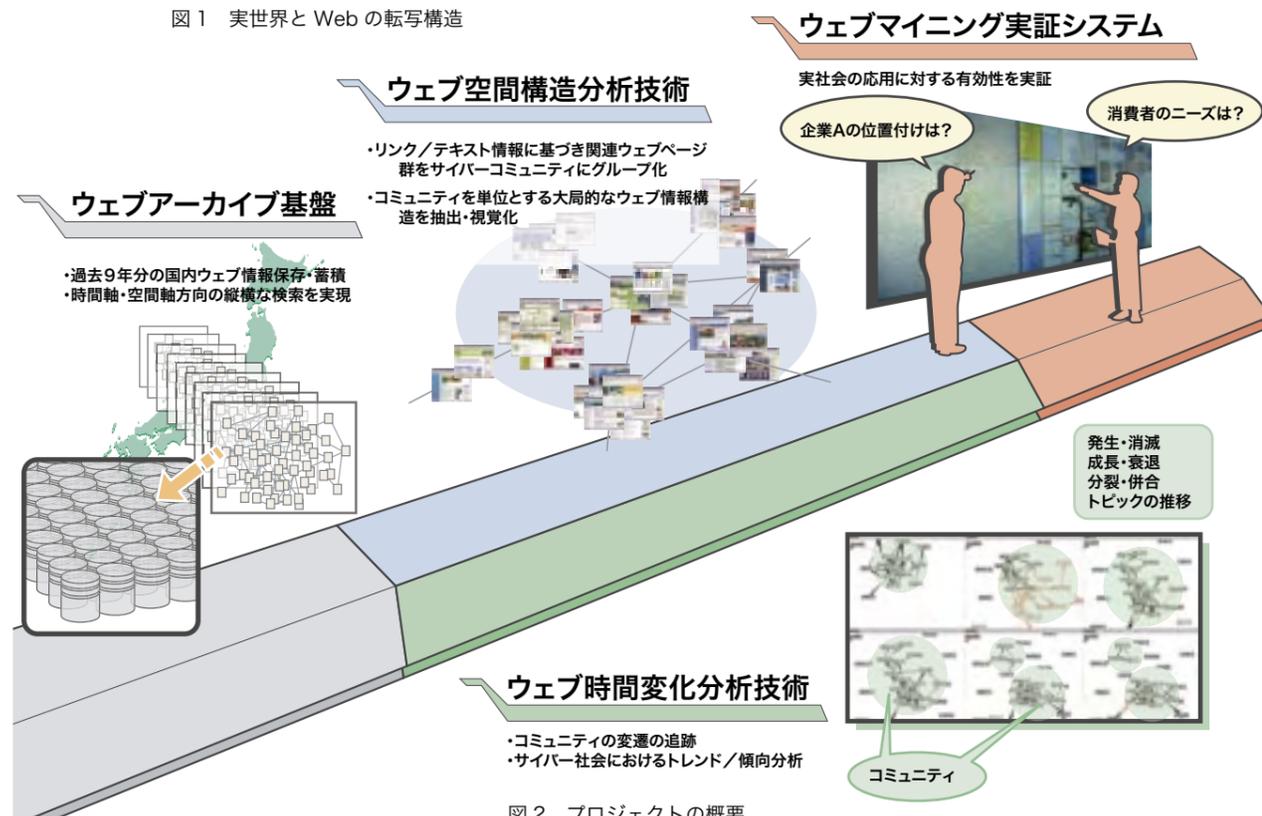


図 2 プロジェクトの概要

を可能とすべく、現行のサーチエンジンとは全く異なる新しい Web 解析システムの創出を目的とし、Socio-Sense なるシステムの開発を進めている (図 2)。以下にその開発項目を示す。

