

サンプリングを応用したビッグデータの新処理方式 New processing method of big data by applying the sampling

高柳 健太^{*}, 山浦 恒央^{*}
Kenta Takayanagi, Tsuneo Yamaura

現在ビッグデータが注目されている。インターネットにつながる様々なコンピュータから生み出されるビッグデータは、合計すると数百テラ (1テラは約1兆) バイト以上とされており、2020年には40ゼタ (1ゼタは1兆の10億倍) バイトになると予測されるほど、データ量が膨大なため処理時間の長期化、処理のオーバーフローなど様々な懸念も指摘されている。本稿では、ビッグデータをサンプリングによりスモールデータ化、全件処理の結果が出るよりも早い段階で実用的な精度を持った「選挙速報」を出す処理方式を提案する。

Abstract: Big data generated from various computers will lead to. Internet Big data is currently of interest, which is (1 Terra is about 1 trillion) bytes or more several hundred Terra In total, in 2020 40 Zeta (1 in the. this paper about the expected Zeta and become one billion times) bytes of one trillion, concern various prolonged processing time, such as an overflow of processing has also been pointed out for the amount of data is enormous, by sampling the big data I propose a processing method to issue with a practical accuracy in the early stages of small data result, the all-powerful processing than leave the "preliminary election".

Keywords: Big data, Small data, Analysis

1. はじめに

近年、ビッグデータが注目されている。ビッグデータとは、一般的に「既存の一般的な技術では管理するのが困難な大量のデータ群」と定義されることが多い。ビッグデータを管理するのが困難なのは、3つのキーワード、Volume(量)、Variety(多様性)、Velocity(発生速度や更新頻度)で表わされるビッグデータの特性にある。

ITで処理されるデータの量は飛躍的に増えており、その種類もブログや動画サイト、またはfacebookやtwitterといったSNS(ソーシャル・ネットワーク・サービス)の利用者の増加により、パソコンやスマートフォンなどのコンピュータから、文字だけでなく、音声や写真、動画などのデジタルデータが、インターネット上の様々なスーパーコンピュータに蓄積されている。更に、それらのコンピュータには、利用状況や通信記録などのログと呼ばれるデータが日々生成されている。これらインターネットにつながる様々なコンピュータから生み出されるビッグデータは、合計すると数百テラ (1テラは約1兆) バイト以上とされており、米国のIT専門調査会社IDCによれば、2020年には40ゼタ (1ゼタは1兆の10億倍) バイトになると予測されるほど、データ量が膨大である。

ビッグデータの始まりと考えられる出来事は、2004年にgoogleが大規模検索サービスを実現するための基本技術であるGFS(Google File System)やMap Reduce[1]を論文形式で公表し、大きな注目を浴びた。その後も、googleは大規模データ処理技術の開発を行い、(1)Sawzall(大規模データに対する特定の処理を効率よく記述できるプログラム言語)、(2)BigTable(列のデータ処理を効率化した大規模対応データベース)、(3)Dremel(大規模データを分散システムで解析する技術)(4)Percolator(大規模データベースに短い間隔でアクセスできる方式)(5)Spanner(新しいパラダイムで実装された大規模データ分散システム)などの最先端技術を次々に公表した。

これらの技術には、ビッグデータ活用において、処理・分析に時間がかかる、サーバーの台数も数百から数千台必要となるためコストも高くなる、大量のデータが発生するため処理のオーバーフローが発生するという懸念がある。

そこで、本論はランダムサンプリングを応用し、ビッグデータを全件処理するプロセスと同時進行的にビッグデータをスモールデータ化することでニュースの選挙速報のように正式の結果が出る前に実用性のある精度でいち早く利用者が分析結果を得ることができる処理方式を提案する。

この処理方式を使うことで、これまで1週間から数日かかっていた処理・分析を1時間で行い、精度の保証、さらに処理に必要なサーバーの台数の問題、処理のオーバーフローの問題、コストの問題の解決に期待でき、ビッグデータの普及に効果が期待できる。

