1A1-OS-17a-1

アジア情報 HUB プロジェクト（第一報）

ビッグデータ時代における次世代コミュニケーションプラットフォームの実現に向けて

Asian Information HUB Project（1st Report）

岩爪 道昭[*1] 藤井 秀明[*1] 原口 弘志[*1]
Michiaki IWAZUME Hideaki FUJII Hiroshi HARAGUCHI

泥谷 誠[*1] 岩瀬 高博[*2,1] 中村 哲[*3,1]
Makoto HIJIYA Takahiro IWASE Satoshi NAKAMURA

[*1] 独立行政法人 情報通信研究機構
National Institute of Information and Communications Technology

[*2] 株式会社 神戸デジタル・ラボ
Kobe Digital Labo, inc

[*3] 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

The Universal Communications Research Institute (UCRI), NICT conducts research and development on universal communication technologies: multi-lingual machine translation, spoken dialogue, information analysis and ultra-realistic interaction technologies, through which people can truly interconnect, anytime, anywhere, about any topic, and by any method, transcending the boundaries of language, culture, ability and distance. To realizing universal communication, UCRI collects diverse information including huge volumes of web pages focusing on information from Asia. This paper introduces NICT's vision and strategies for Asian information hub as a platform for collecting, storing, analyzing large-scale information and providing advanced communication services in Big Data Era.

1. はじめに

近年，ライフログ，ソーシャルネットワークサービスや Internet of things と呼ばれるセンサ，電子タグ等と接続したネットワークから非定型かつ大量のデータがリアルタイムに生成されるようになってきた．このようなデータは，短時間のうちに数億から数百億のデータエントリー数，ペタバイト級のデータサイズに及ぶことから「ビッグデータ（big data）」と呼ばれる．ビッグデータの蓄積，解析および有効な活用によって，新しいビジネスサービスやイノベーション創出のためのフロンティアとしてアカデミアだけでなく，産業界，経済界において広く関心と期待が高まっている．

しかし，流通する情報の量やコミュニケーションチャンネルが増加しただけでは，グローバルな問題解決のための意思決定や真に人と人の意思疎通，相互理解には必ずしも直接的に貢献するとは限らない．

独立行政法人情報通信研究機構（以下 NICT）では，言語，知識，距離などの障壁を乗り越え，いつでも，どこでも，誰とでも意思疎通を可能とするユニバーサルコミュニケーションの実現に向け，多言語翻訳技術，音声コミュニケーション技術，情報分析技術，超臨場感コミュニケーション技術などの高度な情報通信技術の研究開発に取り組んでいる．

これらの各研究課題では，大規模なコーパスや各種の言語資源，ネットワーク上のコンテンツやセンサデータから得られる多様かつ大量の情報などを取り扱うこと前提としているが，各課題で個別的にデータを集積し，基盤技術やアプリケーションサービスを研究開発を行うことは，コストがかかるとともに，新しいイノベーションも生まれにくい．

本稿では，最近の ICT の潮流になりつつあるビッグデータを多言語翻訳や情報分析等の高度な情報通信技術の研究開発に戦略的活用するとともに，技術開発を通じで構築された資源やアプリケーションをオープンなサービスとして社会に還元するための知識インフラ[長尾 10，CSTP 11]としてアジア情報 HUB の戦略とビジョンを紹介する．また，同構想実現のためのインターネット上の大量の情報を効率よく集め，蓄え，用途に応じて高速に取り出すためのより基盤的なレイヤーとして現在研究開発中の大規模情報基盤について概説する．

2. ビッグデータ時代における ICT 研究の諸課題

2.1 ビッグデータの特徴

ICT 分野においては，2011 年以降ビッグデータという言葉が頻繁に出現しはじめた．その背景としては，ソーシャルネットワークサービスやスマートフォンを介してユーザから生成されるドキュメント，画像，動画，通信・アクセスログ，GPS データ，電子マネーの購入情報等に加えて，各種センサデータ，カーナビやテレマティクス等の「モノのインターネット（Internet of things）」から送受信されるデータ等も急速に増加し，膨大になってきたことがあげられる．その定義は立場によって所説があるが，単純に「大量のデータ」という捉え方だけでは不十分である．ビッグデータを特徴づけるものとしては，いわゆる「3 つの V（the three Vs）」があげられる[O'Reilly 12].

