

高信頼性組み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

組み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、高信頼性は組み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、組み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み込みソフトウェアの開発に投入し、組み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み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) 組み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) 組み込みシステム向け基盤ソフトウェア (早稲田大学)
- (3) 組み込み用実時間Java技術 (京都大学)

2. 組み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てこられ、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、組み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・検証しながら組み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。あわせてこの環境を踏まえた、設計検証のアプローチについての検討や実事例に基づく評価なども進めた。

3. 組み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プ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プライアンスには、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、これらを実現するためのミドルウェア、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. 組み込み用実時間Java技術

組み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、職人芸的技術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、信頼性の高い組み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。

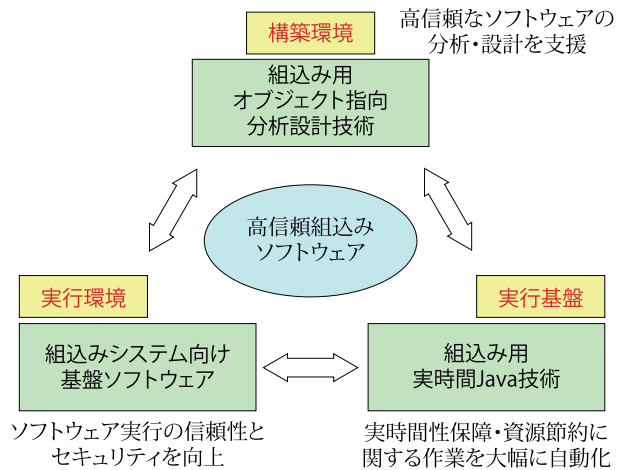


図1 プロジェクトの全体像



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

<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、さらに組込みソフトウェアの検証のために、周期イベントに基づいた並行タスクの振る舞いの検証法や、モデル検査技術のリアルタイム・スケジューリングへの応用について検討した。

2. 主要な成果

以下に本研究課題の主要な成果について報告する。

2.1. UML 検証ツールの開発と適用

UMLで組込みソフトウェアの設計を記述し、その検証を行うことを支援するツールを開発し、またいくつかの企業とその適用を進めてきた。MARTE (リアルタイム・組込みソフトウェア向けのプロファイル) が作られるなど、組込み分野でのUML適用の環境も整備されつつあり、実際の適用も徐々に広がってきている。本ツールはそうした技術動向との整合性を意識しつつ、企業での試行結果を踏まえたリファインを進めてきた。また同一モデルに対して様々な利用シナリオでの検証を行ったり、モデルの一部のみを改変した検証を行ったりすることを支援する機能など、実際の検証の施行を踏まえた各種支援機能の充実を行い、実用性を高めるとともに、典型的な適用アプローチについての整理・提言や、実事例による評価を進めてきた。ツールは複数の企業で試行され、設計段階での問題発見により、全体としての障害件数を減らすことができたという報告もなされている。

2.2. 組込みソフトウェアのモデル検査法

組込みシステムを対象とする自動検査の方法としてモデル検査法に期待が集まっている。しかし、不用意に使うと、状態爆発の問題にぶつかって検証ができないという問題点がある。一方、産業界が用いるUMLステートダイアグラムやリアルタイム・スケジューリングではその特徴を生かすことで、状態爆発の問題を避けられる可能性がある。本研究テーマでは、産業界において、既存のモデル検査ツールを組込みシステムの検証に適用する場合の問題点について、UMLステートダイアグラムのモデル検査、リアルタイムタスク・スケジューリング法、不安定な外界の影響を考慮する連続時間の取り扱い方を検討した。その結果、実用的なモデル検査ツールSPINを用いて前二者の問題を取り扱えることを示した。連続時間への拡張はSPINでは難しいため、RT-Maudeを用いた実験を行った。また、産業界で形式手法を利用するための参考とする際に、形式手法の技術を整理し、現状を概観するレポートを配布するなどの情報発信を行った。

2.3. リアルタイムオペレーティングシステムへの応用

組込みソフトウェアで用いられるリアルタイムオペレーティングシステム(RTOS)には、ソフトウェア実行の共通基盤として用いるため、高い信頼性が求められる。一方で、パフォーマンス向上の目的から、内部操作をインタリーブさせるなどして、複雑な実装がされている場合が多い。このような実装の正しさの確認は非常に困難であり、実際、それが原因で誤りが発生している。そこで、我々は、モデル検査手法を用いてRTOSの振る舞いを検証する手法について研究を行った。対象は μ ITRON仕様のRTOSで、モデル検査ツールSpinを用いた。モデル検査により検証を行うには、検証対象の振る舞いだけでなく、検証する動作列や性質についてもモデル化する必要がある。これはテスト手法におけるテスト仕様やテストケースの作成と同様のものであるが、モデル検査を適用することを目的としている点で異なる。そこで、モデル検査が得意で人間が苦手とする仕様横断的な性質の検証に焦点を当て、検証用モデルの作成法とモデル検査による検証手法に関する実験を行った。結果として、いくつかの知見を得ることができた。

組込みシステム向け基盤ソフトウェア

早稲田大学 理工学術院 中島 達夫

<http://www.dcl.info.waseda.ac.jp/>

筑波大学 システム情報工学研究科 追川 修一

<http://www.real.cs.tsukuba.ac.jp/>

1. はじめに

組込みシステムは今後益々複雑となり、システムの信頼性の向上はより困難な課題となっていく。特に、現状の組込みシステムでは、使用する組込みOSも μ ITRON仕様に基づく比較的単純なものからLinux等のより複雑なものを使用するようになってきている。しかし、既存のRTOS上のソフトウェアをLinuxに移行する場合、それらのソフトウェアを単純に実時間タスクとして実行すると、従来のタイムシェアリングに基づくスケジューリングを前提に作られたアプリケーションの応答性が極度に低下する可能性がある。

また、多くの人に魅力的と思われるアプリケーションは既に提供されてきているため、今後は、各個人の好みに応じてパーソナライズしたアプリケーションの提供が重要となる。その場合、ユーザが気に入ったアプリケーションを必要に応じてダウンロードできるようになる。各アプリケーションのリソース使用量をあらかじめ見積もることが困難なため、あるアプリケーションがリソース使用量を間違えるとシステム全体の動作が不安定となり、使用することが不可能となる可能性も考えられる。そのため、特にダウンロードすることにより後から追加したアプリケーションが使用するリソース量を制限することにより、はじめからインストールされているアプリケーションの動作が不安定にならないようにしないといけない。

本研究課題では、次世代組込みオペレーティングシステムのためのQOS支援とセンサー管理に関する2つの研究に取り組んだ。以下、2つの研究の概要を記載する。

2. 組込みOSのためのQOS支援

アカウントティングシステムは、アカウントティングオブジェクトという抽象化の提供をおこなう。アカウントティング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。アカウントティングオブジェクトをバインドされたプロセス群は、それらが周期毎に決められた実行時間のみを消費することを保証する。

アカウントティング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ー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。

アカウントティングシステムは、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、アカウントティングシステムの維持を容易におこなうことが可能となる。

3. 組込みOSのためのセンサー管理

センサーは外界の情報を取得するアプリケーションを構築するために必要不可欠なデバイスである。センサーを利用することにより、ユーザが誰でどこにいるか、何をしようとしているかなどの実世界情報を取得し、アプリケーションの動作をユーザの振る舞いに適応することを可能とする。ユビキタスコンピューティング環境の発展に伴い、組込みオペレーティングシステムにおいてどのようにセンサーを利用するかを明らかにすることは非常に重要になってきている。

我々は、東京ノキア研究センターと協力して様々なセンサーを搭載した携帯端末を作成した。本システム上ではLinuxが動作し、センサーをアクセスするデバイスドライバを介してアプリケーションは様々なセンサー情報を取得することが可能である。センサーデータをアプリケーションから利用可能とするためのミドルウェアを利用することによりコンテキストアウェアなアプリケーションを容易に構築することが可能となる。

我々は、開発したセンサーを利用して多様なサービスを開発し、センサーの有効性を実証した。

4. 研究成果のまとめ

我々が現在開発したアカウントティングシステムは、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における使用に関して検討が進められている。

組込みOSのためのリソース管理に関する研究成果は、JSTの「実用化を目指した組込み用ディベナダブルオペレーティングシステム」でも利用され、組込みOSのためのセンサー管理に関する研究成果は米国のラトガー大学、フィンランドの情報技術研究所等の海外の研究機関との共同研究に発展している。



組込みシステム用実時間 Java 技術

京都大学 大学院情報学研究所 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、実行基盤の開発、実証実験、要求仕様検証技術との統合、の三つのサブテーマからなる。

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り付け、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、公開してきた。さらに、実証実験および要求使用技術との融合の一環として、プロジェクト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。

これらの成果を実用化するために、昨年度から引き続き、携帯電話機に代表されるモバイル端末に適用する活動を進めている。

2. モバイル端末における Java VM

携帯電話機でも高度な Java アプリケーションが利用できる時代になってきたが、そのようなアプリケーションの多くはゲームであり、実時間性に対する要求が強い。メモリ空間が数百 KB から数 MB、さらに数十 MB と拡大するにつれて、携帯電話機用の Java アプリケーションはさらに高度化する傾向にある。従来は、クリティカルな場面で GC が起きないように、ベンダは、アプリケーションの「適切な」箇所に、GC を強制起動するためのコードである `system.gc()` を手作業で挿入していた。挿入箇所は端末機のモデルに依存し、このために開発コストの増大、開発期間の長期化、人材不足といった問題が生じていた。実時間 GC の実装によってこの状況が解決できることは明らかである。

モバイル端末における Java VM としては、Sun Microsystems 社が提供する KVM/CLDC および MIDP が主流である。KVM とは、モバイル端末に適した超小型 (KB クラス) の VM であり、CLDC とは、モバイル端末に適した想定実行環境 (configuration) を、コンパクトな基本的ライブラリによって提供する。MIDP は、描画やキー入力、ネットワーク通信など、端末機に依存するライブラリ (profile) 仕様であり、モバイル端末向けにコンパクトに設計されている。KVM/CLDC は多くのモバイル端末における Java VM のベースとなっており、特に我が国では、携帯電話キャリア三社すべてがこの VM をベースとして採用している。さらに二社が MIDP 準拠の VM を採用している。そこで、本プロジェクトでは、KVM/CLDC と MIDP に実

時間技術を実装することによって、モバイル端末における実用性を検証することとした。

3. 実装の概要

KVM/CLDC と MIDP は、Sun 社 から reference implementation (以下 RI) がソース公開されているので、それをもとに実装を進めた。この RI では一括マーク・スイープ方式に圧縮オプションを加えた GC を採用しており、1 回の GC 処理は、ルート走査、マーク、スイープの三つのフェイズからなる。これらの各フェイズを、これまで開発してきた実時間技術を適用して細分化し、実時間化を達成した。まず PC 上でテスト及び性能評価を行い、より携帯電話端末に近いプラットフォームとして、ARM アーキテクチャを採用した PDA や、多くのモバイル端末が採用している Windows CE 上にも移行して性能検討を行った。

MIDP については、Windows CE 用のコードがなく、Windows XP 用のコードを流用することにしたが、XP 用も完成度が低かったために、まず XP、次に CE 上で動作させるためにかなりの工数を要した。現時点でもサウンドや通信などの機能が未実装であるが、基本的には動作するようになった。一方、メモリ管理機能は JNI (Java Native Interface) で記述されていたために、実時間化のための実装は単純であった。

4. 実用試験に向けて

実際の携帯電話で利用されているアプリケーションを実行するための準備がひととおりそろったことになる。KVM/CLDC については商用の Java VM に匹敵する性能を出しているが、MIDP の RI は実行性能に難点があり、ゲームプログラムなどを実行するにはかなりの改造を必要とする。これは本プロジェクトの本来の目的ではないので、KVM/CLDC および MIDP に準拠した商用の Java VM に開発した技術を実装する計画を進めている。また、実用試験を効果的に行うために、ゲームベンダーからの協力を得る予定である。