2. 従来研究

ビッグデータのスモールデータ化には一部の代表的データを抽出するランダムサンプリング手法が必要不可欠である。さらに、分析の対象や目的によってサンプリング手法を適切に選択、開発することが重要である。

例えば、グラフについてはランダムウォークを基本とするサンプリング手法であるHenzingerらが提案したPageRank[2]に基づくランダムウォークやLeskovecらのForestFire[3]法等があげられる。しかし、これらの手法には、ページ対象の調査と、リンク構造を対象としての調査の違いを考慮に入れていない。つまり、偏った結果しか得られないという問題点がある。さらに、PageRankの用いたサンプリングはコストがかかる問題点がある。

ビッグデータの場合、データの総数が不明なことが多くそのような場合はReservoir sampling[4]が有効であることが知られている。Reservoir Samplingは、母集団のデータが一度しかアクセスできない場合や、新しいデータが無限に出てくる場合に、データ出現の順序に関係なく問う確

率でサンプルを抽出する手法である。サンプルサイズを k として事前に決めておく必要がある。この手法は、非常に大きなデータセットを分析の際、多くの用途において有用である。しかし、この手法には層化抽出法が使われている。層化抽出法は、コストがかかる、適切な層に分けるのが困難、抽出するデータに偏りが生じる可能性が高いという問題点がある。

3. ビッグデータのスマールデータ化

ビッグデータの価値は高く、膨大なデータ量(数百テラバイト、時には数百ペタバイト)であるビッグデータを処理・分析することで、近年様々な分野で価値を見出している。しかし、必ずしもビッグデータ内のデータ全てを処理・分析する必要はない。例えば、選挙では当日の夜にサンプリングにより「当選確定」が出て、翌日に全件分析による正確な得票数が出る。このように、大きなデータセットではなくても、情報の中から価値を見出すことは可能であり、実際に多くの組織が「通常サイズ」(数テラバイト)のデータセットから多くの価値を見出している。つまり、ビッグデータの一部のデータを処理・分析するだけでも多くの価値を見出すことができ、様々な分野でも活用することができるのである。これをビッグデータのスマールデータ化と呼ぶ。

スマールデータ活用の例としてマーケティングについて説明する。マーケティングとは、顧客がモノやサービスなどの商品を購入する理由を、探るために顧客自身やその市場をより深く理解するための活動である。そのためには、顧客や市場に関する具体的な仮説を持つことが不可欠である。これは、アプローチが定量的であろうと定性的であろうと変わらない。ここで仮説と関係のないデータが入っていると、ノイズとなるばかりか、仮説検証を歪める原因にさえなりかねないのである。このようにスマールデータであっても価値を見出すことは可能である。

ビッグデータをスマールデータ化することで、ビッグデータの課題である膨大なデータ量を高速で処理、使用するサーバーの台数も数台、処理のオーバーフローの心配がなくなる、またビッグデータでなくても多くの洞察を得られる効果が期待できる。

4. 適用範囲

ビッグデータの活用パターンの分類を以下に示す。(fig.1 参照)

4.1 個別最適・バッチ型

特定の個人やモノに関するデータを収集して、その人に最適な商品やサービスを推奨あるいは、そのモノに最適な処置を施すケース。

(1) 大阪ガス (電気・ガス業)

コールセンターに寄せられる給湯器などの修理依頼の内容から、交換が必要となりそうな部品を自動で割り出す。過去数百万件の修理履歴や機器の型番などをもとに適切な部品を自動抽出するシステムを自社開発した。修理作業員の業務効率化に寄与する。

(2) カルチュア・コンビニエンス・クラブ (サービス業)

TSUTAYA 各店舗の POS データや、レンタル商品ごとの利

用履歴、3000 万人を超える会員の購買履歴を分析。キャンペーンやサービス向上、ゲームソフトを仕入れる際のメーカーとの交渉用資料作成などに役立てている。IBM の「Netezza」を導入し、分析処理のパフォーマンス低下を解消する。