(1) Volume（サイズおよび数における大規模性）

テラバイトから時として数 10 ペタバイトというデータのサイズだけではなく，twitter にみられるような小サイズながら膨大な数のデータをいかにして処理するかが問題となる．

(2) Variety（データの多様性）

生産管理や会計データのような定型の構造化されたデータだけではなく，インターネット上のソーシャルメディアやモバイル端末等から発信される非定型・非構造の多種多様なデータをいかにして処理するかが問題となる．

(3) Velocity（処理のリアルタイム性）

時々刻々と生成される大量データをいかにリアルタイムに，あるいはそれに近い短時間で処理するかが問題となる．

既存のデータベースやアーキテクチャでは，上記の 3 つの特徴に付随する問題に十分対処できないような規模のデータをビッグデータと言うこともできる[Villars 11]．

2.2 ビッグデータの戦略的活用における諸課題

ビッグデータの戦略的な活用は，我々の生活や社会に大きな便益をもたらすポテンシャルを秘めており，実社会においてもすでに利用されて始めているが，現在は，いかに大量のデータを捌くかということに主眼が置かれているケースが多い． また，個別的な事例を積み重ねている段階で，社会の共通インフラとして活用，AI 技術のようなより高度な基盤技術による付加価値の高いアプリケーションサービスを実現していくためには，以下のような諸課題が存在する．

(1) 技術的課題

情報の質の問題：学術情報や企業内情報のように定型化されたレガシーデータ異なり，ネット上を流通する非定型なデータを取り扱うことになるため，情報の出処が不明でその内容の信頼性・信憑性の判断が困難な場合や，ノイズや SPAM が多く含まれている場合も少なくない．しかし情報収集の段階から過度な情報フィルタリングによる絞り込を行うと，重要な事項の予兆を見落とす可能性もある．玉石混淆の大量データからいかに品質の高い情報を選りすぐるかは，最終的には後段の分析やアプリケーションとも相互に影響を及ぼす重要な課題である．

大規模情報の蓄積とアクセスの課題：大量のデータを収集，蓄積し，必要に応じて高速に取り出す仕組みを実現するには，大規模な計算機インフラや NoSQL，分散処理フレームワーク等のミドルウェアが不可欠であるが，それだけですべての問題が解決する訳ではない．特に，HDD 上のデータの書き込み／読み出し（I/O）における遅延（レイテンシー）が大きなボトルネックとなる情報システム全体のスループットが上がらないという問題に直面することが少なくない．

情報間の相互関連付け：時々刻々と生成，収集される一次情報を段階的に構造化しながら，すでにデジタル化されたレガシーな情報資源と関連付け，新たな知識を抽出・発見したり，領域横断的に検索し活用するための体系的な手法の確立が不可欠である．

浅い分析と深い分析のギャップ：時々刻々発生する大量かつ非定型なデータをリアルタイムあるいは現実的な時間内で処理するためには，1次情報の段階では情報の内容に立ち入らない浅い分析となる．一方，機械翻訳や情報分析など高度な 2 次，3 次の情報処理では，情報の意味内容に立ち入った深い分析が必要となり，一般的に計算コストが高く，バッチ処理的アプローチになる傾向にある．両者のギャップを段階的に埋めるための仕組みが必要である．

(2) 社会的・制度的課題

データサイエンティストの育成：従来からの統計学や機械学習の専門家だけでなく，膨大のデータを持て余すことなく取り扱える技術者，そしてビッグデータから導き出された結果を用いて戦略的，合理的な意思決定を行い，組織を動かすマネージャやアナリストを育成する必要がある．

プライバシーの保護とセキュリティの確保：個人情報がデータ化されることが消費者としての個人に益する一方で，いかにしてプライバシーを保護するかも重要な課題である．技術的には，秘密情報の保護と活用のバランスを適切に管理し，利用価値の高い秘密情報を完全かつ有効に活用することを目的として，プライバシー保護データマイニング技術[Vaidya 06]への関心が高まっているが，今後は技術と制度両面から社会実装のあり方について検討する必要がある．

法制度が未整備： ネットワークを介して容易に組織や国境を超え，複製が可能なデジタルデータにおいて，その所有権や取得／利用方法の問題に対して，安全・安心なビッグデータ利用の法制度を整備する必要がある．欧州ではすでに法制度化の動きがあり，日本においてもそれに準拠した研究開発およびサービスの運用を求められる可能性がある．