- 1 過去 9 年分 (プロジェクト完了時) の日本の Web 情報を蓄積し、縦横に検索可能とした、他に類例のない Web アーカイブ基盤の構築
  - 2 リンク解析およびテキスト解析 (自然言語処理) を核に、関連 Web ページ群 (コミュニティ) を単位として大局的な Web 情報構造の俯瞰・対話的分析を可能とする Web 空間構造分析技術の開発
  - 3 コミュニティの発生・成長・衰退・消滅や分裂・併合・トピック推移などの時系列変化を捉え、トレンドや傾向の分析を可能とする Web 時間変化分析技術の開発
- さらに、これらの要素技術を統合して Web マイニング実証システムを構築し、例えば、企業活動における消費者の嗜好・ライフスタイルのタイムリな把握とそれに基づく意思決定支援など、実社会の応用に対する有効性を実証する。

### 2. これまでの成果

#### 2.1. Web アーカイブ基盤

これまでに、プロジェクト開始時点の保有データと併せ、約 7

年分の国内 Web を中心とするアーカイブを構築してきた。異なる時期に対応する複数のスナップショットは統合して管理し、時間属性や全文に対する検索機構を介して縦横にアクセス可能とした。これにより、URL 毎の内容の変遷や特定キーワードの出現頻度の推移などを容易に取得できるようになった。Web アーカイブの更新は、スナップショット単位での追加から Web ページ毎に随時追加する方式へと移行し、Web の変化により柔軟に対応できるようにしている。また、Web アーカイブ基盤は Web 時空間解析のプラットフォームとしての役割も併せ持ち、並列スキャン機構によりスナップショット全体の高速な解析処理を支援している。

#### 2.2. Web 時空間分析

互いに関連する Web ページは直接あるいはリンク集などを介して間接的にリンクされる傾向にあり、この特徴を利用して Web 空間から稠密なリンク構造を抽出すると Web コミュニティと呼ばれる関連ページの集合が得られる。Web の空間構造分析では、Web のスナップショットからほぼ全てのコミュニティを網羅的に抽出し、それらを互いの関連度に従って配置した Web 空間の俯瞰図 (図 3) を作成した。また、利用者の分析意図に応じたコミュニティの抽出を目指して、リンク解析と自然言語解析を併用した解析方式の検討および評価実験を行い、コミュニティ抽出精度の向上を確認した。

Web の時系列分析では、コミュニティの時系列的な発展過程を追跡する手法を開発し、視覚的に変化を確認する手法を開発した (図 4)。本手法は、Web におけるトピックの推移、新たな情報の発生、Web 上の社会的な現象などの調査に利用することができる。

さらに、空間構造と時系列変化を統合した可視化手法を用いた時空間分析手法を開発し、話題の黎明期からの詳細な空間構造の変化、および成長ステージの把握を可能にした (図 5)。

#### 2.3. Web マイニング実証システム

Web アーカイブ基盤と Web 時空間分析手法を統合し、実証システムのプロトタイプ開発を行った (図 6)。本プロトタイプでは、Web の構造俯瞰図とその発展過程追跡、および時空間分析の結果を可視化して大規模高精細ディスプレイウォールに投影しており、様々な対話的分析が可能となっている。

### 3. 今後の展望

今後は、リンク解析とテキスト解析技術の密な融合による、より精緻な分析を実施すると共に、実応用からのフィードバックに基づいて実証システムの高度化を進めて行く予定である。



図 3 Web 空間の構造俯瞰



図 4 Web の時系列分析



図 5 Web の時空間分析 (i-mode 用検索エンジンの変遷)