(3) 国土交通省関東地方整備局 (公社・官庁)

2012 年に開通した「東京ゲートブリッジ」の異常を検知するのにセンサーデータを活用する。橋脚に多数のセンサーを取り付け、橋のひずみや振動などを検知する。橋の破損状況を遠隔より把握できる。橋を通行する車両の重量から、橋への負担なども予測する。

(4) 東京海上日動火災保険 (損害保険)

保険の勧誘に位置情報を活用する。NTT ドコモと共同で「ドコモワライム保険」と呼ぶサービスを提供する。スマートフォンの位置情報を使い、空港やゴルフ場に到着したユーザーに対して、ケガなどに備える保険を案内するメールを配信する。

4.2 全体最適・バッチ型

多数の個人やモノが発する情報を収集・蓄積し、蓄積したデータを一括して統計的に処理・分析することで、その個人やモノが属するコミュニティや社会全体にとって役立つ統計情報をフィードバックあるいは、最適化を図るケース。

(1) ウェザーニュース (情報・通信業)

会員から寄せられる現地の気象情報を分析。詳細なエリアごとに的確な天気予報を配信できるようにする。ゲリラ豪雨などの局地的な天気も高い精度で予測できる。多い時には 1 カ月あたり 100 万件の情報が会員から届く。観測機の気象データも活用する。

(2) ソフトバンクテレコム (情報・通信業)

携帯電話やスマートフォンの接続率と利用者の満足度を調査することで、費用対効果の高いエリア対策を実施できるようにした。アプリを介して収集する、月間 3 億件の通信情報を分析する。地図と組み合わせたエリアデータは月間 8000 万件に上るといふ。Twitter の投稿から自社の評価も分析する。

(3) 全日本食品 (小売業)

約 1800 店舗の POS データを収集/分析する。発行したクーポンや配布したチラシの効果を最大化する施策の検討に用いる。顧客情報や購買履歴をもとに、最適なクーポンをレシートの裏に印刷して発行する。

4.3 個別最適・リアルタイム型

個別最適・バッチ型同様に特定の個人やモノに関するデータを収集して、その人に最適な商品やサービスを推奨あるいは、そのモノに最適な処置を施すケース。

(1) ANA (空運業)

航空機に取り付けたセンサーを使って機体の重心を測定する。搭乗者数や荷物の重量データも収集し、乗客や積荷などをどう配置すれば重心を最適化できるか割り出す。重心位置を適正に保つことで、燃料費を削減できる。

(2) NTT ドコモ (情報・通信業)

販売店やコールセンターに寄せられた顧客からの問い合わせなどを収集し、マイニング技術を使って分析する。分析結果を新商品の開発やサービス改善などに役立てる。

携帯電話やスマートフォンから収集したデータ「モバイル空間統計」の販売にも乗り出す。

4.4 全体最適・リアルタイム型

多数の個人やモノが発する情報を収集・蓄積し、蓄積したデータを一括して統計的に処理・分析することで、その個人やモノが属するコミュニティや社会全体にとって役立つコンテキストに合わせてリアルタイムにフィードバックあるいは、最適化を図るケース。

(1)QUICK (金融業)

大量かつ高頻度で生成する証券データを高速処理し、顧客にほぼリアルタイムに金融情報を配信できるようにする。東証や大証、海外の取引所などから取得したデータを、為替、銘柄、期間、企業別に顧客に応じた形式で配信する。

(2)TipRanks (金融・情報業)

クラウドベースでアドバイスが参考になるかどうかを判定する仕組みを提供している。同社のアプリケーションを通して例えば米経済誌フォーブスの記事を読むと、アナリストの名前がハイライトされ、そこをクリックすると、そのアナリストの予想がどのくらいあてになるかがパーセンテージなどで表示される。また、その人のこれまでの予想が実際にあたったかどうか、プラスマイナス何%かで表示される。加えて、過去のデータを基に、予想が当たったかどうかによってアナリストをランキングしているため、ある特定の株式銘柄について調査したい場合、これまで予想が的中しているトップランクのアナリストがその銘柄についてどのような意見を持っているかも表示される。[5]