大規模インフラの維持・運用コスト：ビッグデータを前提あるいは対象とした ICT の研究開発を推進していくためには，研究開発のインフラやテストベッドとして機能する一定規模の情報基盤を整備，運用することが不可欠となる．しかし，昨今の経済情勢による財政問題や事務および運用体制，個々の研究チームや機関単独では，研究情報基盤の整備，維持，更新することが難しくなりつつある[CSTP 11]．

3. アジア情報 HUB プロジェクト

3.1 プロジェクト構想の背景

NICT では，アジアにおけるネットワーク型音声翻訳の先端研究の推進を目的として，アジア諸国の研究機関と共同でアジア音声翻訳先端研究コンソーシアム（A-STAR: Asian Speech Translation Advanced Research Consortium）を発足させ，2009 年に世界で初めてインターネットを介して，異なるアジア言語を話す複数話者間で，旅行対話を対象にした実時間音声翻訳システムの実証実験に成功した[中村 11]． また，アジアにとどまらず世界で用いられる技術とするため，ITU-T における国際標準化活動にも取り組みネットワーク型音声翻訳のサービス要求条件と機能，およびアーキテクチャにおける要求条件の 2 件の勧告が承認されている．

一方，NICT では，研究開発を通じて構築された各種コー

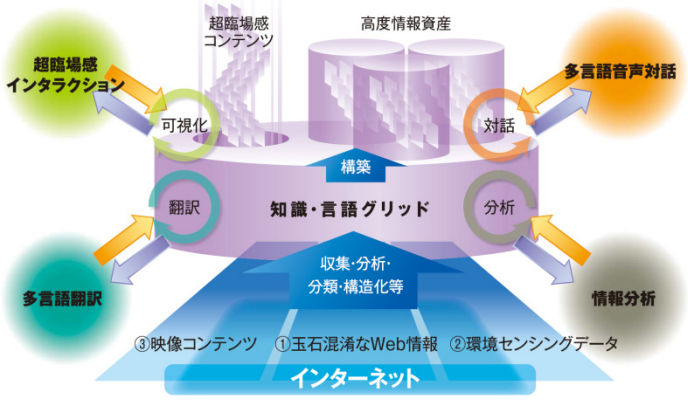

図 1 アジア情報 HUB 構想

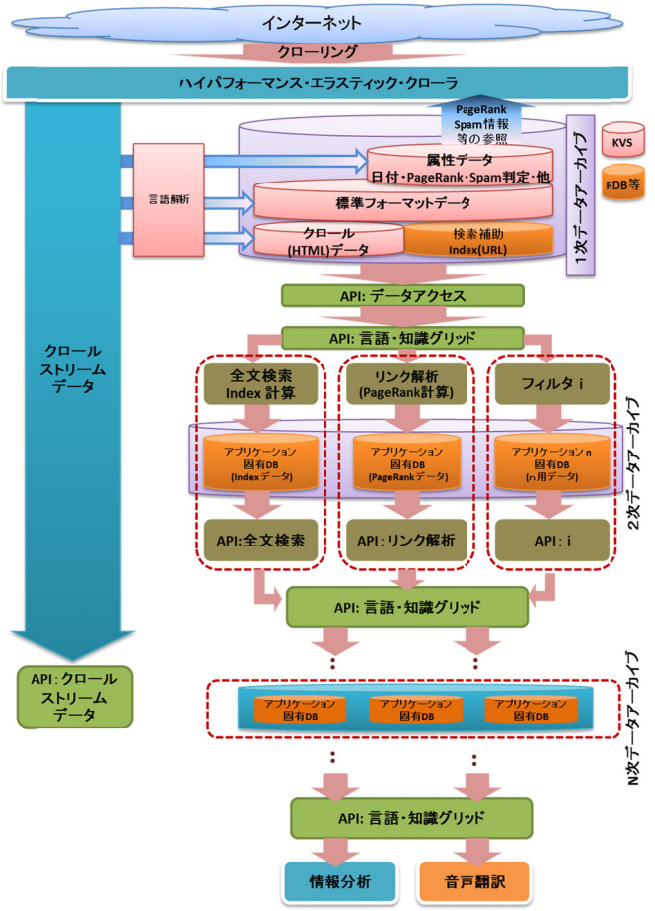

図 2　大規模情報基盤アーキテクチャ概要

パス，辞書，言語処理ツール等の成果を，産業界やアカデミアに配信する仕組みとして，「高度言語情報融合フォーラム（ALAGIN Forum: Advanced Language Information）」を産学官体制で発足させ，これまでに多数の言語資源，ツールを提供してきた[ALAGIN].

