巨大データへの情報システム的アプローチ

Information Systems Approach to Big Data Handling

溝口徹夫
Tetsuo Mizoguchi

要旨
巨大データについての議論が多くなされるが、その多くは巨大データを使って何が可能かという情報技術的アプローチである。本資料では巨大データの特性と課題、特定の分野、航空管制での事例検討を通じて、情報システム的アプローチを考察する。

1. はじめに

巨大データ(Big Data)についての議論が多くなされる[1],[2],[3],[4],[5],[6]。その多くは情報技術的アプローチであり、巨大データを使って何が可能であるかの議論である。資料[5]は多少異なり、特に社会システム、プライバシーに焦点を当てているが、提案されている社会の影響モデルは必ずしも適切だと言えない。本資料では、第 2 章で、巨大データの特性と課題、第 3 章で、特定の分野、航空管制での事例検討を通じて、情報システム的アプローチを考察する。

2. （巨大）処理データに関する課題

本章では、巨大データの特徴としての 3 つの V[4]についてまずまとめを行い、巨大データに関する課題を挙げて、考察を行う。

2.1. 3 つの V(Volume, Velocity, Variety)

1) Volume について

巨大データでは、量と期間(いつまで維持するか)に関心が集まっている。量については、ここで言う巨大データは、画像データのように一個のデータの量が大きいのではなく、データ総数が大きいことが特徴と言える。期間に関する巨大データの費用と効果の評価は当然保持するデータの利用価値に依存する。費用は運用上の費用であり、データを利用可能にするための維持費用である。効果の面を考慮に入れるとすれば、効果的維持のためのデータ管理(価値評価)の費用もあるが、一般にはこれは含まれていない。欠けているのは、保持するデータの利用目的やその効果価値評価である。データ保持のための技術問題に終始している域を超えていない。

2) Velocity について

発生するデータの発生頻度が速い(多い)、遅い(少ない)という特性を指すと考えられる。データで表現しようとする実体の活動速度がそのままデータ発生頻度ではない。実体の活動はサンプル収集されるデータとは無関係に行われるものが一般的であるとしてよいであろう。ここで、データ発生頻度はデータ利用に際して必要となる最低のデータ発生頻度と考えるのが妥当であろうし、データ発生頻度は、与えられるものではなく、必要に応じて人為的に決定可能とするのが妥当であろう。

データ利用によって得られる結果が、実世界での成果の利用に合っているかの考慮が必要であることに気づく。このことは、データ利用がどのようなもので、成果である情報の要件(時間制約なども含めて)が明らかになっていることである。

3)Variety について

多様性は異なるデータ種類があり、それをまとめて一つのデータであるという特性化をした場合である。あるいは、データの整理のないままに、それをデータとして取り扱うような場合である。

例えば、メール文や Web のページの文章を非構造データとして、また巨大データとして取り扱うという場合などである。このような非構造データは内容として多様性をもち、その中から何を抽出するかはデータ利用前には予めわかっていない場合もあり得る。非構造データの持つ多様性はそれ自体別個の課題と取り扱われるべきで、何もかも巨大データとして扱うことは得策でないと考えられる。

2.2. (巨大を含めた)処理データの課題

1) 事象データと軌跡データ

巨大データを議論する場合、想定されるのは軌跡データであることが多い。軌跡データでないものとして、事象データがある。

① 事象型は特定の事象が発生したときにデータの記録が行われるもので、特定の事象は人為的にデータの特性として定義され外部から与えられるものではない

② 軌跡型は外部から与えられるものであり、定義することが主要な行為ではない。定義するとすれば、サンプリング周期や与えられたものからのデータの選択である。

2) 事後解析と実時間的制御

事象データであれ軌跡データであれ、データは過去に発生した記録である。大量の記録を事後に解析し、対象となる実体の挙動などを理解するのに使用される。データ解析によって、過去に判然としなかったことが明確になる場合もある。重要なのは実時間的制御の場合も多い。

