大阪大学でのデータサイエンス基盤構築の取り組み

下條真司[1),2)]、義久智樹[1),2)]、春本　要[2)]、石　芳正[2)]、寺西　裕一[1),3)]

1) 大阪大学サイバーメディアセンター
2) 大阪大学データビリティフロンティア機構
3) 情報通信研究機構

shimojo@cmc.osaka-u.ac.jp

Current Status of Data Science Infrastructure at Osaka University

Shinji Shimojo [1),2)], Tomoki Yoshihisa [1),2)], Kaname Harumoto [2)], Yoshimasa Ishi [2)], Yuichi Teranishi [1),3)]

1)　Cybermedia Center, Osaka University.
2) Institute for Datability Science, Osaka University.
3) National Institute of Information and Communications Technology.

概要

大阪大学ではビッグデータの高度な統合利活用と新たな知的価値の創造を目指して、2016年４月にデータビリティフロンティア機構（以下、機構）を立ち上げた。サイバーメディアセンターは、そのためのインフラであるデータサイエンス基盤を整備する役割を担うとともに、「サービス創出・支援部門」として、新たな価値を持ったサービスの創出に取り組んでいる。本稿では、現在取り組んでいるデータサイエンス基盤構築について紹介する。

1 はじめに

大阪大学ではビッグデータの高度な統合利活用と新たな知的価値の創造を目指して、2016年４月にデータビリティフロンティア機構（以下、機構）を立ち上げた[1]。ここでは、「データビリティ」すなわち、データを持続可能かつ責任を持って利活用する新たな科学の方法を探求するため、人工知能をはじめとする高度な情報関連技術を駆使し、生命科学、医歯薬学、理工学、人文科学など広範な学際研究を推進することを目標としている。サイバーメディアセンターは、そのためのインフラであるデータサイエンス基盤を整備する役割を担うとともに、「サービス創出・支援部門」として、新たな価値を持ったサービスの創出に取り組んでいる。本稿では、現在取り組んでいるデータサイエンス基盤構築について紹介する。

2 超スマートキャンパスプロジェクト

機構では、学内にカメラやレーザーレンジセンサーなど様々なセンサーをばらまくことで、学内の人々の活動状況をセンシングするとともに、それにより省エネや安全安心などの様々なサービスを提供することによって、新たなイノベーションを創出することを狙った超スマートキャンパスプロジェクトを推進している。我々は、これまでの研究開発で培って来た様々なセンサー技術を統合して、これに当たっている。

本計画の第一弾として、すでに産業科学研究所周辺ではカメラの設置が終了している。現在は、工学部の新たにリノベーションをした福利厚生棟とその周辺の整備を行なっている。また、豊中キャンパスのグラウンドには、無線メッシュネットワークによってグラウンド全体をカバーするwifi環境を実現し、医学系研究科を中心として進められているスポーツ庁による「スポーツ研究イノベーション拠点形成プロジェクト（SRIP）」と協調し、様々なセンサーによるスポーツ活動の計測が計画されている（図１）。

九州大学ですでに先行しているP-sen[2]や米国シカゴの”Array of Things”[3]などを参考にしながら設置するセンサーノードの設計を進めている。特に、センサーの設置にあたっては、カメラ、レーザーレンジセンサーなどが計測の所用条件を満たすと同時に、周辺の景観を乱さないようにする配慮や風雨に対する耐性も踏まえて、設計を進めている。