図 6 大規模高精細画面を用いた実証システムプロトタイプ

## ユーザ負担のない話者・環境適応性を実現する自然な音声対話処理技術

奈良先端科学技術大学院大学 情報科学研究科 鹿野清宏

研究分担者：京都大学 河原達也，名古屋大学 武田一哉，和歌山大学 河原英紀，奈良先端大 猿渡洋，名古屋工業大学 徳田恵一，

立命館大学 西浦敬信 協力企業：松下電器，旭化成，日立製作所，松下電工，ASTEM，オムロン

<http://cif.iis.u-tokyo.ac.jp/e-society/database/index.html>

### 1. プロジェクトの目標

携帯電話，携帯端末，PCの入出力，カーナビ，家電制御，秘書ロボットなどを，誰でも容易に利用できることが望まれる。これらを実現するには，人と機械との自然な対話として，音声認識・合成技術が有望である。本格的に利用されるためには，頑健かつ高精度の音声認識基盤ソフトウェアを開発して，かつ廉価に利用できるようにすることが重要である。この音声認識・合成技術の基盤ソフトウェアの普及により，誰でも気軽に，情報機器の利用ができるようになり，デジタルディバイドの社会問題の軽減，IT市場の活性化につながる。

プロジェクトでは，大語彙連続音声認識プログラム，話者環境適応プログラム，ハンズフリー音声認識プログラム，ハンズフリー音声収録 DSP，大語彙連続音声認識プログラムのマイコンへの実装を行う。音声合成では，多様な音声合成プログラムを作成する。これらのプログラムは，単に開発するだけでなく，実環境での応用システムでの実証試験による評価も行って改善をはかる。さらに，音声認識技術の利用法のノウハウも蓄える。当初の計画に付け加えて，静かな音声メディアとして発見した「非可聴つぶやき声 (NAM:Non-Audible Murmur) の音声認識・合成の研究開発も進める。

### 2. プロジェクトの概要

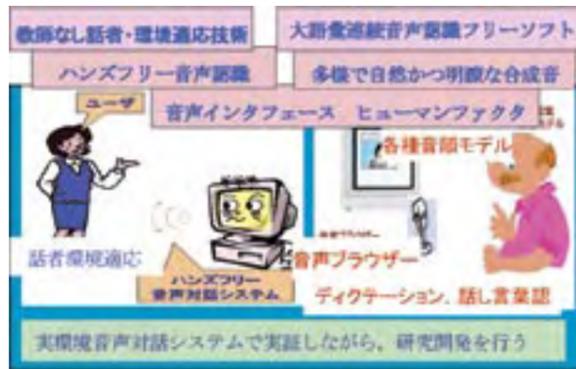
音声認識を本格的な商用化につなげるには，頑健な音声認識システムとして，以下の技術を研究開発することが必要である。

- (i) **利用環境およびユーザに対する負担をかけない適応技術**，
- (ii) **高精度連続音声認識プログラム**および音声認識システム研究開発ワークベンチ，
- (iii) **マイクを意識しない自然なハンズフリー音声認識技術**が必要となる。これらのソフトウェアを研究開発し，廉価に誰もが利用できるプログラム，あるいは DSP/マイコンとして提供する。さらに，開発したソフトウェアを用いて，(iv) 人と機械の**音声対話の実証実験**を行い，ソフトウェアの評価を行うとともに，利用に関するノウハウを蓄積する。音声合成では，(v) **多様な声質**の実現が重要であり，高精度音声分析合成系 STRAIGHT を用いた声質変換プログラムを開発する。上記に加えて，非可聴つぶやき声の音声認識（無音声認識）を

(i) で，変換合成（無音声電話）を (v) で研究開発を行う。さらに，(ii) では，新しく開発された音源分離 (BSS:Blind Source Separation) アルゴリズムによるハンズフリー音声認識も行う。(iii) では，マイコンで動作する大語彙連続音声認識プログラムの実行速度を実時間処理にまだ高める。