3)データの非完備性

ここで言う非完備性は、データが欠落していることを指す。欠落には、二種類あって、一つのデータ源からのデータが一部欠落していることと、データ源がいくつか落ちていることである。ある携帯電話業者による、特定都市での住民の挙動を分析[6]した際に、その業者のサービスに加盟した利用者が電源を切っているためデータが収集されない場合が前者で、他の携帯電話業者のサービスを受けている住民の挙動はこの分析には考慮されていないというのが後者の場合である。

4)データの品質と所要性能要件

データには、特にセンサによって得られるデータには誤差がある。データの精度には次の二つの尺度がある。

- データ内容の精度(accuracy)
- データ容器の精度(precision)

大切なのは、accuracy である。

データの品質、特にセンサによるデータの場合は、データの内容の精度以外に、データがサンプリングされた時刻の誤差ないしはデータが転送されて受信した時刻の遅延もあり得る。複数のデータソースからのデータを周期的に受信する場合、複数データソースのデータ周期は同期しているわけではない。この課題は、データの実時間的利用を行う場合に注目せねばならない。事後のデータ解析を行う場合と異なることは自明である。

5)Human-in-loop の問題

データ利用が実時間的制御である場合、その制御ループに人間が介在する場合もあり得る。その場合は、品質(誤り)や所要時間などについての人的要因を勘案することが求められる。

3. 事例による(巨大)処理データの課題への情報システム的アプローチの考察

航空管制での課題を事例として取り上げ、上記課題の具体的な考察をする。航空管制の主活動は監視と制御である。その課題は、安全でかつ効率的な運航、特に、安全を脅かすことなく、いかに運航の遅延を減らすかである。特に離陸時、飛行時である。

3.1.事象データと軌跡データ

航空管制における事象データとして典型的なものは、運航データ、フライトデータなどと呼ばれ、各フライトの出発空港・目的空港・離陸時刻(計画・予測・実)・到着時刻(計画・予測・実)等からなる。著者独自のデータモデルを図 1 に示す。巨大データに限らず、この種のデータは、多種のデータソース(図

1 の例では、管制機関の各システム、各空港、国内外の航空会社、気象関係機関等)からのデータの統合からなる。データソース間でのデータ単位の不揃い、データ品質(データ欠損を含め)の不揃いに対処せねばならない。

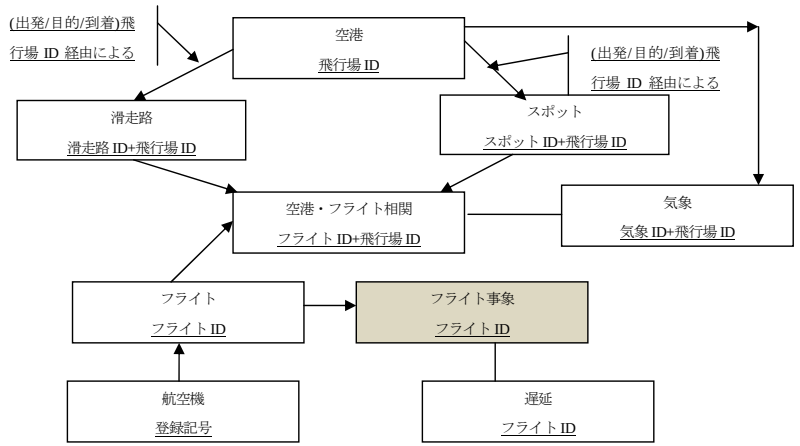

図 1 航空事象データのモデル

一方、各フライトはレーダーなどによってその位置を捕獲され、軌跡データとして利用される。ここで注意すべきは、軌跡データは事象データの一部であるフライト事象の詳細である。軌跡データのデータ利用は事象データと関連づけて有意義となる。軌跡データは各フライトの 4 次元(3 次元空間+時間)であり、データは定点観測値の場合と移動体からの一定周期の観測値がある。軌跡データの利用で大切なことは個別のフライトの軌跡のみでなく、4 次元近傍のフライトとの管制間隔の維持である。また、同じレーダーデータと言っても、空域(航空路、ターミナル領域、空港面)によって収集データの周期が異なる。