ビッグデータの分類において、スモールデータ化に最も適用されやすいのは、個別最適・バッチ型であると考えられる。

なぜなら、この個別最適・バッチ型のように蓄積されたデータで、個人・モノに最適な処置を施すのが一番適しているからである。その最たる事例がネット選挙である。これは、有権者の投票行動を文字通り一人一人割り出すシステムが開発された。ツイッターやフェイスブックなどのデータはもちろん、商品の購入履歴などの情報を消費者情報の会社から収集することで政治的な傾向を割り出す。このような事例に、スモールデータ化を適用すれば、さらにシステムが機能すると考えられる。

ビッグデータのスモールデータ化に適しているのは、個別最適・バッチ型であると考えられる。

5. ランダム性の保証と精度

ここでは、ビッグデータをスモールデータ化する際、ランダムにサンプリングされるデータのランダム性の保証と精度について例を用いて説明する。

5.1 視聴率の計算

ビデオリサーチが視聴率を上げるまでの流れの中の、視聴率調査のサンプル世帯の選び方を紹介する。[6]

関東地区の標本抽出の場合、関東地区の全 17,022,000 世帯 (2013 年 10 月 1 日の国勢調査の速報値を基に、ビデオリサーチが独自推定したテレビ所有世帯) が入った大きな鍋をよくかき混ぜ次のようにサンプル世帯を無作為に選び出す。

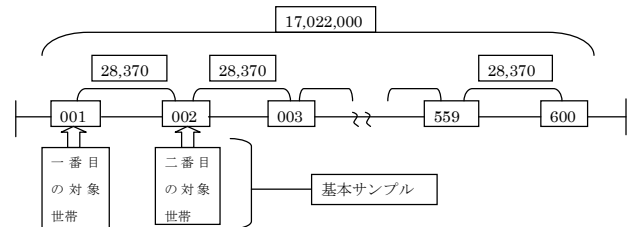


Fig. 2 サンプルの選び方

- ① 関東地区の調査地域に含まれるすべての国勢調査の調査区を行政単位別、調査区の順に並べる
- ② 調査地域内の世帯数を合計し、総世帯数を求める
- ③ 総世帯数を調査サンプル数(600)で割り抽出間隔を求める
- ④ 乱数表を用いて一番目のサンプルとなる世帯を決める
- ⑤ 一番目のサンプルとなった世帯に、③で求められた抽出間隔を順に加えて、調査地点の抽出番号を求める
- ⑥ 国勢調査の調査区一覧の世帯数を加算して、⑤で求めた抽出番号を含む調査区を選び出す
- ⑦ ⑥の調査区の地点地図を作成する
- ⑧ ⑦の地点地図に一定のルールで番号をつけ、抽出番号に該当した世帯をサンプルとする

視聴率の精度は、600 世帯を調査して 20% という視聴率が出たとしたら、実際の視聴率は 95% の確率で 16.7%~23.2% の範囲に収まる。

6. sd 方式の概要

ここでは、スモールデータ化の一例として、系統抽出法を応用した sd 方式を紹介する。

これは母集団を構成するものに通し番号を付けておき、初めのデータだけランダムに選びあとは一定間隔で系統的に抽出する方式です。sd 方式のデータ構造の構築を 5.1 に、サンプリングする際の精度を 5.2 で述べる。