本プロジェクトの研究開発の概要を図に示す。さらに，研究開

発項目ごとの成果の状況および予定を表にまとめておく。



人にやさしい自然な音声インタフェース

研究開発の成果の状況および予定

| 研究開発項目           | 成果の状況および予定  |
|------------------|---|
| ユーザ負担のない話者・環境適応  | 教師なしオンライン話者適応 (H18.3)<br>非可聴つぶやき声 (NAM) 認識 (H19.3)  |
| 大語彙連続音声認識ソフトウェア  | 話し言葉認識モデルとプログラム (H18.3)<br>大語彙連続音声認識プログラム Julius(H18.3)<br>マイコン SH-4A Julius(実時間動作)(H19.3)        |
| ハンズフリー音声認識       | ハンズフリー音声収録 DSP(H19.3)<br>BSS音源分離オンラインプログラム (H18.3)  |
| 実環境音声対話システムの構築   | たけまるくん音声情報案内構築キット (H16.7)<br>自動車内音声認識プログラムキット (H19.3)   |
| 多様な声質の音声合成ソフトウェア | オンライン音声変換プログラム (H18.3)<br>高精度音声変換プログラム (H19.3)<br>HMMベース音声規則合成システム (H15.10)<br>無音声電話プログラム (H19.3) |

(青字で書かれた成果は，当初計画になく，追加した成果予定)

### 3. H17年度の進捗

非可聴つぶやき声の音声認識で電子情報通信学会の最優秀論文賞「猪瀬賞」やハンズフリーの BSS 音源分離で音響学会の「独創研究奨励賞 板倉記念」を受賞した。また，非可聴つぶやき声や BSS 音源分離技術は，注目を集めて NHK や新聞などでも報道された。研究項目ごとに H17 年度の進捗をまとめておく。

- (i) 教師なし話者適応プログラム：話者適応アルゴリズムの処理速度と精度を向上させ，数秒で教師なし話者適応が実行できるオンライン話者適応システムを構築した。音声情報案内システム

内システム「たけまるくん」の 2 年間分の 30 万発話の書き起こしを終了し，各種音韻モデルの精度の向上を確認した。この音声データベースを用いて，新たに考案した膨大な音声データベースからタスクに適した音韻モデルを自動構築するアルゴリズムの有効性を実証した。非可聴つぶやき声 (NAM:Non-Audible Murmur) による音声認識（無音声認識）の研究，および，声を出さない電話（無音声電話）の研究も進展した。

- (ii) 大語彙連続音声認識ソフトウェア：種々の不特定話者音韻モデルを音声対話システムにおける音声認識性能で評価した。Web から収集した関連テキストから話し言の言語モデルを構築する汎用的な手法を開発した。音声対話による情報検索システムを指向して，重要語に重きを置いた認識手法を研究開発した。音声認識プログラムの SH-4 マイコンへの実装を進め，高精度版の動作を確認した。京都市バス運行情報案内システムを引き続き運用し，音声対話データの収集を行った。

- (iii) 空間スペクトル演算アレー SSA の素子誤差の適応補正，残響マッチド音響モデル，実時間かつ低演算量の方位推定法について検討し，音声認識性能の改善を達成した。PC 上での実時間 SSA 処理系の実装にもつづき，SSA を DSP 上に実装して，実時間処理の可能性を確認した。また，音源間の独立性に基づいて音声抽出するブラインド音源分離に関して検討を行い，音質劣化が少ない高精度分離アルゴリズムの DSP 実装および評価を行った。

- (iv) 音声認識のフィールドテスト：フィールドテストにより収集された音声データを用いて，雑音環境の変動に対処するため，SNR に依存した音響モデルを構築するとともに，それらを選択的に利用する手法を開発した。インターネット楽曲ダウンロードを音声対話により行うシステムのフィールドテストを実施した。楽曲の購入，認識辞書のカスタマイズ，入力音声のアップロードの機能を実装し，不特定多数の利用者の入力音声と利用ログを収集可能な音楽検索用フィールドシステムの運用実験を開始した。対話システム構築のガイドラインをマニュアル化した。実環境での音声対話の実証実験として，駅に音声情報案内システムとロボット案内システムを設置した。