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/smlsharp/ja/

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認する必要がある。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、システムに甚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ラ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つの技術を実現するプログラミング言語システムを構築し、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。

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の開発をめざす。

1. 次世代高信頼プログラミング言語

現時点で最も優れた高信頼言語はML系プログラミング言語と言える。MLでは、プログラムの静的解析技術により、プログラムの整合性の解析と検証を行い、プログラムに潜在する多くのエラーをコンパイル時に自動的に検出する、しかしながら現在のML系言語は、その基礎とする型理論の制限や処理系の実装上の制限から、

- ・基本データ構造であるレコードの扱いが不十分
- ・CやJavaなどの既存言語やデータベースなどの相互運用性の欠如

などを含むいくつかの問題点があり、実用アプリケーション開発用言語として広く普及しているとは言いがたい。

本プロジェクトの主要な具体的な目標は、上記を含む実用上の問題点を克服しかつ最先端の機能を実現する新しいML系次世代

高信頼プログラミング言語, SML#, の開発である。

2. プログラム開発のためのツール群

SML#を実用システム開発に活用するためには、既存言語とともにSML#を使いこなし、効率よく高信頼プログラム開発を行うためのプログラム開発環境が必要である。本プロジェクトの具体的な目標の二つ目は、そのようなプログラム開発環境実現のための以下のツール群の開発である。

- (ア)SML#で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)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ツール及びライブラリ
- (ウ)Javaで書かれたプログラム部品を組み合わせる際の不整合を自動検出するツール

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もなると期待される。特にSML#言語は、従来のML系言語の弱点を克服する先端機能を世界ではじめて実現するプログラミング言語であり、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。

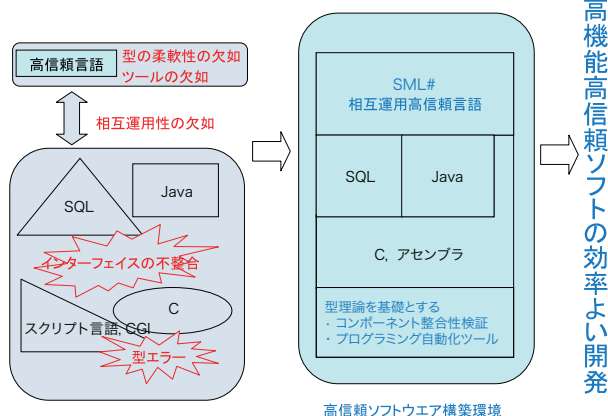


図1. プロジェクトの概要



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. 本プロジェクトの成果

4.1. SML# 言語

本プロジェクトの主要な成果である SML# 言語は、既存の ML 系高信頼言語の特徴を完全に継承し、かつ、以下の機能を実現する新しい関数型プログラミング言語である。

- ・多相型レコード演算
- ・ランク1多相性
- ・相互運用のための多相型システム
- ・レコード多相を基礎とするデータベースとの連携
- ・型主導コンパイルによる自然なメモリー表現

SML# のこれら新機能は、いずれも、我々独自の基礎研究に基づく最先端のものであり、その理論的な有効性は確認され、ACM TODS, ACM SIGMOD, ACM TOPLAS, ACM POPL, ACM ICFP 等のフォーラムで発表され認められているものであるが、いまだに実装がなされていない最先端技術である。SML# は、これら基礎理論を基にした設計と開発により、前節で論じたレコードの扱いの制限および相互運用性の欠如の問題を含む現状の ML 系言語の問題点を克服し、複雑なソフトウェア開発の生産性と信頼性の向上に貢献する実用性の高い言語を実現している。

本プロジェクトでは、それら独自の基礎理論に基づく先端機能を実現する実装方式や最適化方式の開発からシステム詳細設計、プログラム開発に至るまでのすべてを独自技術で行い、SML# 言語コンパイラ、および、ごみ集めモジュール等を含む SML# 実行時処理系の開発に成功した。

この SML# 言語処理系の開発に加えて、開発の過程で、コンパイラやコード生成、最適化などに関する新しい理論や方式などが多数産み出されている。その代表的なものには、

- ・機械語コードとコンパイルのための証明論的基礎 (2007年, ACM TOPLAS に発表)
- ・証明変換の考え方に基づく新しい系統的なレジスタ割付方式 (2004年, Sci. Comput. Program. に発表)
- ・不動点昇格という新たな考え方に基づく軽量関数融合方式 (2007年, ACM POPL に発表)

などがある。これら基礎理論や基本方式は、SML# ばかりでなく、堅牢で高性能な高信頼言語の実現に幅広く寄与し得る一般性の高いものである。

4.2. プログラム開発のためのツール群

本プロジェクトのもう一つの成果は、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可能にする以下の各種ツール群およびサポートライブラリである。

1. SML# プログラム開発ツール群

- SML# のプログラム開発を支援する以下のツール、
- ・ドキュメント自動生成ツール (SMLDoc)
 - ・清書プログラム自動生成ツール (SMLFormat)
 - ・プログラムテスト支援ツール (SMLUnit)
 - ・SML# デバッガ

2. SML# 相互運用サポートツール及び各種ライブラリ

C や Java, 言語, COM オブジェクトなどと連携するための種々のインターフェイス自動生成ツールや、多バイト文字処理ライブラリ LMLML を含む種々のライブラリ

3. Java プログラムの不整合を自動検出するツール

プログラムの静的解析により、Java の部品の結合に関する脆弱性を静的に検出するツール。

5. まとめと今後の展望

当初の予定通り、種々の先端機能を装備した新しい ML 系関数型高信頼言語 SML# のコンパイラおよび実行時処理系の開発に成功し、平行して開発した SML# を中核とするプログラム開発のための各種プログラミングツールやサポートライブラリと共に、以下の URL よりリリースした。

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

SML# は、これまでの ML 系高信頼言語の弱点であった不十分な相互運用性や不完全なレコードの扱いなどの実用上重要な課題を、その基礎をなす理論から実装技術までをすべて開発することによって解決し開発した先進性と実用性の高い、日本発の新しい言語である。また、この開発過程で、新しい理論や技術がうみだされた。SML# が示す先進性と実用性は、国内の学会の関心を得、2007年度ソフトウェア科学会大会、日本ソフトウェア科学会 PPL2007ワークショップ、2005年の日本ソフトウェア科学会 FOSE ワークショップなどで招待講演を依頼されている。

今後、SML# コンパイラをより高機能でかつ高性能な言語処理系に発展させるための開発を進め、SML# 処理系を継続してリリースしていくと共に、高信頼ソフトウェア生産に関わる産学官の研究開発者との連携を深め、本プロジェクトを、SML# が持つポテンシャルを生かした高信頼ソフトウェア開発基盤の構築へと発展させていく予定である。

安全なシステム記述言語および高信頼 OS 記述言語

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
 協力企業 日立製作所, とめ研究所, エーエヌラボ
<http://www.yl.is.s.u-tokyo.ac.jp/e-society>

1. 安全・高信頼な基盤ソフトウェアの開発技術

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りつつある。今後、さらなる情報化は不可避であり、社会基盤としてのコンピュータシステムの安全性を保証することは急務である。

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いるが、それらの多くは、ソフトウェアの安全性の問題が社会的に顕在化する以前に開発されたプログラミング言語やOSなどの基盤的ソフトウェアに依存している。このような深刻な問題に対処するためには、対症療法的な解決ではなく、理論的な方法・系統的な方法に基づいて、安全性や信頼性が保証されなければならない。

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やプログラム検証・証明技術の開発によって、既存の基盤ソフトウェアの安全性・信頼性を強化する技術を開発しその有効性を示した。

2. プロジェクトの目標・手法・成果とその検証

2.1. 目標

本プロジェクトの具体的な研究目標は次の二つである。

1) 安全なC言語システムの構築

C言語は、ソフトウェアの安全性が社会問題化する以前、今から三十年以上前に開発されたプログラミング言語であり、安全性に問題があることが知られている。しかし、言語機能の柔軟性が高いため、未だ多くの基盤的ソフトウェアが、このC言語を用いて実装されている。この問題に対処するため、我々は2種類の安全なC言語のコンパイラを構築した。

2) 安全/高信頼なOSを構築するための記述システムの開発

OS(オペレーティングシステム)はコンピュータシステムで最も基礎的なソフトウェアであり、このOSに問題があると、コンピュータシステム全体の安全性・信頼性に大きく影響する。これに対し我々は、安全/高信頼なOSを構築するため、型理論に基づいたシステムソフトウェア記述システムの開発を目指した。

なお、本プロジェクトで実現を目指した安全性/高信頼性は具体的には次の三つである。

1) メモリ安全性

プログラムが不正なメモリアクセスを行わないこと。

2) 機密情報の漏洩防止

プログラムが機密情報を(例えその一部でも)外部に漏洩しないこと。

3) 実行継続可能性

不正なメモリアクセスを検出した後でも、機密情報が外部に漏洩することなく安全に実行を継続できること。



2.2. 目標達成のための手法

本プロジェクトでは以下の四つの手法について研究を行った。

- 1) 安全なコードを生成するC言語コンパイラの開発
- 2) C言語ソースコード上の情報流の系統的解析手法の考案
- 3) OS記述用アセンブリ言語の型システム/型検査の考案
- 4) OSの安全性の証明手法/検証手法の考案

上記四手法の概要については、第3節以降で述べる。

2.3. 研究開発の主な成果とその検証

1) メモリ安全なコードを生成するANSI標準C言語のコンパイラの構築と公開¹

検証: OpenSSL(ネットワーク通信の暗号化や通信相手の認証等を行うプログラム, ソースコード約300,000行), BIND9(インターネット上での名前解決に広く用いられているサーバ, ソースコード約350,000行), などの基盤的システムをコンパイル・実行した。

2) メモリ安全で機密情報漏洩を防止できるコードを生成するC言語(+付加記述)のコンパイラVITCの構築

検証: tthttpd(小規模ウェブサーバ, 約1万行)をコンパイル・実行した。

3) メモリ管理やスレッド管理など、今まで不可能とされていたOS機能の型付き言語による記述に成功

検証: 新しい型付きアセンブリ言語TALKを考案し、それにより上記機能を実現した。

4) 高次論理系に基づく証明支援系Coqの枠組で「分離論理」体系²を定式化し、ヒープメモリやポインタへの操作の安全性を証明する体系を構築した。

検証: 教育目的で実用されているTopsy OSのメモリ管理モジュールの安全性を、Coq証明支援系を用いて証明・検証し、Topsy内のバグも発見した。

¹この開発は平成16年度末まで米澤研究室で実施され、その後の主たる開発は、産業技術総合研究所で継続され公開された。
<http://www.rcis.aist.go.jp/project/FailSafeC-ja.html>

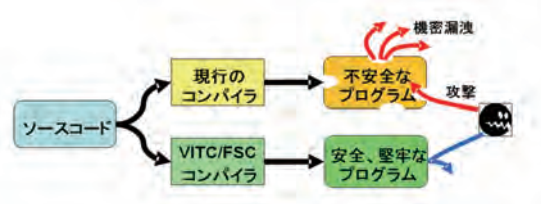
3. 高安全なC言語コンパイラの開発

C言語は、その実行速度やハードウェアメモリモデルについての記述力から、Javaなどの新言語登場後の現在もシステムプログラムの記述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一般的なバグにより誤動作したり、メモリ管理のバグによってセキュリティーホールが出現し、攻撃を受けることでシステムが破壊されたり機密情報を盗まれることが多発している。これは、C言語がメモリ保護機構を欠くためや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に起因する。

3.1. メモリ安全C言語システム (FSC) の開発

メモリ安全性を保証し、コンピュータウイルスによるメモリ脆弱性攻撃を検知、緊急停止するコードを生成するC言語コンパイラを設計し実装した³。我々の研究成果によって得られた型システムに基づくアルゴリズムを用いて、ANSI規格を包含するC言語プログラムを機械語コードにコンパイルする段階で、安全性を検査するコードを自動的に挿入するコンパイラを構築した。