また、航空機の空港面での移動データの場合、航空機の位置情報が障害物のため、一部受信されず欠落していることがある。あるいはパイロットが出発走行前ではデータ送信器の電源を切っていることもある。また、航空機の位置情報を発信するためには、航空機に装備が必要であるが、航空機によってはその装備がないため、欠落している場合もある。前者の場合は、一つのデータ源の中での補間が必要になる。補間は時系列解析で行われている。後者の非完備性の場合は、データの利用法がデータの完備している場合と異なるであろう。管制の対象となる全ての航空機の位置情報が得られなければ管制の意味をなさないであろうが、いくつかの航空機の移動時間を知るためには、データが非完備であっても構わないこともある。

3.2.事後解析と実時間的制御

巨大データの利用には、事後におけるデータ解析によるデータ利用と実時間的な制御を行うためのデータ利用がある。巨大データを利用した事後解析による安全性の向上例が示されている[7]。多数のフライトの軌跡を管制担当官に図示する(図 2)ことで、安全に関する全体的な視野を与えるというものである。この結果を実時間的制御へと展開するという意図もあるが、本質的にはフライトの事後解析である。過去において、離陸時の地上走行は離陸機が出発するスポットから滑走路までの距離に比例すると想定(予測)されていた。しかし、この予測は余り正確ではなく、走行時間は離陸機が走行を開始し、離陸するまでに滑走路から離陸した離陸フライト数(つまり滑走路までの混雑度)に比例する傾向があることが事後解析によって判明した。また、直感的にも(見てわかるように)、離陸のための最終合流誘導路以後での走行に変動する遅延が集中していることも事後解析で判明している。このように走行時間を例にとると、走行絶対時間の大小もさることながら、走行時間の変動要因も解析の対象になる。もう一つの例

は、フライトの空港面での離陸走行時間の予測を行うために、データの分析によって走行時間に影響を与えるのは、先行して離陸したフライト数(待ち行列の長さ)が大きな要因であることが判明し、待ち行列と走行時間の相関を平均的に求めて、走行時間の予測値とするものである[8]（図 3)。これは事後解析から実時間制御のための走行時間予測を行うという試みである。但し、待ち行列の長さはフライトが離陸した後でしか得られない。そこで巨大データから統計的に待ち行列の長さとそれに対応する平均的走行時間を推定する。この走行時間は平均であり、同一待ち行列の長さでも、走行時間に変動がある。想定したいくつかの前提が将来とも成り立つかどうかも実証を必要とする。別の分析では、走行時間の変動要因は走行開始頻度そのものが過大であること、走行開始順と離陸順の逆転によることが判明している。また、パイロットは最適(最少)時間で走行していないことが多いとの分析結果も得られている。変動に影響を与える要因は、走行時間だけでも複数あることが分かる。

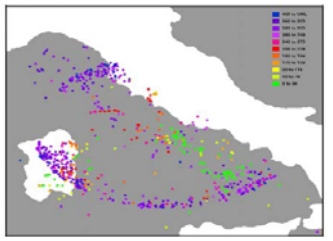

図 2　軌跡データの表示例[7]

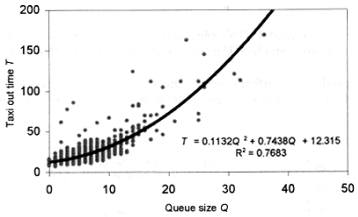

図 3　Taxi-Out 時間予測(先行フライト数と走行時間相関)[8]