- (v) 多様な音声合成プログラム：STRAIGHT 分析合成システムの C 言語による実装を行い，リアルタイム動作可能なプログラムを開発した。声優や一般人による音声素材データベースを拡充した。HMM に基づく合成音声用の韻律制御システムを中心とする音声合成システムは，国際会議併設の音声合成システムコンテストにおいて，最優秀のシステムとしての評価を得た。その他，非可聴つぶやき声 NAM (Non Audible Murmur) から通常の音声やささやき声への変換に応用する技術を開発した。

## 教師なし話者・環境適応技術

奈良先端科学技術大学院大学 情報科学研究科 鹿野清宏

<http://spalab.naist.jp/>

### 1. ねらい

人と機械との自然な対話を実現するためには，高精度な音声認識技術が必須である。とくに，利用環境とユーザへの適応技術が必要となり，かつ，ユーザに負担をかけない適応技術が望まれる。このような環境と話者への適応を実現のために，教師なし話者・環境適応技術に関して研究開発を行う。

### 2. 研究の概要

雑音環境下で，任意の 1 文発声に基づくユーザに負担をかけない教師なし話者適応アルゴリズムの研究開発を進める。具体的には，雑音に頑健な音声認識アルゴリズムと HMM(隠れマルコフモデル) 十分統計量に基づく教師なし話者適応アルゴリズムの研究開発を行う。最終的に，オンラインで動作する話者・環境適応プログラムを完成させる。さらに，新しい静かな音声メディアとして，つぶやき声 (NAM) の認識（無音声認識），声を出さない電話（無音声電話）の研究を行う。

### 3. H17 年度の進捗

非可聴つぶやき声の認識が高く評価され，電子情報通信学会の最優秀論文賞「猪瀬賞」を受賞した。平成 17 年度における研究進捗を以下に示す。

- (i) 話者適応アルゴリズムの精度を向上させ，かつ，計算量も削減して，数秒で教師なし話者適応が実行できるオンライン話者適応システムを構築した。
- (ii) 音声情報案内システム「たけまるくん」の運用を継続するとともに，2 年間分の 30 万発話の書き起こしを終了し，各種音韻モデルの精度の向上を確認した。
- (iii) 音韻モデルの構築のコストを減らすことを目指して，膨大な既存の音声データベースからタスクに適した音韻モデルを自動構築するアルゴリズムを考案した。高齢者および幼児の音韻モデルの構築を行い，有効性を実証した。
- (iv) 非可聴つぶやき声 (NAM:Non-Audible Murmur) は，話し手の近くでも聞こえない声である。NAM による音声認識（無音声認識）の研究，および，声を出さない電話（無音声電話）の研究が進展した。



音声情報案内システム「たけまるくん」

## 大語彙連続音声認識プログラム

京都大学 学術情報メディアセンター 河原達也  
<http://www.ar.media.kyoto-u.ac.jp/>

### 1. ねらい

ユーザに負担のない自然な音声対話を実現するには、音声認識システムが、できるだけ広範な話者層や言い回し、特に話し言葉に対応できる必要がある。このような大語彙連続音声認識を行うオープンソースのプログラムを開発する。また、組み込み機器にも利用できるような、マイコンへの実装も行う。

### 2. 研究の概要

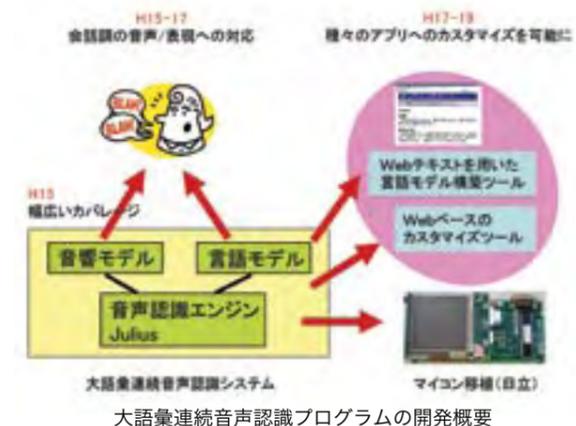
ディクテーションシステムは読上げ音声を対象としていたが、音声対話を指向して発展させる。音韻モデルも、多様な話者、対話調の音声に対応できるようにする。言語・発音モデルについても、話し言葉に対応できるようにする。さらに、音声認識エンジン Julius についても音声対話システム向けの機能強化を行うとともに、マイコンに移植できるような効率化を図る。