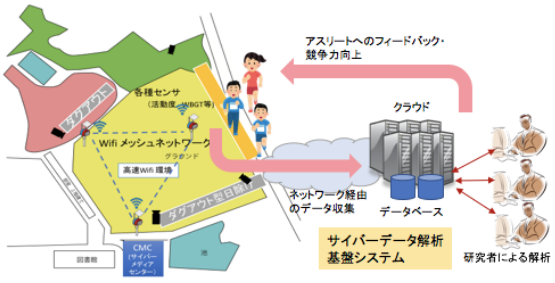

図１　豊中グラウンドにおけるスポーツ計測

設置されたセンサーの管理やエッジにおける特徴抽出などが可能なように、データビライザと呼ぶセンサ集約装置の開発を進めている。図 2にデータビライザと呼んでいる現状の設計案を示す。POE給電を利用して、様々なセンサーをつなぎこみ、備え付けられた小型コンピュータで、簡単なデータ処理や暗号化、データ伝送を行う。小型コンピュータがLED等の複数の出力装置を制御できるように設計しており、利用者に簡単な情報提示を行える。小型コンピュータや、センサーへPoE給電することで、給電に伴う配線を省略し、容易に設置できる。図 2の右側に、ポール内にレーザセンサと小型コンピュータを設置した場合の設計例示す。配線ケーブルが通行人の邪魔にならないよう、ポール下端から配線することを考えている。

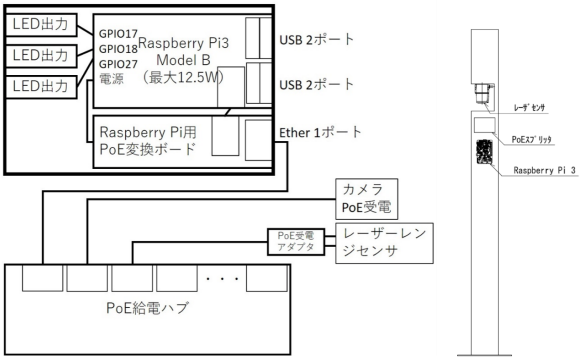

図２データビライザの設計案

今後、データビライザの詳細な設計を詰め、動作検証、システム構築を考えている。

3 データビリティ基盤システム

センサーより取得したデータを蓄積し、深層学習などを用いた分析を行い、また、場合によっては長期保存を行うためのデータビリティ基盤システムの整備も同時に進めている。これは、図 3 のような取得したデータを高速に分析する分析コンピュータ群と、それらのデータを保存するストレージシステム、そしてこれらを結ぶ高速ネットワークから構成される。

データ分析用コンピュータとして、汎用的なデータ処理を担う 24 台の仮想サーバホストと主に深層学習によるデータ分析処理を行う深層学習用コンピュータを3台備える。仮想サーバホストでは、大阪大学構内の各地に設置されたセンサーからの継続的な情報取得を行うとともに、さまざまなデータビリティ参画プロジェクトの要望に応じて仮想サーバ環境を提供することも行う。これにより、参画プロジェクトはそれぞれ自由にデータ分析環境を仮想サーバ上に構築し、柔軟なデータ分析を行う事ができる。他方の深層学習用コンピュータにおいては、 Tesla P100 データセンター GPU を 8 機搭載し、170 TFLOPS の演算処理性能を持つ高速深層学習用コンピュータを 1 台、残る 2 台は同 GPU を 4 機搭載する深層学習用コンピュータとなる。

ストレージシステムにおいては、分析コンピュータにおけるデータ分析に十分なデータ格納容量と高速性を提供する大容量ストレージと、長期間のデータ保存を低コストで行える大容量光ディスクアーカイバを備える。大容量ストレージは、約 1.2 PB の物理容量を持つ高速ストレージシステムを導入し、先述した仮想サーバに対し、任意サイズのデータ保存領域を提供する。このデータ保存領域は、センサーより得られたデータの一時格納や、データ分析時の中間データ格納、仮想サーバのディスクイメージといった動的なデータの格納に汎用的に利用できる。また、大容量光ディスクアーカイバは、 435TB の保存容量を持ち、センサから得られた観測値のオリジナルデータや、データ分析により得られた処理結果データといった、更新される可能性が低いデータの長期保存を担う。データの記憶媒体として Blu-ray ディスクの特性を継承した Archival Disc カートリッジを用いているため、50年以上にわたるデータ保存寿命を実現するとともに、データの保持に電力を必要しないため長期間にわたるデータ保管を低消費電力・高信頼・低コストで実現できる。