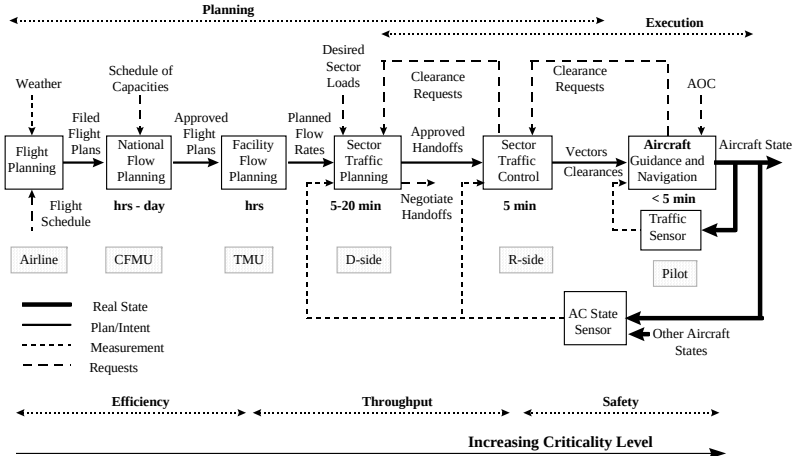

図 4　航空管制の計画・実行の段階[10]

国内での成果としては、中間報告で、「混雑空港における空港面運用の詳細なスケジューリングによって空港面交通流を管理し、空港容量の最大活用を行うことで空港面運用を効率化することを目的として、技術的な可能性を検討するために、実運航データを用いた空港面運用の分析を行って空港面交通流の特徴や滞留の発生メカニズム等を把握し、スケジューリングによる空港面交通量管理が運航効率化のための有力な施策であることを確認する」としている[9]。ここで言う実運航データ(このデータは一部の関係者にしか利用可能ではない)に含まれるのは、空港面を移動している各フライトからの毎秒単位での位置データである。但しデータ量は大きいとは言えない。ここで気付くことは、「技術的な可能性を検討」という表現である。技術的以外にこの種の課題の重要な点は、制度的制約であろう。

航空交通に類似した、道路交通での各種解析の成果が示されている[6]。発表者の言によれば、巨大データは一部の関係者のみに利用可能で、そのためもあり、解析結果が一般には評価されない、多くの人が解析に興味を持っていないという閉鎖状態である。解析者が解析結果は有効であることを繰り返し示すことによって、開けた巨大データ利用へと進むことを期待するしかないという現状でもある。

道路交通の領域で、ある種の予測を実施する試行が行われている[2]。処理するデータが大量であることもあって、予測は近似的な手法を用いている。近似的手法が悪いわけではないが、結果が妥当であることを保証することが極めて困難と思われる。これは結果の利用目的に依存することであるが、試行によって妥当であると結論付けることが保証になるかどうか疑問である。

より広い視野から、各フライトについて航空管制での計画・実行の時間進行が図 4 に示されている[10]。中ほどにセクタでの計画と管制が位置づけられ、複数の航空機の管制が行われる。R-side はレーダー側、D-side はデータ側で、完全にはデータ処理のみで管制が可能でないことを示し、更には loop に人間(パイロットと管制官)が入っていることである。計画段階では、事後解析の結果が利用され、実行段階では、実時間的制御が行われると解釈できる。

3.3. 管制のための 4 次元予測

事象データや軌跡データで与えられるものは、過去のデータである。離陸管制を行うために、事後の離陸時刻は意味がない。効率的運航を行う管制のためには、時刻を事前に予測することが求められる。例えば離陸に関しては、逐次走行準備完了する離陸フライトの走行時間を短くするには、状況に合わせた適切なスケジューリングが必要であり、また成果の評価ができる必要がある。スケジューリングの検討/実装、それに伴う走行時間推定/評価が試みられ、その成果として離陸時刻の予測法も検討されている。予測により必要に応じて、走行開始を遅らすという管制方法も提案されている。完全に実施されているものではないが、図 5 で予測の時間関係を示した。(スポットに)到着後、折返し飛行する場合は、各種の地上活動(機内清掃、燃料搭載、等)の後、走行準備完了となる。これは航空会社の活動で、管制側が走行準備完了を正確に知る方策として、管制と航空会社の間の協調的意思決定という方策も試行されている。このことで離陸時刻の予測もより正確なると考えられるが、一方走行時間を短縮するために、走行開始を遅らすという方策もあり、予測は時間的に前方、後方へと波及することもある。