### 3. H17 年度の進捗

平成 17 年度における研究進捗を以下にまとめる。

- (i) これまで構築してきた種々の不特定話者音韻モデルを音声対話システムにおける音声認識性能で評価した。
- (ii) 任意の情報検索タスクに対して、Web から収集した関連テキストから話し言葉調の文章を選択することにより、効率的に言語モデルを構築する汎用的な手法を開発した。
- (iii) 音声対話による情報検索システムを指向して、重要語に重きを置いた認識手法を研究開発した。
- (iv) 音声認識エンジン Julius の機能・性能強化を行うとともに、Web ベースのカスタマイズツールの運用を開始した。
- (v) 音声認識プログラムの SH-4 マイコンへの実装を進め、高精度版の動作を確認した。
- (vi) 京都市バス運行情報案内システムを引き続き運用し、音声対話データの収集を行った。

上記に加えて、ソフトウェア普及のための講習会を行った。



## ハンズフリー音声認識

奈良先端科学技術大学院大学 情報科学研究科 猿渡洋  
<http://spalab.naist.jp/>

### 1. ねらい

ユーザに負担をかけない自然な音声入力系として、ハンズフリー音声認識システムを構築する。特に、マイクロフォンアレー (図参照) による音声収録技術に着目し、音声認識性能の向上を目指す。また、コンパクトかつ廉価なマイクロフォンアレーアルゴリズムの開発を行う。

### 2. 研究の概要

ユーザからの距離 1m 以下で高性能に動作するハンズフリー音声認識システムを、8 チャンネル以下のマイクロフォンアレーを用いて開発する。認識性能は、1m 離れた音声入力で、従来の接話マイクとほぼ同等の認識性能を目指す。さらに、マイクロフォンアレーのコストを 20 分の 1 以下にするため、ハンズフリー音声収録用 DSP を開発する。

### 3. H17 年度の進捗

平成 17 年度における研究進捗を以下に示す。

- (i) 空間スペクトル演算アレー SSA(Spatial Subtraction Array) に関して、素子誤差の適応補正、残響マッチド音響モデルについて検討し、音声認識性能の改善を達成した。
- (ii) 実時間かつ低演算量の方位推定法の検討を行い音源定位の評価を行った。また、方位情報と発話検出を組み合わせた頑健な発話区間検出法を開発した。
- (iii) PC 上に実時間 SSA 処理系を実装し、ハンズフリー音声対話デモシステムを構築した。想定される規模 (4~8 素子アレー) において実時間処理が可能であることを確認できた。さらに、DSP 上に実装して、実時間処理の可能性が確認できた。
- (iv) 音源間の独立性に基づいて音声を抽出するブラインド音源分離に関して検討を行い、音質劣化が少ない高精度分離アルゴリズムの DSP 実装および評価を行った。この研究で、音響学会の「独創研究奨励賞 板倉記念」を受けた。



直線状マイクロフォンアレー

## 音声認識システムのフィールドテスト

名古屋大学 情報科学研究科 武田一哉  
<http://www.sp.m.is.nagoya-u.ac.jp/>

### 1. ねらい

ユーザにとっての音声対話システムの性能は、認識性能だけでなく、様々なヒューマンファクタに支配されている。開発したプログラムやモデルを実環境下で運用することで、開発成果の検証を行うとともに、運用結果からユーザ負担の少ない音声対話システムの設計指針を得る。