VITC/メモリ安全Cコンパイラ

3.2. VITC言語システムの開発

C言語プログラムにおけるメモリ安全性を保証することに加えて、機密情報の漏洩を防止できるコードを生成するVITC言語コンパイラを構築した。ここでは、プログラム内の機密情報の流れを追う、いわゆる情報流解析を行う機構をコンパイラに導入している。この情報流解析では、変数の持つ値の機密性の高さを「型」と見なすことによって、新しい型システムを考案し、この型システムによってプログラムを解析している。

また、情報流解析により機密情報を漏洩する危険がなくなるため、VITCでコンパイルされたプログラムは、従来の攻撃にさらされても安全に実行を継続することができる。

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、情報流もほぼコンパイル時に、すなわち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。我々のVITCの研究では、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

難しいため、実行時の情報流の動的検査を導入している。また、情報の機密度を表す型は、オリジナルなC言語には備わっていないため、アノテーションの付記が必要だが、記述量は必要最小のものとなっている。

4. 安全/高信頼なOS構築用記述システムの開発

4.1. OS機能の型付きアセンブリ言語による記述

近年の型理論研究の進歩により、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#)を用いて既に作成されるようになっている。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、実行時に予期せぬメモリエラーを生じない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で重要なプログラムであるOSは、未だに強く型付けされていない言語を用いて作成され続けている。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り、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、OSカーネルの重要要素(メモリ管理機構やスレッド管理機構など)の記述ができる、強く型付けされた安全なアセンブリ言語を設計・実装した。これで、バッファオーバーフロー等のセキュリティ脆弱性の原因がOSに存在しないことを保証する方法を与えることに成功した。

4.2. ヒープメモリ・ポインタ操作の安全性証明の体系の構築

広く利用されている既存のOSは、既にC言語で構築されてしまっているが、これを一から安全な言語で構築し直すことは常に現実的であるとは限らない。このため、既存のOSの安全性を検証することは重要である。これに対し本研究は、分離論理²にもとづいて、ヒープメモリ・ポインタ操作の安全性を数理論理的に証明する体系を構築し、これを用いて実際のOSカーネル(Topsy)のメモリ管理システムの正当性の検証を行った。

具体的には、定理証明支援系 Coq を用いて分離論理の体系を定式化し、ヒープメモリ・ポインタ操作の安全性を検証するための体系を構築した。これを用いてTopsyのメモリ管理システムのソースコードをこの定式化の枠組内で検証したところ、Topsyのメモリ管理システムには、メモリーーク、すなわち本来ならば再利用できるはずのメモリを再利用せず、結果としてシステムのメモリ全てを消費してしまうという問題が生じる可能性があることなど、幾つかの問題があることを発見できた。

²互いに交差しない(分離された)メモリ領域上で使われるプログラム変数やポインタの性質を自然に表現できる論理体系。

³前出の脚注1を参照。

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 奈良先端科学技術大学院大学 飯田元, 松本健一；大阪大学 井上克郎, 楠本真二

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 (EASEプロジェクト: Empirical Approach to Software Engineering Project) 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エンピリカルアプローチ (実証的手法) の実践を目指した。以下では、その主な成果と活動を紹介します。

2. エンピリカルアプローチのためのツール群

エンピリカルアプローチでは、ソフトウェアの生産性と品質の向上を目的として、ソフトウェア開発に関するデータの「収集」と「分析」、そしてそれらに基づく「改善」という3つのフェーズが繰り返される。

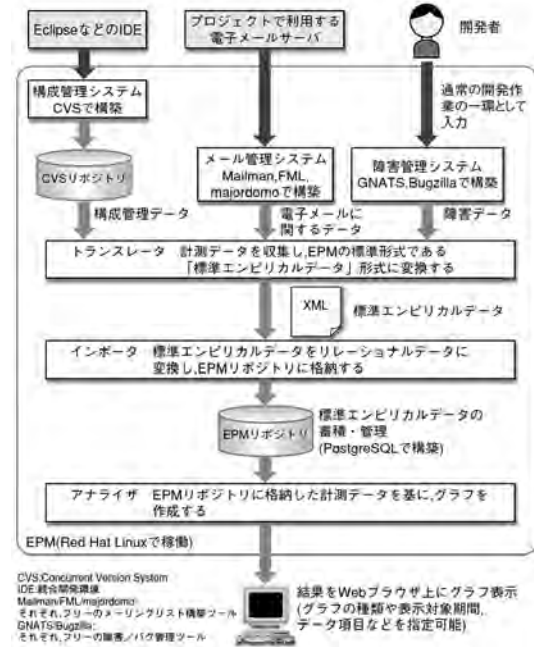
■**収集**: 開発作業をリアルタイムに記録する。ソースコードの改編、障害の対応、電子メールのやりとり等を表すデータを、人為的操作が加えられていない状態で収集する。

■**分析**: 収集されたデータを基に、ソフトウェアの機能、品質、コスト、進捗などを、開発者、管理者、ユーザなど様々な視点から分析する。

■**改善**: 分析結果とそこから導き出された改善案をソフトウェア開発プロジェクトにフィードバックし、生産性と品質の向上を図る (開発支援を行う)。分析結果と改善案は組織の知識として蓄積され、再利用される。

EASEプロジェクトで開発し、オープンソース化したEPM (Empirical Project Monitor) は、エンピリカルアプローチを実践する環境の一つであり、「ソフトウェア開発における自動的なデータ収集と分析のためのプラットフォーム」である (図1)。EPMでは、エンピリカルアプローチにおけるソフトウェア開発者の負担を小さくするために、ソフトウェア開発プロジェクトで広く使用されている、構成管理システム、メール管理システム、障害管理システムからデータを収集する。収集データをXML形式に変換した後で分析を行う方式を採用しており、データ収集や分析をEPM以外のシステムやツールで行う等、既存の開発環境や社内システムとの連携も可能である。

本プロジェクトでは、EPM等と並行して、データ分析やそれに基づく改善を支援するツールを開発した。これらをEPMと合わせてEASEツール群 (Tool Suites for Empirical Approach to Software Engineering) と呼ぶ (図2)。主なツールは次のとおり。



■プロセス計測・分析

- エンピリカルデータ収集・分析プラットフォーム: EPM
- エンピリカルデータ分析支援ツール: EPM Pro*
- エンピリカルデータクラスタリングツール群: EP-Cluster Suite
- マイクロプロセスデータ可視化ツール: MPA-Plot /type 1
- ソフトウェアプロジェクト仮想再現ツール: Project Replayer
- 開発作業記録システム: EASE.CORE

■プロダクト計測・分析

- ソフトウェアレポジトリ自動分類システム: MUDABlue
- 流用部品検索ツール: Stigmata for Web
- コードクローン履歴閲覧システム: CHAN
- 分散型コードクローン検出システムツール名: D-CCFinder
- 統合開発環境向けリアルタイムクローン検出アドイン: SHINOBI

■プロジェクト分析

- ワンクリック見積り&データ診断ツール: Magi
- 相関ルール分析ツール: NEEDLE

■開発管理支援

- プロセスデータガイドブック: EPDG2
- 開発知識協創支援システム: D-SNS

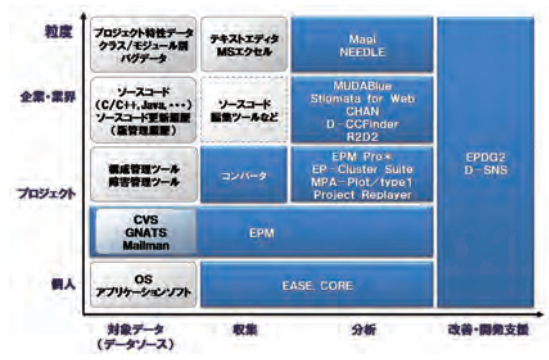


図2 EASE ツール群

3. 産官学連携活動

EASEプロジェクトでは、エンピリカルソフトウェア工学ラボを開設すると共に、企業技術者をメンバーとするEASE産業部会を設置し、産学官連携の下、EASEツール群の開発やソフトウェア開発プロジェクトへの適用と評価を行った。連携企業等はのべ120社を超える。例えば、EPMの開発は、同ラボに常駐する協力企業4社の技術者と共同で行った。また、EPMの最初の適用実験と評価は、日立公共システムエンジニアリングと三菱スペース・ソフトウェアの多大なる協力を得て実施した。

平成17年度からは、独立行政法人 情報処理推進機構 (IPA) のソフトウェア・エンジニアリング・センター (SEC) との連携が本格化した。SECが実施した「先進ソフトウェア開発プロジェクト」では、EPMをはじめとするEASEツール群をベースとすることで、広域マルチベンダー開発における定量的な開発データの自動収集、分析、フィードバックというプロセスの構築に成功した。その有用性は、ベースとなったEASEツール群と共に高く評価され、SECでは、EPMを産業界へ展開する活動を推進している。

EASEプロジェクトでは、その活動や成果を産業界に発信、普及するための活動にも注力してきた。ウェブサイト <http://www.empirical.jp/> の開設、ニュースレターやアンニュアルレポートなど紙媒体での情報公開等に加え、平成16年4月からは、東京田町のキャンパス・イノベーション・センターにて「エンピリカルソフトウェア工学研究会」を開催してきた(第12回のみ大阪で開催)。参加者数は全16回でのべ約800名(1回あたり約50名、EASEプロジェクトメンバーは除く)にものぼる。同研究会では、EASEプロジェクトの研究成果の単なる報告だけでなく、連携企業における研究事例や産学連携活動の紹介、エンピリカルソフトウェア工学分野の著名な研究者による講演や最新の技術動向に関するチュートリアルなど幅広く実施した。同研究会がトリガーとなり共同研究開発に発展したケースも多数あり、産と学の距離を近くした意義は大きく、産業界から高く評価されている。

4. 国際連携活動

EASEプロジェクトでは、海外の研究者との連携を重視してきた。4名の著名な研究者(Prof. Victor R. Basili, Prof. Barry W. Boehm, Prof. Dr. Dieter H. Rombach, Prof. Ross Jeffery)を海外アドバイザーとして迎え、計13回の会合を通じて、海外の研究開発動向をいち早く把握すると共に、EASEプロジェクトの活動や成果を海外に向けて発信し、National ICT Australia (NICTA) 等における後発の同種プロジェクトに影響を与えるなど、エンピリカルソフトウェア工学研究を国際的にもリードしてきた。

海外の研究開発の動向を国内の研究者に伝える場としてEASE国際フォーラムを2回開催した。第1回(平成15年11月、東京で開催、参加者約400名)では、4名の海外アドバイザーとEASEプロジェクトメンバーによる基調講演や国内の産学の研究者を加えたパネル討論を通じて、ソフトウェア開発におけるエンピリカルアプローチの現状や将来性について活発な議論や情報交換が行われた。第2回(平成18年5月、東京で開催、参加者約150名)では、5件の招待講演(インド1件、米国2件、日本2件)とパネル討論、そしてEASEツール群のデモ展示を行った。特に、インドからの発表では、平成17年8月にオープンソースとして公開されたEPM英語版がインドでも注目されており、自国のあるソフトウェア開発企業の自社向け開発支援環境に同様の機能が組み込まれたとの紹介があった。

5. おわりに

以上のように、EASEプロジェクトでは、EPMをはじめとするEASEツール群を開発すると共に、産官学連携や国際連携の活動を通じて、ソフトウェア開発におけるエンピリカルアプローチの実践を推進してきた。その活動や成果は、EASEプロジェクト外部評価委員会はもとより、国内外で高く評価されている。その結果、平成18年度経済産業大臣表彰(情報化促進部門)をはじめとして、学術論文誌や国際会議における最優秀論文賞等を、プロジェクトメンバーが多数受賞している。