6.1 sd 方式のデータ構造

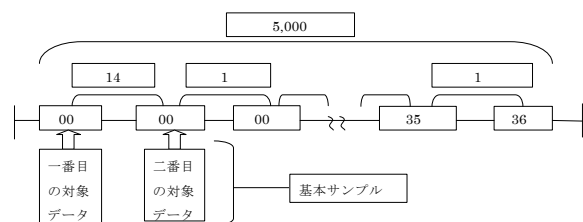


Fig. 3 sd 方式の構造

- ①ビッグデータの中からデータをランダムに取り出す方法を説明するために、ここでは仮にビッグデータを5,000とする。
- ②すべてのビッグデータを対象データ数で割り、インターバルデータを決める。
全ビッグデータ：5,000
抽出データ：360
 $5,000 \div 360 = 14(13.88\cdots)$ (インターバルデータが求められる)
- ③14より小さな数字をランダムに一つ選びスタートデータとする。
このデータを、速報とする。

・エクセルによってランダムに数字を選ぶ (例 300)

コンピュータの表計算ソフトなどで、一つのセルに「=INT(RAND()*300+1)」と入力し、「Enter」キーを押すとそのセルに1以上300以下の数字が表示される。数が表示されたセルの右下のかどをクリックし、適当なセルまでドラッグしてはなすと、その間のすべてのセルに1以上300以下の数が表示される。

※各セルに表示された数のうち、同じ数は捨て、上から必要な個数を標本とする。

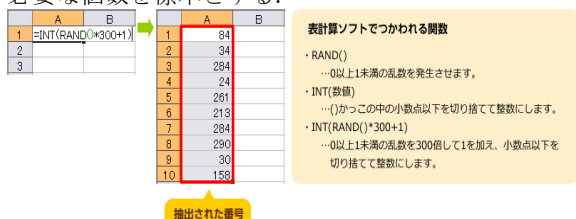


Fig. 4 エクセル上のランダムに数字を選ぶ

- ④スタートデータにインターバルデータを加算していき、選ばれるデータを求める。

6.1.1 対象データの根拠

今回は、紹介したランダムにデータを抽出する方法は系統抽出法というものである。例として5000のビッグデータの中から360のデータを抽出した。ここでは、Carl Drottのよく知られた論文をもとに、5000の中からなぜ360のデータを抽出したのか、抽出したデータの精度について図や表で説明する。

・360個系統抽出法の統計学的根拠[7]
360個のデータを系統抽出法で抽出するには以下のような統計学上の計算式に基づく。

$$\text{抽出サンプル数} \geq \frac{\text{調査対象蔵書の総点数 (母集団)}}{\left[\frac{\text{誤差の最大値}}{\text{信頼度係数}} \right]^2 \times \frac{\text{調査対象蔵書の総点数}-1}{\text{母比率} \times (1-\text{母比率})} + 1}$$

Fig. 5 抽出サンプル数の計算式

これを、エクセルなどの計算表ソフト上で書くと以下の表のようになる。

抽出対象データの総個数(母集団)	
母比率	50% ←これで固定
誤差の最大値	5 ←これで固定
信頼度係数	1.96 ←これで固定
抽出サンプル数	

Fig. 6 エクセル上の抽出サンプル数の計算式

ここで、抽出対象データの総個数とは文字通り、抽出対象になるデータの全データ (ビッグデータ) のこと。

母集団全部を対象にした悉皆調査を行った場合でも、誤りは発生する。ましてサンプルを抽出しての調査ならば、その分析結果を全体に反映させた時に、実態とは開き (誤差) が出てくるのは理解できる。もちろん、その誤差はなるべく小さいほうがよい。上記の表では5%にしている。これが誤差の最大値である。95%の確率で当たっている (信頼度) が、±5%の幅で間違っている (誤差) 可能性があるという意味である。最大値を±1にしてもかまわないが、サンプル数が膨らむし、蔵書のサンプル調査ではほとんど意味がない。95%±5%で充分である。

次に、母比率というのは、ある調査項目に該当するものが母集団に占める比率の予測である。それを知りたいがためのサンプル調査であり悉皆調査なのに、なぜ調査前に予測して入れておかなければならないのか、戸惑うのは当然である。しかし統計的にはこれは必須項目になる。普通は一項目だけの抽出はありえず、複数の項目が設定される。複数の項目にまたがる母比率を予想して入れなければならない時には、それぞれの項目では比率に凸凹があったとしても、全体としてならば半分ぐらいを占めるであろうとして50%に設定するのが定石になっている。