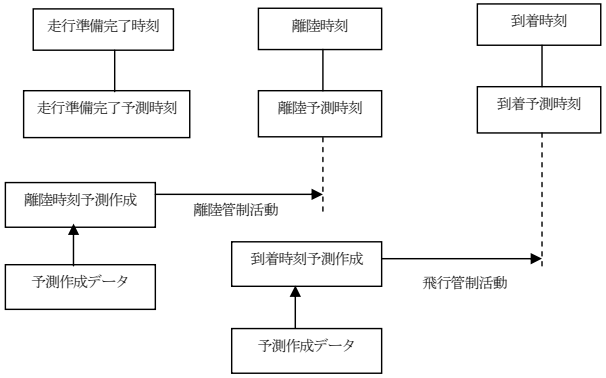

図 5　時刻予測と事前予測作成の時間前後関係

4. まとめ

別の文脈であるが、「技術は「可能なこと」と「必要なこと」を見分けて常に自らを監視しなければならない」[11]との警告がある。

巨大データで重要なのは(必要な)データ利用の目的を明確にすることであることは多く指摘されている。そのためには情報システム的アプローチ、何が必要なことなのかを探求する必要がある。これは技術的可能性を追求するよりも、対象分野での洞察を必要とし、容易ではない。更に、データ利用目的の明確化を困難にしているのは、巨大データを収集する立場では、データを収集することに責任はあるが、他の人が行うデータ利用の目的には関心を持たないし、その成果を評価することもない。従って、開かれたデータ提供が行われず、データ解析も行われず、データ利用の効果も生じない。

開かれたデータ提供が行われれば、自動的に有効なデータ利用が行われるわけでもない。対象となるデータの特性の理解が必要であろう。データから目的を選択するというよりは、データの理解を基に目的からデータの選択を行うのが本来であろうが、データと目的との双方向の関連付けを同時進行して行うことが有効と思われる。本資料でデータモデルを示したのはデータ特性の理解を強調するためである。本資料で取り上げた実時間的制御を必要とする場合は、事後解析で行われるデータの理解のみでなく、事象の事前予測、事象に影響を及ぼす要因の判定が必要になる。航空管制に関して、現在多くの国や機関で検討が行われている。

これらの予測や制御法は試行実験の実施と評価を経て(パイロットや管制官など人間が介在する場合はこれらの活動は困難を極めるし、アプローチの選択も重要な判断である)、実稼働となる。

参考文献

[1] 吉田圭吾、松田和賢「ビックデータ斜め読み　流行に惑わされないための要点と将来展望」情報処理、2012,9, 968-973

[2] 喜田弘司、藤山健一郎、磯山和彦「ビックデータをリアルタイムに処理するデータストリーム処理技術」情報処理、2012,9,962-967

[3] Tony Hey, Dennis Gannon, and Jim Pinkelman, ‘The Future of Data-Intensive Science’ Computer, IEEE, May2012, pp81-82

[4] Sam Madden, ‘From Databases to Big Data’, Internet Computing, IEEE, May/June 2012

[5] Alex “Sandy” Pentland, ‘Society’s nervous system: building effective government, energy, and public health systems’, IEEE Computer, Jan. 2012, pp31-38

[6] 「都市をマネジメントするビックデータの可能性」、情報処理学会連続セミナー「ビックデータとスマートな社会」#3, 2012.9.25

[7] Simone Pozzi, et. al, ‘Safety Monitoring in the Age of Big Data : From Description to Intervention’, Ninth USA/Europe Air Traffic Management Research and Development Seminar (ATM2011), Berlin, June 2011

[8] Idris, H., et.al ‘Queuing model for taxi-out time estimation’, Air Traffic Control Quarterly, Vol.10 (1) 1-22,2002

[9] 山田泉、他「空港面交通管理の評価手法に関する考察」電子航法研究所研究発表会、2012/6

[10] Aslaug Haraldsdottir, et.al ‘Air Traffic Management Concept Baseline Definition’, NEXTOR Report # RR-97-3 October 31, 1997, http://www.boeing.com/commercial/caft/reference/documents/coe_report.pdf

[11]赤川次郎「ネット社会の闇」図書、岩波書店、2012 年 9 月号　pp32-34

情報システム学会　第 8 回全国大会・研究発表大会 []