EASEプロジェクトを実施した5年間で得られた研究成果は、産学連携の共同研究や学術論文等として既に広く公開している(招待論文・講演46件、学術論文誌掲載論文34編、国際会議発表69件をはじめ外部発表は計340件余り)。また、EASEツール群は、一部を除いて、解説ドキュメントと共にCD-ROMでの配布やウェブサイトでの公開を行う。これらの成果を産業界に根付かせ、日本のソフトウェア産業の発展に資するために、プロジェクトメンバーが中心となって、「技術移転合同会社」を、平成19年度末をめぐりに設立する。エンピリカルソフトウェア工学研究会も継続し、学会との連携等により、産からだけでなく学からも多くの参加者を集める。それらを核として、エンピリカルソフトウェア工学研究をさらに発展、加速させていく。



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
 協力企業 (株)ジャストシステム
<http://www.psdlab.org/>

1. 高信頼構造化文書変換技術研究の概要

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に代表されるこれまでの技術は、発展の著しい情報環境下で既存技術を拡張してできたものを事実上の標準としているが、その言語的な概念は十分に整理されているわけではなく、そのような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に欠ける文書情報を蓄積する一因となっている。本研究課題はこうした問題を解決するために、体系的な文書処理技術の基盤を確立して高信頼XML文書処理(Dependable XML Processing, DXP)を実現することを目的としたものである。そこでは、文書処理に信頼性の高い技術基盤を与えたとともに、文書を作成する利用者とソフトウェア開発者に有用なシステムを提供することとした。

本研究開発によって得られた基盤技術の成果は学術論文として発表しており、開発したソフトウェアはWEBを通じて提供している。代表的な利用者向けソフトウェアとして、プログラマブルXML文書編集システムbiXfyとWEBサイト編集支援システムVu-Xを、また、ソフトウェア開発者向けにXML文書の双方向変換を実行する双方向変換Bi-Xサーバを公開している(<http://www.psdlab.org/>)。

2. 高信頼構造化文書処理の基盤技術

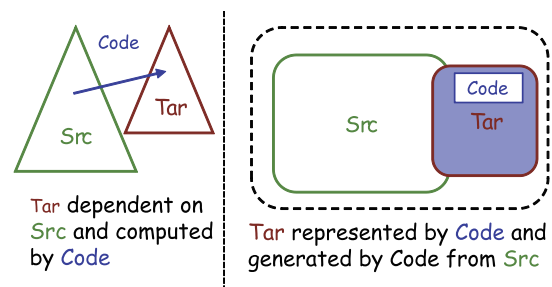
本研究開発では、文書内にプログラムの記述を含むプログラマブル構造化文書(Programmable Structured Document, PSD)の枠組みと、構造化文書の双方向変換(Bidirectional transformation)の基礎となる双方向変換言語Bi-Xの設計とその処理系の開発を基盤技術としての中核において革新的な発展を導いてきた。

PSDは文書内で相互に依存する項目間の整合性を確保して文書情報の信頼性を高めることを目指したものである。また、構造化文書の双方向変換は、さまざまな場面に適用できる文書間の変換の枠組みであり、一方向の変換プログラムをBi-X言語で記述すれば、対応する逆方向の変換プログラムもそれと同一のもので実現されるというものである。これにより、両方向の変換相互の整合性を保証して信頼性を確保すると同時に、文書変換ソフトウェアの開発効率を高めることになる。このように、これらの基盤技術は、構造化文書処理において信頼性を高めると同時に、ソフトウェアの生産性を高めるために有効なものといえる。

3. プログラマブル構造化文書

本研究開発で開発したプログラマブル構造化文書(PSD)とその計算機構の基本は、文書内項目間の依存性を一般化して扱うとこ

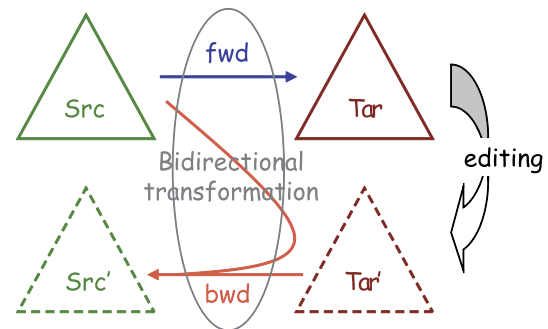
ろにある。文書内のある部分Tarが同一文書の他の部分Srcに依存して決まるという例は、表計算において列のいくつかの数値要素の総和が列の要素から決まることや、書籍の目次が本文の章の題目から定まるといったところに見ることができる。PSDの考え方は、このような依存性をプログラムCodeによる計算によって表現し、表計算のような計算機構を付随させて構造化文書処理を実現しようとするものである。



PSDの計算機構では、数値計算だけではなく文書の一部を結果とするような計算も許容する。具体的には、SrcからCodeによってTarを生成するものとして、文書中のTarにあたる部分にCodeを埋め込んでPSD文書を構成する。このようなPSD文書内のCodeを記述するためのプログラミング言語は一般的なものでよい。本研究開発では、PSD文書としてのXML文書と、その内部に置かれたCodeを計算する評価系を繋ぐ一般的なインタフェースをサーバ・クライアントモデルに基づいて開発し、PSD文書の実現例として木構造計算シートTreeCalcやインタラクティブ文書iDocumentを公開した。

4. 双方向変換による構造化文書処理

双方向変換は、原文書とそれから抽出等の変換によって得られた文書との間の両方向の変換を一体的に表現するものである。



構造化文書処理においては、原文書Srcの一部を抽出したり項目を組み替えたりする変換fwdによって目的に合った文書Tarを作成することが頻繁に行われる。データベースの問合せ処理の場合には、問合せに相当する変換fwdによってデータベースSrcか

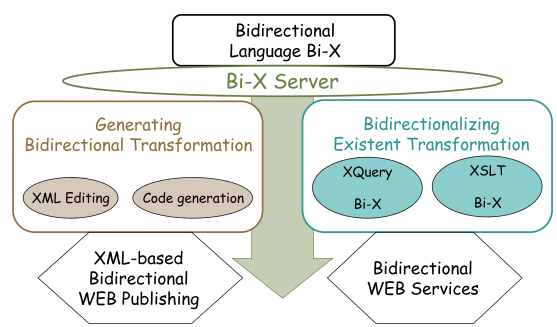
らビュー Tar を得て情報を手にすることになる。このような場面で、ビュー Tar 上で変更を行って Tar' を得たときに、それに見合う変更をデータベースに反映させた Src' を得るというビュー更新も一つの構造化文書処理である。このときには、fwd とは逆方向の変換 bwd が働くことになる。もちろん、順方向変換 fwd に対応させて個別に逆方向変換 bwd を用意すれば双方向の処理は可能であるが、両者の整合性を保証することは容易ではない。本研究開発では、これらの両方向の変換を一つの双方向変換として記述するための言語 Bi-X を設計し、Java によるライブラリ Bi-XJ を開発した。

言語 Bi-X は、プログラムを組み合わせるより高度なプログラムを作ることができるように設計されていて、Bi-X プログラムはポイントフリーと呼ばれる形式をとっている。この形式は一般のプログラムには馴染みの薄いものであり、プログラム開発の便を図るために、一般の XML 処理言語 XQuery や XSLT で記述されたプログラムを Bi-X の双方向変換プログラムへ翻訳するツールも開発して提供している。

5. 高信頼構造化文書処理サービスの提供・公開

双方向変換 Bi-X サーバ

双方向変換は、更新を伴う2つの構造化文書の同期処理をはじめとして広範囲に展開できる技術であり、多方面での利用が見込まれるので、この機能を提供する Bi-X サーバを提供して WEB を通じて利用できるように公開している。



プログラマブル XML 文書編集システム biXfy

プログラマブル XML 文書を作成するためのシステムとして、本研究開発の協力企業である(株)ジャストシステムの xfy Client

(<http://www.xfy.com/jp/>)のプラグインとして機能する biXfy を開発して提供している。これは、xfy が扱う XML 文書を用いて PSD 文書を作成するとともに双方向変換を利用した編集機能を支援するものである。ここでは、XML Schema の定義に Bi-X のプログラムを記述することにより、それに適合する XML 文書 Src に埋め込む Code を生成し、必要に応じて Tar を得るものである。ここで、xfy Client の編集機能を用いて Src, Tar のいずれの部分をも編集できるが、Tar を生成する双方向変換を実現する Bi-X プログラムによって、その変更は統一的に一貫性をもって XML 文書に反映さ

れる。プラグイン biXfy は小規模データベースの構築と各種アプリケーションを文書作成の感覚で容易に構築することができるという利点を与えている。これを用いた図書の帯出管理システム等の例とともに提供している。



WEB サイト編集支援システム Vu-X

構造化文書に対する双方向変換に基づく実用的なアプリケーションとして、XML-based WEB パブリッシングを支援するシステムを開発し、WEB を通じたサービスとして公開している。XML-based WEB パブリッシングは、複数のページからなるサイトで整合性を保持しつつ更新してゆくには、共通の情報を XML データ Src として管理して、ページごとに順方向変換 fwd によって XHTML 文書(ビュー)Tar を作るのが望ましいという考えによる。しかし、この方式では各ページの更新に対応する Src 上の編集箇所を的確に得ることが難しいという問題がある。これに対して Vu-X では双方向変換によってこれを解決している。ここでは、ビュー Tar 上の操作によって Src の更新や変換プログラム fwd の生成を行って WEB サイトの編集を支援する。この Vu-X は Bi-X サーバを利用しており、標準の WEB ブラウザ上で編集操作を行うことができるようになっている。



6. おわりに

本研究開発は 2003 年度～2007 年度の 5 年間にわたって実施したものである。この間にも関連技術の発展は目覚ましいものがあるが、その中であっても本成果は高信頼構造化文書処理の先進的技術として位置づけることができる。

高信頼 WebWare 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

WebWareは、(1) XMLやHTMLなどのコンテンツ記述、(2) レンダリングのためのCSSやXSLT、(3) JSP やJava Servletなどのサーバサイドプログラム、(4) JavaScriptなどのクライアントサイドプログラムなどから構成される、大規模でヘテロジニアスな分散ソフトウェアである。その広がりはいまさら触れるまでもなく、水道・ガス・電力と同様な社会のインフラストラクチャとして位置づけられるため、WebWareの信頼性を確保することは社会的課題である。

WebWareを構成する各要素の機能が向上している一方で、WebWareに対してその全体を記述する枠組みは欠けており、このことはWebWareの信頼性を低下させる要因である。本研究の目的はWebWareの信頼性と安全性を保証しつつ、デザイナーが行なうレンダリング・エディトリアル作業とエンジニアが行なうシステム構築作業を高度に統合するWebWare開発環境を構築することにある。WebWare開発環境は(1) テスト支援システム、(2) 細粒度リポジトリ、(3) 作成支援システムから構成される。

以降、WebWare開発環境を構成する各サブシステムについて本研究の成果を述べる。

2. WebWareのテスト支援システム

WebWareのテスト支援システムは、大規模WebWareに対する信頼性検査を自動的に行うシステムである。本システムでは、テストケースの生成、テスト実行などを自動化することにより、WebWareのテストに必要な工数をこれまでの1/5に削減することを目標とした。

本システムに関する主たる成果は以下の通りである。

(a). JSPの単体テスト支援フレームワーク

MVCモデルに基づき構築されるWebWareのViewプログラム単体テストに対して、テスト生成とテスト実行を自動化するテスト支援フレームワークを開発した。

協力企業にて、実際の開発プロジェクト3件に対して、本フレームワークを利用して構築したJSP単体テスト支援ツールを適用した結果、単体テストの工数が約1/3に削減された。

(b). WebWareのテストシーケンス生成ツール

WebWareのユーザインタフェースプログラムであるページ生成プログラムのテストに対して、ページ網羅をはじめとする3種類のカバレッジ基準を提案し、WebWareの機能テストとしてカバレッジを100%にするテストシーケンスを生成するツールを試作した。本ツールが生成するテストシーケンスにテストデータを与えること

で既存のテスト自動化ツールと連携させ、WebWareに対する機能テストの自動実行が可能である。

ページ数30のアプリケーションに対して本ツールを適用した結果、カバレッジを100%にする20本のテストシーケンスが生成された。生成に必要な時間は約2分であり、本ツールは現実的な時間でテストシーケンスを生成できる。