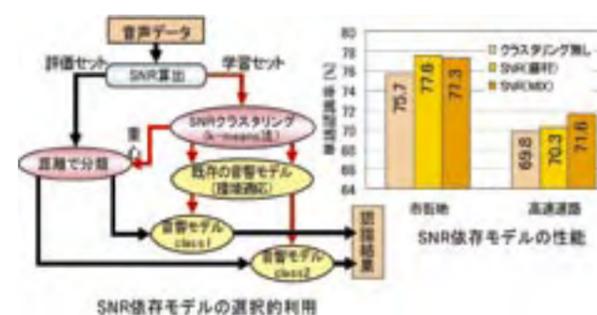
### 2. 研究の概要

開発した音声認識プログラム、各種モデルを用いて、実環境下での音声対話システムのフィールドテストを行い、開発成果の検証を行うとともに、多様なアプリケーションを効率的に作成する方法や、音声認識利用のヒューマンファクタに関するノウハウを蓄積する。

### 3. H17 年度の進捗

平成 17 年度における研究進捗を以下に示す。

- (i) フィールドテストにより収集された音声データを用いて、多様な環境に適合した音響・言語モデルを学習する方法を開発した。雑音環境の変動に対処するため、SNR に依存した音響モデルを構築するとともに、それらを選択的に利用する手法を開発した。音声区間検出を行うことなく、高精度に SNR を推定する方法を開発し、モデル選択手法と組み合わせ、走行中のバイクで収集された音声でその有効性を確認した。
- (ii) インターネット楽曲ダウンロードを音声対話により行うシステムのフィールドテストを実施した。前年度までに行ってきたフィールドテストシステムの機能を拡充し、楽曲の購入、認識辞書のカスタマイズ、入力音声のアップロードの機能を実装し、インターネットを介して不特定多数の利用者の入力音声と利用ログを収集可能な音楽検索用フィールドシステムの運用実験を開始した。
- (iii) 対話システム構築のガイドラインをマニュアル化した。



## 多様な音声合成プログラム

和歌山大学 システム工学部 河原英紀  
<http://www.wakayama-u.ac.jp/~kawahara/>

### 1. ねらい

本プロジェクトでは、機械と人間との対話を自然なものとするために、人間のように多様な声質を有する音声を合成することの出来るプログラムを開発する。

### 2. 研究の概要

本プロジェクトでは、我々の開発した高精度音声分析合成系 STRAIGHT を用い多様な声質の合成音声を作成する。この研究手段を用いて、まず、話し手による声質の違いや話し手の感情や話し方による声質の違いを取り出すためのデータベースを整備する。これらのデータベースの解析結果を利用して、様々な声質を合成音声に付与することの出来るプログラムを開発する。開発するプログラムは、高品質で精密な声質の付与が可能ではあるが非リアルタイムのものと、品質には制限があるがリアルタイムで動作するものの二系統とする。

### 3. H17 年度の進捗

本年度も研究計画に従い下記の項目について研究を推進した。

- (i) Matlab で実装されている STRAIGHT 分析合成システムのコードを整理し、C 言語による実装のためのインタフェース仕様を確定した。
- (ii) 上記で確定した仕様と整理されたコードに基づき、STRAIGHT の C 言語による実装を進めた。
- (iii) この C 言語により実装された STRAIGHT に基づき、リアルタイム動作可能なプログラムを開発し、動作を確認した。
- (iv) 平成 16 年度に作成した音声素材データベースを用いて、名古屋工業大学で開発された音声合成用システム用の合成データを作成した。
- (v) 声優や一般人による音声素材データベースを拡充し、音声多様性データベースを構築した。
- (vi) HMM に基づく合成音声用の韻律制御システムを中心とする音声合成システムを、フリーソフトとして配布した。また、このシステムに STRAIGHT の技術を取り入れることにより、企画した音声合成システムコンテストにおいて、最優秀のシステムとしての評価を得、知名度の向上と普及の促進を得た。
- (vii) これら当初計画に加えて、STRAIGHT に基づく音声変換法を、他人には聴こえないレベルのつぶやき声 NAM(Non Audible Murmur) から通常の音声やささやき声への変換に応用する技術を開発し拡張した。