このように，NICT では，単なる先端的な基盤技術の研究開発に留まらず，その成果を広く社会に還元し，イノベーションを創出させるための仕組みづくりに取り組んできたが，これらの取り組みを加速させ，音声・言語コミュニケーション技術，コンテンツ・サービス基盤技術，超臨場感コミュニケーション技術の各技術の統融合効果を狙うとともに，ビッグデータ時代にも対応したより大規模な知識インフラの構築が不可欠であると考えている．

そこで，我々は，アジアを中心として諸外国の生活，経済，産業等に関する情報を Web，ニュース，新聞等の情報メディアより収集し，翻訳，要約，情報分析技術等により入手した情報を整理・構造化し，質の高い基盤的情報資源として**アジア情報ハブ**の構想している（図 1）．

アジア情報 HUB の情報システムとしての機能的要件としては，①映像コンテンツ，②玉石混淆なネット上の情報，③環境センシングデータなどの大量かつ多様な非定型データを収集・蓄積し，高速アクセスを可能とする**大規模情報基盤**，言語意味解析，データマイニング，3 次元画像レンダリング，検索インデクシング等の解析処理を施すことで構築される言語および音声・映像コーパス，概念辞書等の言語資源などの**高度情報資産構築・管理機能**，さらにこれらの資源と翻訳，対話，情報分析技術等を柔軟に組み合わせオープンなサービスとして提供するためのプラットフォームである**知識・言語グリッド**[寺島 10]等が想定されている．

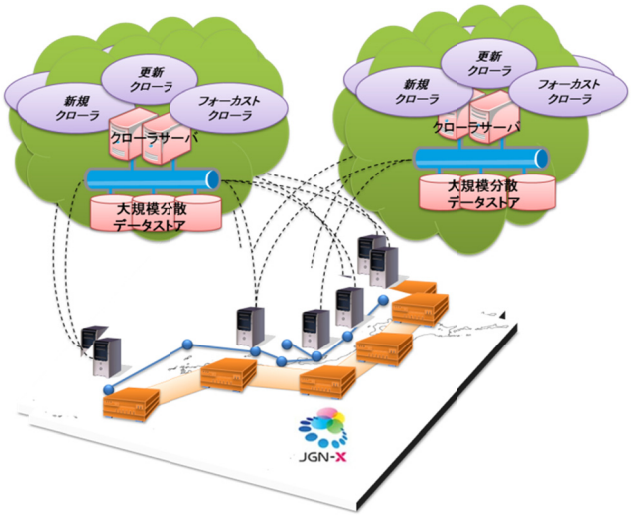

図 3　ハイパフォーマンス・エラスティック・クローラ

一方，アジア情報 HUB が有効に機能するためには，技術的な視点だけでは不十分である．情報システムがサスティナブルに機能し，社会に還元，実装されるためには，運用体制，社会展開，各基盤技術の改良・改善のためのフィードバックなどの仕組み自体もデザインする必要がある．

4. アジア情報 HUB のための大規模情報基盤

本章では，アジア情報 HUB 実現にむけ，最も低層のレイヤーとして重要な役割を果たす大規模情報基盤の概要について概説する．大規模情報基盤は，主に，①ハイパフォーマンス・エラスティック・クローラ，②大規模分散データストア（1 次データアーカイブ），③大規模計算機基盤により構成される(図 2 参照)．本稿では紙面の都合上主に，①，②の概要について述べる．

4.1 ハイパフォーマンス・エラスティック・クローラ

インターネットの普及，特に Web の普及，発展により大量の情報が容易に得やすくなり，機械学習，自然言語処理等の統計的アプローチによる情報処理技術が大きくしてきた．アジア情報 HUB においても，大量のデータ収集が大前提となっており，クローラの果たす役割は重要である．クローラの実行原理は，Web 文書のリンクを順次たどるだけの極めて単純なものであるが，近年の SNS やマイクロブログ等のソーシャルネットワークサービスの普及，高度なスパム情報や自動生成情報サイト等の増加などにより，従来のアプローチだけでは有用な情報を収集することは必ずしも容易ではない．

また情報メディアによっては，ほとんど更新されない情報源もあれば非常に更新頻度が高いものや，プッシュ型，ストリーム型の情報メディアも存在する．一方でネットワーク環境や取得先の Web サイトに過度な負荷をかけないよう十二分な配慮も不可欠となる．

NICT では，これまで情報分析システム WISDOM [WISDOM]の情報基盤として，Web クローラを開発，運用してきたが，現在アジア情報 HUB のコアの情報資源として 40 億ページ以上の Web アーカイブの構築するため，より質の高い情報を効率的に収集するためのクローラ制御の高度化，計算機やネッ

トワーク環境に応じて柔軟にスケールする非同期的な並列分散収集機能備えたハイパフォーマンス・エラスティック・クローラ[藤井 12]の開発を進めている(図 3 参照).