(c). 宣言的ルール記述によるコーディング規約検査器

宣言的な検査規約の記述により検査機能を追加可能なJSPプログラムに対するコーディング規約検査器を開発した。

本ツールを利用して、協力企業の開発プロジェクトで適用されたコーディング規約を記述した結果、78%のコーディング規約が本ツールにより検査可能であった。

(d). 検査項目のプラグイン化によるコーディング規約検査器

Ajax技術を利用したWebWareにおけるJavaScriptプログラムに対するコーディング規約検査器を開発した。本ツールは検査機能をプラグインとして追加可能である。

本ツールを協力企業のAjaxフレームワークに含まれる44,000行のJavaScriptプログラムに適用し、327件の見逃されていたエラーを検出した。協力企業の開発者によって検出したエラーの妥当性の確認を行い、本ツールの有用性を確認した。

3. WebWareを対象とした細粒度リポジトリ

WebWareを対象とした細粒度リポジトリは、WebWareを構成するコンテンツ記述やクライアント/サーバサイドのプログラムを統一的に解析するシステムである。本システムでは、WebWareを構成する各要素の解析に加え、要素間の関係の解析とそれらの情報が容易に取得できることを目標とした。

本システムに関する主たる成果は以下の通りである。

(a). WebWareモデルとリポジトリAPIの構築

WebWareのテスト支援および作成支援の基盤環境として、WebWareの構造とデータフローを表現するモデルを確立した。また、例として、Strutsフレームワーク上に構築されるWebWareを対象に、その実装からWebWareモデルを生成する解析系と高度なビューを提供するAPIを開発した。

(b). 解析能力およびビューの有用性評価

Strutsフレームワーク上に構築された22のオープンソースのWebWareに対して解析を行い、ページ遷移やデータフローの情報が得られることを確認した。また、開発現場における本システムの有効性については、現在、協力企業により評価中である。

4. WebWareの作成支援システム

WebWareの作成支援システムは、デザイナーが行う視覚的表現の構成作業とエンジニアが行う情報処理プログラムの構成作業の連携をとりつつ、WebWareの構築を支援するシステムである。本システムでは、デザイナーによって配置・構造化される情報が、エンジニアが扱うアプリケーションロジックに適切に反映されることを支援する機能の開発を目標とした。

本システムに関する主たる成果は以下の通りである。

(a). JavaScriptアプリケーションの作成支援

ユーザインタフェースの構築技術として、JavaScriptとHTMLの関係を考慮した解析技術を、サイト構造などの情報を視覚化するツールである「Ridual」に組み込んだ。75のページから構成されるWebサイトに適用し、従来の技術では得られなかったJavaScriptによる32のページへのリンク構造を正しく取得できることを確認した。

(b). CSS記述の作成支援

CSS記述を対象にしたリファクタリング手法を検討し、類似したルールセットの組を見つける技術を開発した。実際に6つのCSSファイル(75のルールセット)を持つWebサイトに適用し、9組の類似したルールセットを統合して5つのCSSファイル(66のルールセット)へのリファクタリングが可能であることを確認した。

(c). スタイル変換記述に基づくWebWareフレームワーク

バックエンドとユーザインタフェースを、XMLによるデータ表現とXSLTによる変換技術により連携させるフレームワークとして、中間的なデータ構造を利用した連携手法(二相XSLT構造)を提案した。本手法を適用した結果、109行のJavaScriptアプリケーションに分散して出現していた37の分岐条件が、XML(タグを含め262行)では3~10行の3つの条件群にまとめて記述することができ、条件分岐を中心としたモジュール化が確認できた。

(d). WebWareのページ構成決定支援ツール

ページ構成の決定を支援する技術として、Webサイトを構成する要素と要素間の関係に対する設計ルールの定義および、ページ構成決定プロセスの提案とツール化を行った。Webサイトを構成する要素は、``リンク``と``リンクによって結びつけられる情報のかたまり``、``外部リンク``、``カテゴリ``、``外部から放射的に参照される要素(トップページなど)``である。

最大4階層、36種類のページで構成されるサイトを対象に提案手法に基づいてリデザインを行なった結果、最大5階層、38種類のページになった。第1階層のカテゴリ数が10項目から8項目に整理される一方で、1つのページに含まれていた複数の情報が分割され、サイト利用者の目的ごとに情報へのパスを分けることができた。

5. おわりに

本研究では、信頼性と安全性を保証しつつ、デザイナーとエンジニアの作業を高度に統合するWebWare開発環境の構築を目的とした。WebWare開発環境の各サブシステムで掲げた目標に対して、それぞれ以下のように目標を達成できた。

(1). テスト支援

単体テスト支援フレームワークおよびテストシーケンス生成ツールにより、テストの設計、実行にかかるコストを大幅に削減できる。また、コーディング検査規約器により、バグの検出がテスト実行前に可能となり、テスト総数を削減できる。これらにより、テストに必要な工数を1/5に削減できた。

(2). 細粒度リポジトリ

HTMLやJSP、JavaScriptなどWebWareを構成する各要素に対する解析器を開発し、それらの解析結果を統合してWebWare全体の構造と制御フロー、データフローを表現するモデルを開発した。本リポジトリを利用することで、WebWareの整合性検査器やコーディング検査器などが容易に開発できた。

(3). 作成支援

JavaScriptやCSSの作成支援、ページ構成決定支援では、デザイナーによって配置される情報の構造化および再構成の支援を行った。さらに、スタイル変換に基づくWebWareフレームワークにより、デザイナーとプログラマが扱う情報の変換が容易になり、モジュール性の高い分業を実現できた。

本研究における研究発表成果として、これまでに論文誌12件、国際会議12件、国内研究発表82件を発表している。また、本研究における産学連携の効果として、オブジェクトワークスやInterstageなどの協力企業の製品に対して、既に本研究の成果の一部が組み込まれている。

本研究において開発したソフトウェアの一部は、以下から入手可能である。

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

・<http://www.sapid.org>

・<http://sdlab.sys.wakayama-u.ac.jp/jsanalyzer/>

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.yama.info.waseda.ac.jp/~yamana/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144.5億のWebページをもとに推定すると、2008年1月時点で15,558万台のWebサーバから合計で350億ページとなる¹。このような膨大なWeb情報の中には、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。本プロジェクトは、こうした膨大なWeb情報の効率的な収集と収集されたWeb情報の利活用を目指して実施した。

プロジェクト実施期間を通して収集されたWebページ総数は144.5億ページに上る。これは、商用検索エンジンやインターネットアーカイブでの収集に次ぐ規模であり、研究用途としては世界最大である。また、2006年9月からは収集済の144.5億ページから日本語で記述されたページを抽出し、日本語ページを持つ約100万のWebサーバに対して毎月更新を行った。こうした実験を通してWebサーバに負荷をかけることなく効率的にWebページを収集するための仕組みを構築した。

収集されたWebページの利活用については、様々な解析や実験を試みた。まず、全世界のWebページの分布状況の解析を世界に先駆けて実施した。また、e企業調査プロトタイプシステムを構築し、Web上のデータを解析することで各企業の特徴、戦略、評判などをビジネスの場で利用可能であることを示した。

2. 研究課題と成果

本プロジェクトでは、(1) 平均して1ヶ月以内の新しいデータを更新することを可能とするWWWクローラを開発すると共に、(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た。

2.1. 世界最大のWebデータ収集

総収集ページ数は14,456,201,906ページである。世界中で発見したWebサーバ数は約13,468万台であり、内8,116万台の収集を完了した。なお、8,116万台の内、実際に収集できたのは5,548万台であり、2,568万台(収集済サーバの約32%)は既にIPアドレスが存在しない等の理由からアクセスができなかった。また、256万台(収集済Webサーバの3%)については、/robots.txtによりクローラによるアクセスが禁止されていた。残り約5,000万台については未収集である。

表1は収集済の144.5億ページ中の10,696,996,553ページについてトップレベルドメイン(TLD)別の構成割合を示したもので

あり²、図1はWebページの記述言語分布を示す。表1より、“.com”に属するWebページは、全体の38.05%を占めており、Webが“.com”に偏っていることがわかる。また、図1より日本語(jp)が英語(en)に次いで多くのページ数を占めており、インターネット上の第2言語となっていることがわかる。

表1 全収集ページのTLD分布

TLD	国名	取得ページ数	割合	平均ページ / サーバ
.com	-	4,070,092,124	38.05%	339
.net	-	890,604,259	8.33%	169
.de	ドイツ	878,838,449	8.22%	459
.org	-	745,984,032	6.97%	(※)
.jp	日本	543,654,556	5.08%	609
.ru	ロシア	407,169,769	3.81%	(※)
.pl	ポーランド	321,209,334	3.00%	185
.uk	イギリス	240,244,507	2.25%	(.com, .net, .jp, .de 以外の平均)
.edu	-	232,132,978	2.17%	
.nl	オランダ	215,722,380	2.02%	
.cn	中国	185,907,711	1.74%	
.it	イタリア	156,657,707	1.46%	
.kr	韓国	151,025,640	1.41%	
.us	アメリカ	143,135,686	1.34%	
.fr	フランス	129,326,495	1.21%	
other	-	1,385,290,926	12.95%	

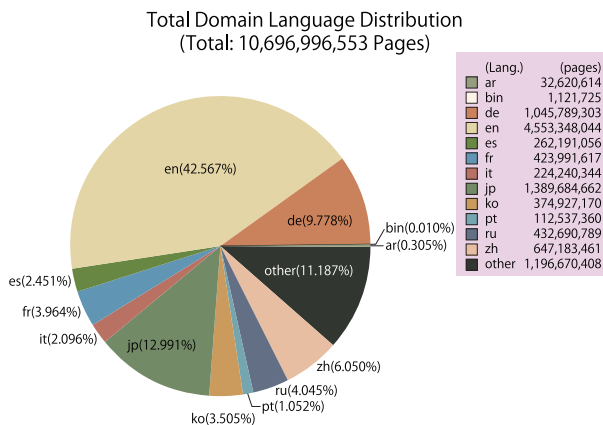


図1 全収集ページの言語分布

TLD分布において、上位15位にランクされる国は、ロシア・ポーランド・中国を除き、全てインターネット普及率が2007年6月現在50%以上であり、インターネット普及率との関連性が強いことが理解できる。さらに、ドメインによってWebサーバあたりの平均Webページ数が大きく異なることが明らかになった。従来、1台あたりの平均Webページ数は200ページ前後と考えられていたが、日本やドイツでは、これを大幅に上回るページ数が確認された。これは、インターネット普及率の高い国では、CGI等で動的に生成されるWebページ数が急激に増えていることが理由だと考えられる。

¹ 平均224.3ページ/Webサーバである。ただし、各サーバのトップページからリンクを最大15階層(CGIページからのリンクは1階層のみ)まで辿るとし、1サーバあたりの収集上限を5万ページとした場合のデータである。

² 機器故障のため約37.5億ページが失われた。

図2は、富士通株式会社と共同で開発した分散収集型クローラである。図に示すように、現在どの地域のWebページを収集しているかをリアルタイムで表示できる。本クローラは、複数拠点での分散収集ができ、約30台のPC（メモリ1GB）で全日本語Webページ（約100万サーバ）を1ヶ月以内で収集できる。

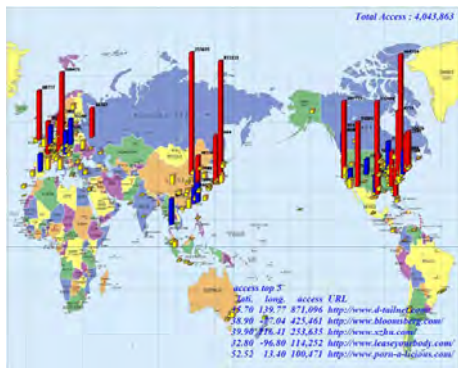


図2 収集先リアルタイム表示システム（富士通との連携）

2.2. 知識フィルタリング技術の開発

知識フィルタリングの技術開発では、超大規模データを対象としたデータマイニング手法としてTF2P-growth手法を開発すると共に、ブログ検索システム、日本全国IT導入事例分析可視化システム、Webコミュニティ抽出システム、e企業調査プロトタイプシステムなどの具体的なアプリケーションを設定し知識フィルタリング技術の有効性の検証を行った。