これらの装置を結ぶネットワークは、装置間のデータ通信に対して十分な高速性を備えるとともに、プロジェクト間の通信分離やセンサーへの通信制御、遠隔地のセンサーに対する通信暗号化など、高速かつセキュアな通信基盤を提供する。

以上のデータビリティ基盤システムにより、大阪大学構内の各フィールドで実施されるデータビリティ実証実験で得られる多種多様なデータのセキュアな保管と、それら多様なデータを融合した多角的な解析や、多量のデータのより高速な分析により新たな価値の発見や知識の創生をサポートする。

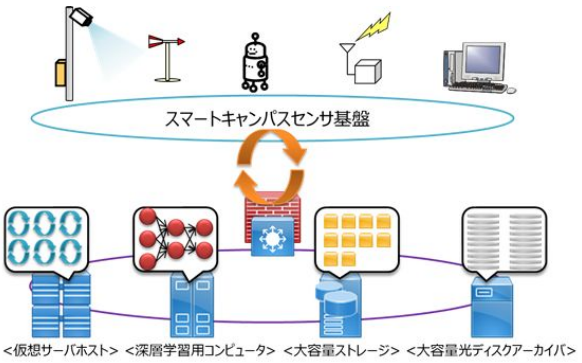

図３　データビリティ基盤システム

4 システムの要件

このようなスマートキャンパスのためのセンサー情報を収集するインフラについては、以下のような要件があると考えている[4]。

1) 迅速性（Agility）
従来、情報の蓄積にはRDB(関係データベース）が使われてきた。RDBでは、最初にデータベースの構造を定義し、収集したデータをその構造に従って形式化し、蓄積していく。しかし、スポーツ医科学のような発展途上の研究においては、当初から構造を決めることが難しい場合が多く、研究が進むに従って構造を柔軟に変化させることが求められる。従って、構造を予め定めずにデータを蓄えていき、利用方法とともに構造を発展させていくような柔軟性をもったデータ管理が有効であると考えられる。例えば、Elasticsearchのような、あらかじめ構造を与えなくてもデータを蓄積して検索することができるデータストアとRDBとをうまく組み合わせ、それぞれの利点を活用できるようなデータ蓄積基盤が有効であると考えている。また、データを利用するアプリケーションも、プロジェクトの進行とともに容易に進化していくことを支えることのできるインフラでなければならない。
2) 多様なステークホルダーからのアクセス (Access control by multi-stake holder)
プロジェクトに関わる人は、様々な研究グループからやってくる。センサーデータはカメラ画像や心拍など、いわゆるパーソナルデータも多く、情報にアクセスする人の属性や情報提供者の同意の状況に応じて適切に情報の流れを制御していく必要がある。
3) セキュリティバイデザイン (Security by Design)
常に進化していくシステムであり、多様なステークホルダーがアクセスするシステムであればこそ、最初からセキュリティ要件について考えたシステムデザインを行っておく必要がある。また、設置して終わりではなく、常にシステムの状態を監視し、セキュリティに対処する運用が必要である。
4) ストリームとデータの融合
センサー情報などの多くは時間的に連続したデータであり、これらをストリームと呼ぶ。これに対して、通常のデータは単発であり、時間的な規則性はない。これら両方のデータを統合的に扱う必要がある。

現在、このような要件を考えながら、システムの設計を行なっているところである。基盤システムとしては、クラウドによる仮想化、SDN (Software Defined Network)によるネットワーク仮想化、オープンソースの活用を最大限行うことで、迅速かつ軽いインフラ整備を目指している。

参考文献

[1] http://www.ids.osaka-u.ac.jp/
[2] 九州大学共進化社会システム創成拠点、「都市OSの創り方1　P-Sen」, coi.kyushu-u.ac.jp/contents_designer/widgets/file.../Topic/.../都市OSの創り方1
[3] https://arrayofthings.github.io/
[4] 下條真司、義久智樹、春本要、石芳正、寺西裕一、「ビッグデータスポーツ医科学のための情報インフラ構築」、大阪大学医学部学友会会誌2017、平成29年12月（掲載予定）。