4.2 大規模データストア(1次データアーカイブ)

40 億ページ規模の Web アーカイブを含む数億～数百億エントリー, ペタバイト級のデータ量に及ぶ大量かつ非定型なデータを高速に蓄え, 必要に応じて取り出す機構が大規模データストアには求められる. しかし, 従来の関係データベースシステム(以下, RDBMS)は, このような用途には必ずしも適しておらず, 仮に実現しようとするとハード, ソフトともに高いコストが伴う.

そこで, 近年 NoSQL と呼ばれる新しいデータベース技術が注目されています. NoSQL は, RDBMS のように関係モデルに基づく固定的なデータ構造ではなく, データや計算機資源の増加に応じてスケールアウトしやすいシンプルなデータ構造とシステムアーキテクチャを採用しています. NoSQL は, そのデータ構造や検索方式によってドキュメント指向型, カラム指向型, キー・バリュー型などのタイプがあるが, 当該プロジェクトでは, 非定型なネット上の収集, 蓄積により適した, 国産オープンソースの分散キー・バリュー・ストア(以下, 分散 kvs) **okuyama** を採用している[岩瀬 12].

分散 kvs okuyama は, データを格納する「データノード」, 多重化されたデータノードを管理するとともに, 分散 kvs へアクセスするインタフェースを提供する「マスターノード」, アプリケーションから分散 kvs にアクセスするための「クライアント」から構成されており, データや計算機資源の増加に応じた柔軟なデータノードの追加や自動障害復旧が可能となる. また, 取扱うデータのサイズによっては, データを全てメモリ上に格納することで, 毎秒数万～数十万件のデータストリームにも対応可能な超高速なインメモリデータベースも実現が可能となる.

5. まとめ

本稿では, ビッグデータに対応した高度なコミュニケーション基盤技術およびアプリケーションサービスの研究開発とその成果の社会実装のプラットフォームとしてアジア情報 HUB の構想を提案した. またアジア情報 HUB 実現に向けて, 最も低層の基盤となる大規模情報基盤について概説した.

参考文献

[ALAGIN] 高度言語情報融合フォーラム, http://www.alagin.jp/

[CSTP 11] 第4期科学技術基本計画(閣議決定), p39, 2011.

[O'Reilly 12]O'Reilly Radar Team, Planning for Big Data A CIO's Handbook to the Changing Data Landscape, O'Reilly Media, Inc.2012.

[Vaidya 06] Jaideep Vaidya, Chris Clifton, and Michael Zhu, "Privacy Preserving Data Mining", Volume 19 in Advances in Information Security, Springer, New York, 2006.

[Villars 11] Richard L. Villars, Carl W. Olofson, Matthew Eastwood, "WHITE PAPER:Big Data: What It Is and Why You Should Care", http://sites.amd.com/us/Documents/Big-Data-WP-06-2011.pdf, 2011.

[WISDOM] NICT 情報分析システム WISDOM, http://wisdom-nict.jp/

[岩瀬 12] 岩瀬他: ビッグデータ・イン・メモリ, 2012 年度人工知能学会全国大会, 1A1-OS-17a-1, 2012.

[寺島 10] 寺島, 総務省「グローバル時代における ICT 政策に関するタスクフォース」国際競争力強化検討部会最終報告書(案), p36, 2010.

[長尾 10] 長尾, 知識インフラの構築, 総合科学技術会議基本政策専門調査会ヒアリング資料, 2010.

[中村 11] 中村他: Web 時代の音声・言語技術, 電子情報通信学会論文誌, Vol. 94. No.6, pp.502－517, 2011.

[藤井 12] 藤井他: ハイパフォーマンス・エラスティック・クローラ, 2012 年度人工知能学会全国大会, 1A1-OS-17a-1, 2012.