図3は、クローラと連動して動作するカテゴリ別データ抽出ツールであり、ある固有表現が現れた日と地域を検索できる。

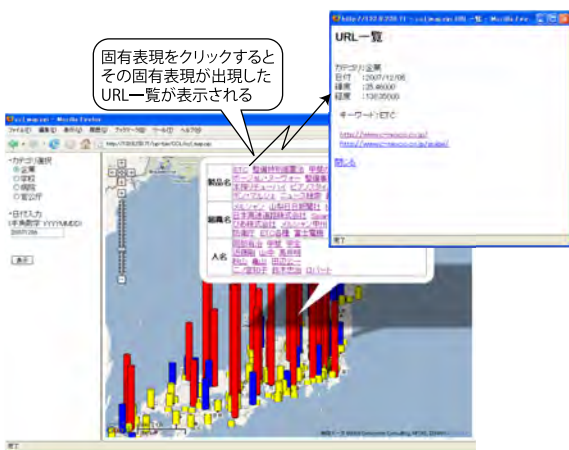


図3 カテゴリ別データ抽出ツール（富士通との連携）

図4は、アクセラテクノロジ株式会社と共同で開発したe企業調査プロトタイプシステムであり、Web上のあらゆる情報から特定企業に関する情報を集約して表示させることができる。



図4 e企業調査（アクセラテクノロジとの連携）

3. 実施体制と技術移転

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、図5に示す実施体制で進めた。これまでの技術移転としては、図6に示す2製品、及び図下部のe企業調査情報のオンライン提供がある。

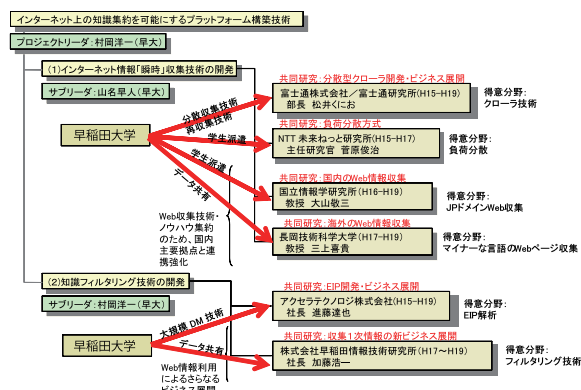


図5 実施体制

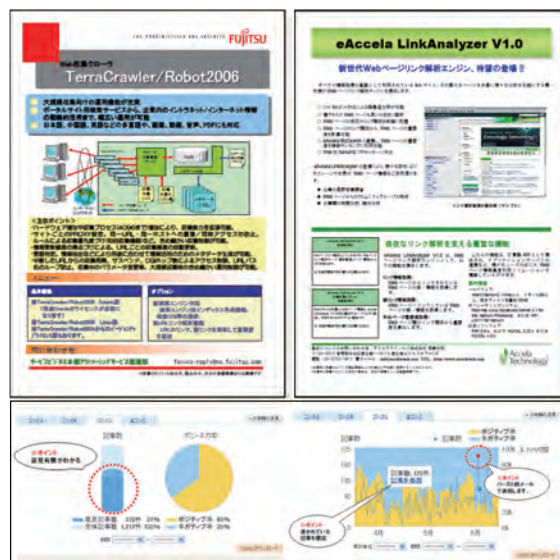


図6 技術移転

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、情報システムのデータ格納庫であるストレージの役割はサーバのそれ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ならびに管理容易化技術の実現は必須の課題である。

本研究開発では、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、ストレージフュージョン (Storage Fusion) と名付けたストレージをサーバ上のデータベース・アプリケーションと融合させる次世代ストレージ技術の確立を目指した。戦略的競争力を実現すべく、焦点を絞り、広域災害に対し強固なデータ保護を実現する高度ディザスタリカバリ機構、高レベルのソフトウェア動作知識を利用する事により入出力性能を大幅に向上させるストレージ超高速アクセス機構、ストレージとデータベースの高度な性能管理を容易に実現可能とするストレージ管理コスト低減機構を開発した。

2. 高度ディザスタリカバリ機構

テロやハリケーンなどの災害による業務の停止は、社会や国家に甚大な影響を与えることが広く認識され、業種によっては国家レベルの法制度によって業務継続が義務付けられるようになりつつある。業務継続には、ITシステムによる完全なデータ保護と早急なサービス復旧が肝要であり、遠隔地にバックアップシステムを設け、災害時に当該システムで業務を継続するディザスタリカバリシステムは極めて重要である。

本研究開発では、DBMSの記憶空間の構成に着目し、データベースのデータをバックアップシステムに非同期転送し、ログを同期転送する手法を提案するとともに、更に、当該方式を発展させログのみを同期転送し、バックアップシステムで転送されたログをデータベースに適用する新たなログ同期レプリケーションを提案した(図1左)。

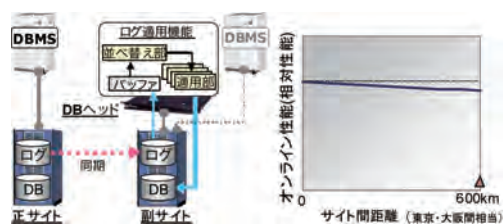


図1 ログ同期レプリケーション (左) と距離特性 (右)

ログ同期レプリケーションは、データベースの更新ログの書込み特性を活用することにより、通信遅延が処理性能へ与える影響を

軽減し、同時に、負荷変動を含む環境変動にロバストな安定運用を可能とする。回線エミュレータと商用ストレージ装置を用いた実験を行い、トランザクション処理ベンチマークに関して東京・大阪間相当の通信遅延の下、提案手法では性能低下を高く1割程度に抑制することを明らかにした(図1右)。又、負荷変動に対して応答時間の悪化を抑える高い性能安定性を実現出来ることを明らかにした。(図2左)。

ログ同期レプリケーションの下では、バックアップシステムのストレージに格納されたデータベースは、常に一貫性を有することが保障されている。このようなトランザクショナルストレージとしての特性により、災害時の復旧時間を大幅に削減することが可能となる。さらに、ログ同期レプリケーションはデータベースの更新ログのみを転送することから、自ずと従来の全てのデータを転送する方式と比較して通信量が大きく削減される。本研究開発で実施した実験では、6割程度の削減効果を確認した(図2右)。安価な低帯域回線などの活用も期待され、ディザスタリカバリシステムの広範な普及が期待される。

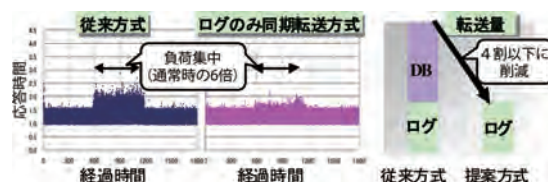


図2 性能安定性 (左) とトラフィック削減効果 (右)

3. ストレージ超高速アクセス機構

爆発的に増大し続けるデータの戦略的な活用するためには、飛躍的なIO性能の向上が必須である。本研究開発では、サーバ上のDBMSが有する高レベルのソフトウェア実行情報とストレージの内部挙動情報を組み合わせて活用するストレージアクセスの新しい高速化手法を開発した(図3)。

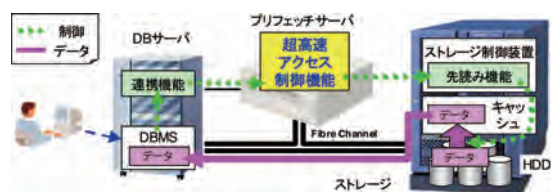


図3 クエリプラン利用先読み

本研究開発では、DBMSがクエリの処理に先立ち生成するクエリプランをストレージ装置が活用することにより、参照データを事前に高精度で予測し、先読みを行うクエリプラン利用先読み方式(プロアクティブキャッシング)を考案した。先読み方式は一次先読みと二次先読みから構成される。一次先読みでは、クエリプランと参照データに関する情報に基づき、クエリがアクセスする索引

データを解析し、以後に参照されるレコードを予測し、当該レコードの先読みを要求する。二次先読みでは、一次先読みによる予測結果に基づいて更なる索引データの解析を行い、より深いアクセス予測と先読み要求を行う(図4)。

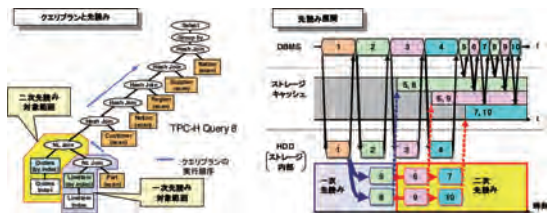


図4 クエリプラン利用先読みの方式

商用ストレージを用いてクエリプラン利用先読み方式のプロトタイプを開発し、TPC-HベンチマークのクエリQ8を対象に評価を行ったところ、約7倍の性能向上を確認した(図5)。また、当該先読み方式は精度が極めて高いことから、かなり少容量のストレージキャッシュで十分な性能向上を達成することが確認された。加えて、多重クエリ処理に関する評価も実施し、マルチユーザ利用時における有効性も確認した。

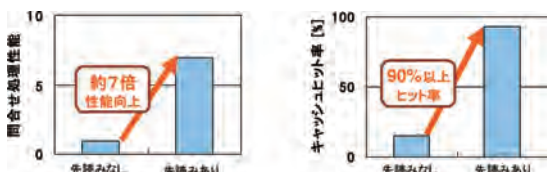


図5 クエリ性能(左)とキャッシュヒット率(右)の向上

4. ストレージ管理コスト低減機構

データ量が爆発的に増大し、ストレージシステムの規模と複雑性が拡大する中で、システム管理者の増員は困難であることから、管理者あたりの負担は急増している。本研究開発では、種々の新しい管理コスト低減機構を開発した。

一般に、データベースでは更新によって構造劣化が生じ、性能を低下させる恐れがある。管理者は、生じた構造劣化に応じて適時にデータベースを再編成し、性能を管理する役目を負っているが、このような構造劣化の管理は容易ではない。本研究開発では、データベースの構造劣化管理をストレージシステム内で自立的に行う自己再編成ストレージを考案した。高精度のオンライン構造劣化モニタリング方式を開発するとともに、ストレージレベル分離型オンラインデータベース再編成方式を確立し、商用及びオープンソースDBMSを対象にプロトタイプを実装した。TPC-CおよびTPC-Hベンチマークを用いた実験により、高精度な構造劣化の計測が可能であると共に(図6左)、自立的なデータベース再編成の有効性を明らかにした。また部分再編成なる新しい手法も確立した。自己再編成ストレージはストレージシステムに内在する高いIO処理能力を活用する事が可能であり、サーバ上のソフトウェアによる再編成と比較した実験では約1桁の高速性が確認出来た(図6右)。構造劣化管理に付随する管理業務の大幅な軽減を実現出来ることを明らかにした。

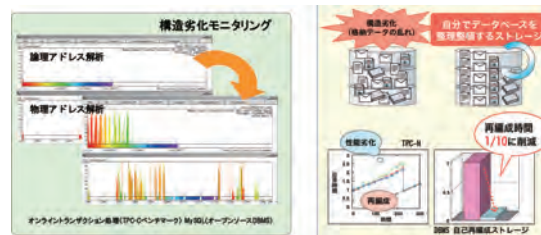


図6 構造劣化監視(左)と自己再編成ストレージ(右)

