Cyber-Physical Systems (CPS) と大規模データ解析

秋　岡　明　香[†1]

Cyber-Physical Systems (CPS) が提唱された当初，想定される環境は組み込みシステムが構成するセンサーネットワークと，センサーネットワークが提供する位置情報などの比較的単純な情報を解析するための計算基盤であった．その後，携帯電話やソーシャルネットワークサービスなどの発展・普及にともない，情報収集基盤も収集可能な情報も多様を極める時代となり，より高度な知識や情報の抽出が求められるようになった．本稿では，こうした高度に複雑化した環境でのディペンダブルなCPS 構築に向けた大規模データ解析技術について議論したい．

Cyber-Physical Systems (CPS) and Big Data

Sayaka Akioka[†1]

Originally, sensor networks with dependable systems, and computational infrastructures are supposed to form Cyber-Physical Systems (CPS), and the information collected in CPS was simple data such as GPS data. After a while, celluar phones, and social network have quickly grown popular, and these devices, and services boosted the descriptions, and the complexity of the information exchanged inside CPS. The change also enforced more sophisticated knowledge or information to be extracted from CPS with the higher expectations. This position paper discusses technologies for big data analyses in such a next-generation CPS to respond to these requests.

1. はじめに

Cyber-Physical System (CPS) が提唱された当初と比較して，CPS での情報収集基盤は多様なデータを大量に収集可能となった．例えば携帯電話は，利用者の位置情報，行動履歴，その時々の利用者の心情などを GPS やソーシャルネットワークサービスなどを通して収集可能なセンサーと見なすことが可能である．こうした，従来よりもリッチで大量のデータを，如何に安定して解析し，新たな知見や価値を提供できるか，という点は，実用的な CPS を実現する上で不可欠な視点である．同様の経緯を辿り，ビッグデータ解析に取り組んでいるのが，データマイニング分野である．本稿では，ビッグデータ解析と，その周辺の技術動向を整理すると同時に，CPS におけるビッグデータ解析はどうあるべきかを議論したい．

2. エッジ・ヘビー・データ

従来，データ解析に必要なデータは，データセンター等に収集・保存され，解析を行なう段階になって取り出して利用する，という形態であった．これに対して，丸山らは，CPS やビッグデータなどが主流になる現代や次世代のアーキテクチャは，ネットワークの辺縁部（エッジ）にデータが蓄積・処理される「エッジ・ヘビー・データ」の時代が来ると予測している[1)]．丸山らは,「エッジ・ヘビー・データ」の例として，監視カメラ画像，生体センサー情報，スマートフォンに保存される情報，天文台から得られる天文学データなどを挙げている．これらのデータは，1) ネットワーク経由で 1 カ所に集めるにはデータの転送コストが大きすぎる，2) プライバシー保護の観点からデータを 1 カ所で管理するのは危険である，3) 収集した全てのデータが同価値を持つわけではないので全てのデータを均等にデータセンターに蓄積することがコスト的に合わない，といった 3 つの理由から,「エッジ・ヘビー・データ」の典型例であるとしている．

こうした考え方は既に広まっており，Altera 社は「moving computation to the data」として，ネットワークやストレージに近いところにデータ処理専用の FPGA を配置し，CPU や OS によるボトルネックを避けることで大規模データ処理の高速化を図る方針

†1 明治大学
Meiji University

を打ち出している [⋆1]．IBM の Netezza は，ディスク I/O 層に FPGA を採用することで，データベースアクセスの高速化を図っている [⋆2]．

3. データ取捨選択の問題点

「エッジ・ヘビー・データ」や「moving computation to the data」の主な共通点は，データを 1 カ所に集めるコストの大きさを考慮し，このオーバーヘッドを避けるために，末端でデータの取捨選択を行ない，あるいは末端でデータの前処理を行ない，データ転送コストを下げ，より複雑かつ詳細な計算を行なうためにデータを集積し，計算基盤を用いて解析を行なおうとしている点である．しかし，この考え方は，昨今のビッグデータ解析の考え方の真逆のアプローチである．

ビッグデータ解析が注目を浴びる以前，データ解析の主なアプローチは次のようなものであった．

(1) 解析対象データについて，利用者が過去の経験や解析結果に基づいて，注目すべきいくつかのパラメータを決定する．
(2) 解析対象データが大きすぎる場合には，一部のデータをサンプリングして (1) で決定したパラメータを元に解析を行なう．
(3) 解析結果から明確に新たな知見や有用な結果が得られない場合，パラメータを変えたりデータのサンプリング手法を変えるなどして再度解析を行なう．

これに対して，ビッグデータ解析は従来のデータ解析手法の欠点を補い，従来は得られなかった知見を得ることができる解析手法として，次のような特徴があると期待されている．

(1) 計算基盤の発達により，データ解析時にパラメータの絞り込みを行なうことなくパラメータスイープ的に解析を行なうことができるケースが増えてきた．
(2) 計算基盤の発達により，データ解析時にデータをサンプリングすることなく，全データについて解析処理を行なうことができるケースが増えてきた．
(3) 計算基盤の発達や情報技術の発展により，収集できるデータの種類が増え，より詳細なデータを対象とした解析を行なうことが可能となった．
(4) 従来はひとつのドメインに閉じた解析が主流であったが，入手可能なデータが増え，計算基盤が発達したことで，ドメインをまたいだ解析が可能となった．

つまり,「エッジ・ヘビー・データ」の考え方に基づくエッジでのデータの蓄積と前処理は，データの転送コストを下げるという利点はあるが，従来のデータ解析と同様に，その解析の範囲を固定的かつ限定的にしている．さらには，末端で不要と判断されたデータは，解析される機会を失ない，ビッグデータ解析で期待される「これまでにない気づき」を提供する機会を失なうとも言える．

4. データ転送コスト再考

かつて,「壁のコンセントにコンセントを挿せば誰でも電力が得られるように，ネットに繋げば誰でも計算パワーを得ることができる」環境の構築を目指したグリッドコンピューティングという考え方があった[2]．近年では，グリッドコンピューティングのような大規模計算基盤に該当するのはクラウドコンピューティングであると考えられる．もし，グリッドコンピューティングのような身近さで，クラウドコンピューティング環境が CPS の各センサーに直結していたら，データをエッジでフィルタリングしたり前処理したりする必要があるだろうか．

データを収集する末端のセンサーとバックエンドのクラウド計算基盤が十分に近く，データ転送コストが極端に大きくない状況，あるいはデータ転送コストと比較してクラウド計算基盤での計算コストが十分に大きく，データ転送コストが隠蔽可能な状況であれば，将来必要であるかもしれないデータを廃棄してまで，データ転送コストを下げる必要はない．エッジでのデータフィルタリングが不要となる．CPS でもビッグデータ解析のメリットを利用することが可能となる．むしろ，CPS のように多様な情報を多角的かつ大量に収集する基盤でこそ，ビッグデータ解析のアプローチが不可欠であると考える．

参 考 文 献

1) 丸山宏: エッジ・ヘビー・データとそのアーキテクチャ，情報管理，Vol.56, No.5, pp.269–275 (2013).
2) Ian Foster, Carl Kesselman: The Grid 2, Morgan Kaufmann (2003).

⋆1 http://www.flaggmgmt.com/hpc/PPT/2013%20HPC%20Session%207-Nick%20Finamore%209.16.13.pdf
⋆2 http://www-01.ibm.com/software/data/netezza/