近年の仮想化と共有化が進化したストレージ環境では、システムソフトウェアは多数のレイヤから構成されており、性能ボトルネックの同定は困難を極めていた。本研究開発では、DBMSとストレージに渡る多レイヤの監視を可能とする統合モニタリング方式を考案した。データベースクエリと物理ディスクドライブの稼働率を対応付けることなどが可能となり、実験ではデータ配置の改善に要する時間を約3割低減できるなどの結果を得た(図7左)。又、DBMSでは排他待ちにより性能問題が最も複雑で解決に時間を必要とすることから、排他待ちイベントを補足しトランザクション単位のバックトレース解析を可能とする解析ツールを開発し、問題解決時間を大幅に短縮化可能であることを明らかにした(図7右)。

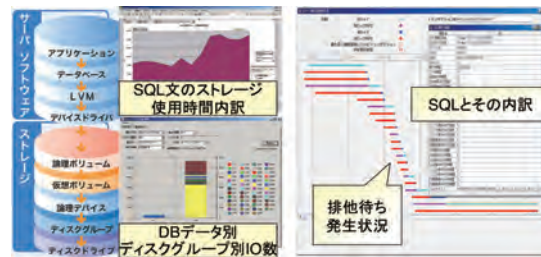


図7 統合モニタリング(左)と排他待ち診断支援(右)

5. まとめ

平成15年度より5年間に渡り、ストレージをサーバ上のデータベース・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を実現するための研究開発を実施した。高度ディザスタリカバリ機構においては、ログ同期レプリケーション方式を考案し、データ保護と安定したオンライン性能確保の両立を実現した。平成16年度および18年度には、協力企業により当該方式の製品化を実施した。ストレージ超高速アクセス機構においては、DBMSが有する高レベルのソフトウェア実行情報とストレージの内部挙動情報を活用するクエリプラン利用先読み方式を確立し、商用ストレージにおいて大幅な性能向上が達成されることを確認した。また、ストレージ管理コスト低減機構においては、自己再編成ストレージシステム、統合モニタリング方式と排他待ち診断支援方式を考案した。平成17年度と18年度には自己再編成ストレージシステムに関する論文が学会から表彰されたほか、平成17年度に協力企業により排他待ち診断支援方式の一部製品化を実現した。次世代のITシステムを支える先進的なストレージ技術として多くの高い成果が得られた。

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予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 解析技術の創出を目的とし、Socio-Sense なるシステムの開発を行った (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た実証システムを構築し、企業活動に対する消費者の反応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 Web マイニングの有効性を示すため実証実験を行った。



図 1 実世界と Web の転写構造

2. Socio-Sense システム

2.1. Web アーカイブ基盤

Web 解析の基盤として日本語 Web ページを中心とするアーカイブの構築を進め、プロジェクト開始時点の保有データと併せ、約 9 年分の情報を蓄積するに至った。Web を網羅しつつ詳細な時間変化の追跡を可能にするため、Web ページ毎に更新頻度を推定する手法を新たに開発し、最適な時間分解能で蓄積を行うようにした。最短で 1 日毎の Web ページの変化を捉えている。また、蓄積されたコンテンツはページ単位で閲覧されるだけでなく、Web 時空間分析を適用するために一括してアクセスされる。そこで、任意時点での Web 空間全体の切り出しや、大量コンテンツの並列解析処理など、Web 時空間分析との機動的連携を可能とするプラットフォームを構築した。

2.2. Web 時空間分析

互いに関連する Web ページはハイパーリンクによって密接に結合される傾向にあり、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は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システムを構築し、評価実験によりその有効性を確認した。

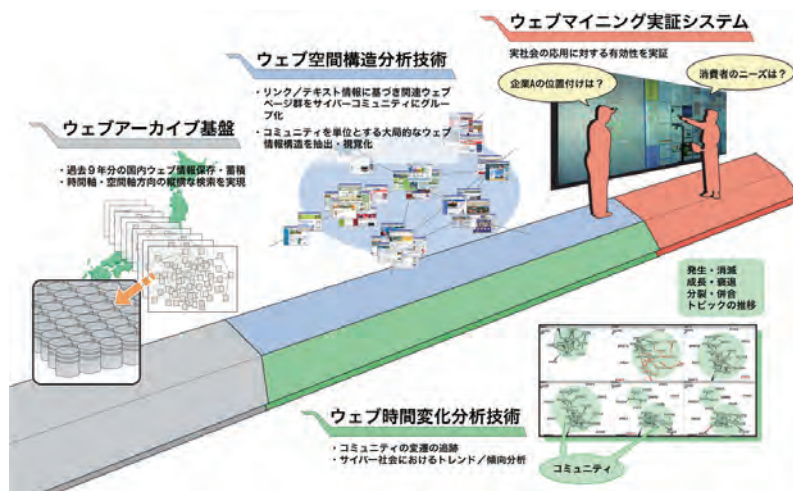


図 2 プロジェクトの概要

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空間構造の変化、および成長ステージの把握を可能にすると共に、ブログ空間などの時間粒度の細かい空間構造変化を追跡可能な従来に無い可視化手法の開発も行った(図5)。

2.3. 高度 Web テキスト解析

Web解析を企業等が利用する際には、当該企業やその製品・サービスに関する客観的な事実よりも、主観的な意見や評価などの記述が重要となるため、評価情報をWeb上のテキスト情報全体から自動的に抽出するシステムを開発した。Web文書集合全体か

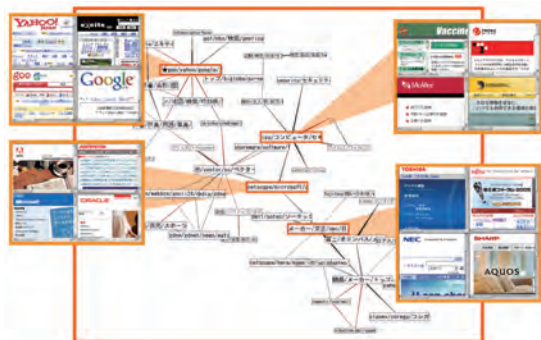


図3 Web空間の構造俯瞰



図4 Webの時系列分析

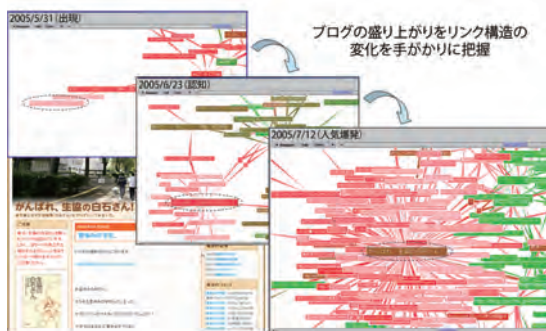


図5 Webの時空間分析

ら評価表現辞書を自動構築することにより、大規模な評価辞書の構築を行い、幅広い評価表現の抽出を可能とした。

また、多様な文書が混在するWebテキストにおいては、語の多義性の問題により、解析に必要な文書が十分に抽出できないケースが多い。これを解決するため機械学習を用いた多義性解消手法を開発し、抽出文書群の適合度を保ちつつ、再現度を大きく向上可能であることを確認した。

3. Webマイニング実証実験

Socio-Senseシステムの有効性を実証すべく、企業における消費者動向分析への適用を想定し、当該分野の専門家(専修大学新井教授、株式会社電通)と共同で実証実験に取り組んだ。同時に、社会学研究への適用を試みるため、お茶の水女子大学ジェンダー研究センターとの共同研究も行った。消費者動向分析の一例として競合ブランド間でブログ記事からの各ブランドサイトへのリンクを比較したものを示す(図6)。大々的な広告により短期間に知名度が向上したブランドや、長期間に亘って口コミで広まってきたブランドなど、それぞれの特性を明確に捉えられることを明らかにした。また、書き込み内容に対するテキスト解析により、商品や作品に対する好不評、および社会問題や不祥事に対する風潮などを把握することができた(図7)。

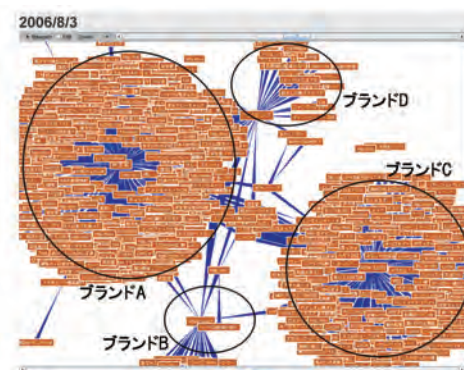


図6 競合ブランド間の被リンク比較

年	言及内容
2007-10	温暖化:深刻 環境問題:深刻 紅葉:遅い データ:不正確 天気:良い 品揃え:豊富
2007-11	紅葉:遅い 環境問題:深刻 効率:よい 地球温暖化:深刻 洗濯槽:清潔 色付き:悪い 温暖化:深刻 暖房:強力
2007-12	エネルギー利用効率:悪い 政府:悪い 地球温暖化:深刻 天気:良い 行方:不明 値段が高い 人:平和 利用効率:悪化 電気代が高い 環境問題:深刻 地球温暖化:悪化 ゴミが減る
2008-01	地球温暖化:深刻 寒さ:厳しい 温暖化:心配 ガソリン:安い 税率維持:不可欠 環境問題:深刻 温暖化:怖い 温暖化:深刻 地球温暖化:心配

図7 環境問題に関する言及分析

4. まとめ

Web情報の大局的な構造と長期に亘る変遷を分析可能にするSocio-Senseシステムの構築を行い、企業の消費者動向分析を始めとする種々の応用を通じてその有効性を実証した。今後、企業活動や政策決定におけるサイバー社会の影響力は強まり続けると考えられ、本システムの重要性は極めて高いものと信ずる。

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, 誰でもが気軽に, 情報機器の利用ができるようになり, デジタルデ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. 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 (NAM: Non-Audible Murmur)」の音声認識・合成の研究開発を進める.

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**研究開発ワークショップ,(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. (ii)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で高める. さらに, (iii)では新しく開発されたブライント音源分離(BSS: Blind Source Separation)アルゴリズムによるハンズフリー音声認識の研究も行う.

本プロジェクトの研究開発の概要を図1に示す. さらに, 研究開

発項目ごとの成果を表1にまとめておく.

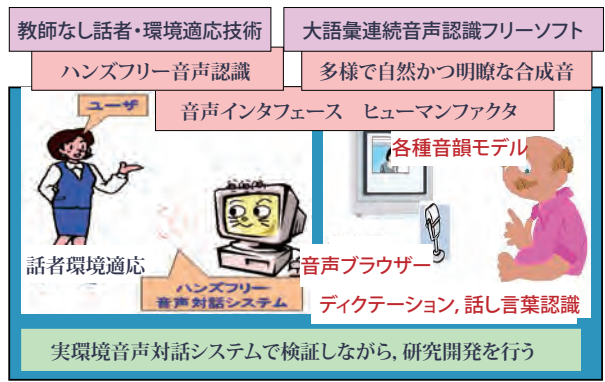


図1: 人にやさしい自然な音声インタフェース

表1: 研究開発の成果

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 非可聴つぶやき (NAM) 認識
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 大語彙連続音声認識プログラム Julius 4.0 マイコン SH-4A Julius (実時間動作)
ハンズフリー音声認識	ハンズフリー音声収録 DSP (SSA) BSS 音源分離オンラインプログラム (BSSA)
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システム構築キット 自動車内音声認識プログラムキット
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 高精度音声変換プログラム (STRAIGHT) HMMベース音声規則合成システム (HTS) 無音声電話プログラム

(青字で書かれた成果は, 当初計画になく, 追加した成果)

3. 研究成果の概要

当初計画以上に研究が進展して, 表1の青字で示したように多くの当初計画以外の成果も達成できた. また, 研究成果が国内外から認められて多くの賞を受賞した.

3.1. 研究成果のまとめ

このプロジェクトで得られた成果を箇条書きにまとめておく.

- (1) 「たけまるくん」などの多くの音声情報案内システムを実環境で運用して, それらの有効性を示すとともに, 幼児から年寄りまでの音声データの収集や話者適応技術により, 音声認識技術の性能の向上を達成した.
- (2) 歪みなしのブライント音源分離と実環境音声データによる音韻

モデルにより、1m以上離れた位置からのハンズフリー音声認識の性能を、接話マイクと同等のレベルまで高めた。

- (3) 話し言葉処理、雑音に頑健な手法などをまとめ上げて、大語彙音声認識プログラム Julius v4.0 を完成した。Julius による講義や議会のディクテーションの性能を向上させ、音声認識の分野を広げた。さらに、Julius のマイコン SH-4A への実装も進め、2万単語の実時間での大語彙連続音声認識を達成した。
- (4) 音声合成の分野では、音声分析合成アルゴリズム STRAIGHT の大幅な改造を行い、精度の向上とともに大幅な計算量の削減を達成し、実時間の高品質音声モーフィングが可能になった。HMM を用いた音声テキスト合成システム HTS を公開して、国際コンペティションでの性能の高さが世界に知られ、世界に普及して多くの研究機関や企業で使われるようになった。
- (5) 新しい静かな音声メディアとして、非可聴つぶやき (NAM) や、歪みなしのブラインド音源分離 (BSS) の原理などの発見により、静かな場所からやかましい場所でのハンズフリー通話や音声認識が可能になり、かつ、NAM による発話障害者の発話補助や、BSS による両耳補聴器の可能性が見出され、音声によるユニバーサルコミュニケーションの研究領域を大幅に広げることができた。

3.2. 研究発表および受賞状況

この5年に、約100件の学術論文、約300件の国際会議論文、約500件の国内大会での発表を行った。受賞の件数も約30件におよび、多くの報道発表もあり、非常に活発な研究活動を行うことができた。以下、主な受賞状況などをまとめておく。

ハンズフリー音声認識の研究では、歪みなしのBSSの原理の発明と実時間処理および企業との共同での装置の開発で、電子情報通信学会の論文賞、IEEE/IROS Best Application Paper Awards、音響学会の独創研究奨励賞板倉記念などを受賞して、国内外で高く認められ評価された。また、2チャンネル雑音分離の国際コンペティションでも優勝した。

非可聴つぶやき (NAM) では、電子情報通信学会の論文賞および最優秀論文賞 (猪瀬賞) を受賞した。また、セキュリティの分野でも SCIS2006 論文賞を受賞した。

音声対話システムでは、たけまるくん、パス運行システム、京都案内システムが高く評価され、情報処理学会山下記念研究賞を合計3回受賞した。

音声合成でも、音声分析合成システム STRAIGHT のデファクトスタンダードとしての地位が強固なものとなるとともに、音声テキスト合成システム HTS の汎用性と優秀さも、国際コンペティションなどで実証され、世界中で利用されるようになった。

リーダーの鹿野は、これまでの研究が評価され、IEEE はじめ情報処理学会、電子情報通信学会の Fellow になった。

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。さらに、新しい静かな音声メディアとして、非可聴つぶやき (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. 研究成果の概要

実環境での音声情報案内システムで技術の検証を行いながら研究を推進した。音声情報案内システムの運用で、情報処理学会の山下研究賞を受賞した。さらに、新しい静かな音声メディア非可聴つぶやきを発見して、電子情報通信学会論文賞と猪瀬賞、セキュリティの分野で SCIS2006 論文賞を受賞した。以下、成果を簡単にまとめる。

- (i) 音声情報案内システム「たけまるくん」を5年間、図2の「キタちゃん」と「キタロボ」を2年間運用し、幼児から高齢者まで利用可能な音声対話システムを実現した。音声データベースも整備して公開の準備中である。
- (ii) より広範な話者が音声対話システムを利用できるように教師なし話者適応オンラインプログラムを完成した。
- (iii) 既存の音声データベースを活用できる十分統計量に基づく音声データ選択アルゴリズムを考案して、音韻モデルの性能の向上と作成コストの削減を達成した。
- (iv) 新しい静かな音声メディアとして非可聴つぶやき (NAM) を発見して、声を出さなくても利用できる無音声電話、無音声認識の可能性を示した。また、統計的音声モーフィングの研究と融合して、発話障害者補助の分野の研究を立ち上げた。

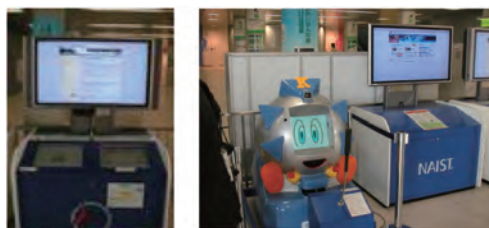


図2： 駅に設置した音声情報案内システム「キタちゃん」(左)と「キタロボ」(右)

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に、マイコンへの実装も行う。これまでのディクテーションシステムが主に成人の読上げ音声を対象としていたのに対して、多様な話者の話し言葉音声を指向して発展させる。

2. 研究成果の概要

話し言葉音声認識の研究が進展して、成果を Julius4.0 などにまとめて公開した。主な研究成果を以下にまとめる。

- (i) 音韻モデルにおいて多様な話者に対応できるように、話者正規化 (CVN, VTLN) や話者適応 (SAT+MLLR)、及び識別学習 (MPE) を実装・評価した。
 - (ii) Webテキストから対象ドメインに合致した話し言葉調の文を自動選択することにより、音声対話システム向けの言語モデルを効率的に構築する方法を提案し、ツールとして実装した。
 - (iii) 音声認識エンジン Julius の音声対話システム向けの機能強化 (音声検出・非音声棄却など)、及び性能改善 (高速化・省メモリ化) を行い、Julius 4.0 を公開した。
 - (iv) SH-4A マイコンに Julius を実装し、2万語彙の連続音声認識の実時間動作を実現した。
 - (v) 京都市バスの運行情報を電話音声で案内するシステムを、4年間にわたって一般公開・運用し、約1万対話を収集した。
 - (vi) 京都の観光情報を案内する音声対話エージェントを構築し、京都大学博物館で3ヶ月にわたり運用した。
- 上記に加えて、ソフトウェア普及のための講習会を毎年行った。

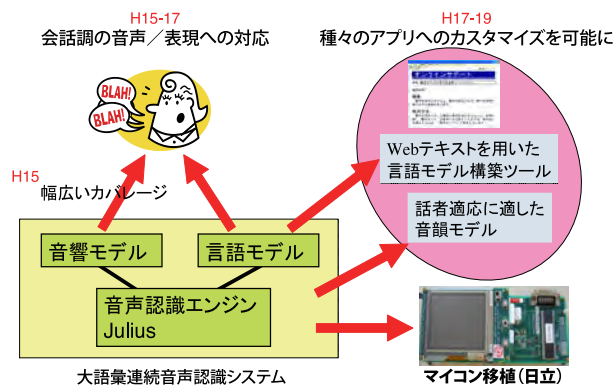


図3：大語彙連続音声認識プログラムの開発概要

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、比較的コンパクトかつ廉価なマイクロフォンアレイシステムを用いて、音声認識性能の向上を目指す。

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イを用いて開発する。さらに、マイクロフォンアレイのコストを下げるため、ハンズフリー音声収録用DSPを開発する。

3. 研究成果の概要

マイクロフォンアレイによるハンズフリー音声認識システムの実現を、新しい歪みなしのブラインド音源分離 (SIMO-ICA) の発見により大幅に性能を向上させることができた。このSIMO-ICAは国内外の学会でも高く評価され、多くの賞を受賞した。主な研究成果を以下に示す。

- (i) 効率的かつ高精度な雑音抑圧アルゴリズムとして、空間スペクトル演算アレイ SSA (Spatial Subtraction Array) を新たに提案し、その実環境評価および改良を行った。
 - (ii) 音声の特徴量および空間的情報を考慮した新しい「重み付き CSP 法に基づく実時間方位推定・発話検出法」を提案し、その実環境評価および改良を行った。
 - (iii) DSPモジュール上に上記(i)(ii)に基づく実時間 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 (図4参照)。
 - (iv) 音源間の独立性のみに基づいて分離を行う歪みなしのブラインド音源分離 (BSS) を発見して、さらに処理の DSP 実時間処理実装・高精度化を行った。
 - (v) 通常室内や駅などにマイクロフォンアレイおよびロボット音声対話システムを設置し、実環境における雑音データを収録および評価システムを構築した。
- 以上により、ハンズフリー環境における高精度リアルタイム音声認識システムの開発に成功した。



図4：マイクロフォンアレイ及び SSA 用実時間 DSP モジュール

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や各種モデルを用いた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におけるヒューマンファクタに関するノウハウを蓄積する。

3. 研究成果の概要

実環境やネットワークを介した音声対話システムを設置して、研究を進め、実用化が可能であることを示した。また、自動車内音声認識の研究も進展した。以下に主な成果をまとめる。

(i) 音声対話サービスの構築と実環境下での運用

レストラン案内システム、楽曲検索システムを構築し、インターネットを用いたデータ収集システムの運用と自動車やバイク運転時のデータ収集実験を行い、主観評価を含めた分析により認識性能に影響を与える要因の調査を行った。さらに公共施設情報案内システムや駅構内案内システム(奈良先端科学技術大学)、京都市バス運行案内システム、コンピュータ利用案内システムや博物館展示案内システム(京都大学)を実環境下で運用し、様々な環境におけるシステム運用データを整備した。

(ii) ネットワークを介した音声認識カスタマイズ法の開発

インターネットサーバと連携して単語辞書や文法の生成とダウンロード機能及び収集データアップロード機能を備えた楽曲検索音声対話プログラムパッケージを作成しインターネットユーザ向けの大規模フィールドテストを行った。発話内容を限定した場合のタスク依存音響モデル構築手法を開発し、ユーザ毎に収集された音声データを用いたMLLRによる話者適応手法と組み合わせることで音響モデルの性能を改善した。

(iii) 自動車内音声認識プログラムの作成、運用及び評価

自動SNR推定手法の開発、単一チャンネル信号を用いた雑音抑圧手法の開発、SNRに基づいたモデル選択による認識性能改善手法の開発を行った。自動車用音声対話プログラムパッケージを作成し、それを用いて車内における楽曲検索・受聴を行う対話システムを構築した。さらに一般道路や高速道路を運転走行するドライバが当該システムを利用する大規模実験を行った。

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を基盤として、話し手の感情や話し方による声質の違いを解析・表現するためのデータベースの整備と、最終的に公開するためのプログラム開発を進めた。また、並行して成果展開のための普及活動と、新たな応用システムの開発を進めた。

3. 研究成果の概要

リアルタイム音声変換プログラムを完成し、さらにSTRAIGHT、HTSという音声合成の分野で世界を先導する技術確立して公開した。主な成果を以下にまとめる。

(i) リアルタイム音声変換プログラム

話し声を、リアルタイムで別の声質や話し方の声に変換するプログラムを開発した。STRAIGHTのプログラムの制御構造などを根本から見直すことにより、普通のPCでの実時間動作を可能とし、応用システムに容易に組み込むことのできるものとした。

(ii) 非リアルタイム高品質音声変換プログラムと応用

応用システムの開発を容易にするために、STRAIGHTのプログラムインタフェースを合理的に設計し直し、広く用いられているC言語により実装した。

(iii) 音声合成システム (HTS) への応用と国際的普及

隠れマルコフモデル (HMM) を用いた韻律制御モデルを中心とした音声合成システム (HTS) にSTRAIGHTを応用するとともに、国際コンペティションを開催した。2005、2006両年度にわたって、このシステムは圧倒的高品質の合成音声を生成できるシステムとして評価された。その他、統計的手法に基づく音声モーフィングアルゴリズムなど多数の応用システムが開発された。このように、本プロジェクトにより、世界最高水準にある多様な音声合成プログラムが、容易に利用できるものとなった。

当初計画には含まれていないが、非可聴つぶやき声による音声コミュニケーションという全く新しい応用分野が、本プロジェクトの音声変換技術を応用することで生み出された。また、継続的な普及活動を通じて、STRAIGHT、HTSは、学術・応用の分野で、実質的な世界標準となっている。このことは、本プロジェクトの成果の今後の応用展開を強く後押しする。既に数社が本プロジェクトの成果に基づいた商用システムへの応用を計画している。