信頼度係数とは特定の信頼度に対して決まっている以下のような定数である (正確には特定の信頼度に対応する標準正規分布の%点という)。

信頼度 90%	係数は 1.95
信頼度 95%	係数は 1.96
信頼度 99%	係数は 2.58

状態調査では信頼度 95%で充分なので、係数は 1.96 になる。

上記の計算式をセル構成ならば

$$=B1/((B3/B4)^2*((B1-1)/(B2*(1-B2))))+1)$$

となる。
表の「抽出対象データの総個数」のセルに5000を入れる。

抽出対象データの総個数(母集団)	
母比率	50% ←これで固定
誤差の最大値	5 ←これで固定
信頼度係数	1.96 ←これで固定
抽出サンプル数	356.82

Fig. 7 5000を入れた場合の抽出サンプル数の計算式

すると「抽出サンプル数」のセルが、356.82になる。総数からこの数(360)だけ抽出して調査すれば、信頼度が95%、誤差が±5%で全体に敷衍出来る統計になる。実際にこの数

よりも多く、切りのよい360個を選べばよい。後々の分析の際にも都合がよく、統計的にも問題ない。もし抽出数が360に満たなくとも、表の数値以上ならば問題ない。360を超えても、もちろん問題ない。

6.2 sd方式の精度

sd方式のランダムにサンプリングされたデータが、利用者の望む分析結果なのかどうかを統計学の区間推定を使って、どの程度の精度で収まるのかを説明する。

6.2.1 区間推定[8]

区間推定とは、母集団の平均値(真の値)を、ある幅を持って推定することです。

このとき、標本を使うので誤差が生じる。その誤差が存在するため、ある範囲の中に母平均があると推測する際、その範囲を信頼区間と呼ぶ。

信頼度95%の時の信頼区間:100回の標本調査のうち95回はこの範囲に母平均がある。

100回推定したら95回は「許容できる誤差の範囲」(許容誤差)の範囲内に収まるという意味である。「許容できる誤差の範囲」5%(信頼度95%)で「統計学上十分意味あり」と言われている。信頼度は信頼係数ともいう。

区間推定の場合、精度を表すことができる。「平均値はだいたいここからここまでの間」とする信頼区間の幅は、もちろん広すぎればあまり意味がない。狭い方が、推定の精度が良いことを示している。

6.2.2 区間推定の事例

ある地域で、あるテレビ番組の視聴率を調べるため100世帯を無作為抽出した所、この番組を見ていたのは20世帯だった。信頼度95%でこの地域のこの番組の視聴率を区間推定する。

平均は $20/100=0.2$

分散は $0.2*0.8=0.16$

母標準偏差は $\sqrt{0.16}=0.04$

となる。

よって視聴率の区間推定は

$0.2-1.96*0.04$ から $0.2+1.96*0.04$ まで

$12.2\%-27.8\%$ となる。

6.2.3 sd方式の精度

6.1で説明した例を区間推定する。

全ビッグデータ:5000

抽出データ:360

母標準偏差

$$=((\text{データ}-\text{平均値})\text{の二乗})\text{の総和}/(\text{個数}-1)$$

標本平均 $\pm 1.96\sqrt{\text{母標準偏差}/360}$

$=24.70372 \leq \text{母平均} \leq 27.79628$

となる。

つまり、 24.7% から 27.7% までとなる。

6.2.2の事例では $12.2\% \sim 27.8\%$ で信頼区間が広いので、データの精度はよくない。しかし、6.2.3のsd方式の例では $24.7\% \sim 27.7\%$ で信頼区間が狭いため、データの精度は良いといえる。つまり、sd方式のランダムサンプリングは95%の信頼度で $24.7\% \sim 27.7\%$ の範囲に収まる。

7. 結論

これまで、ビッグデータのスモールデータ化についての活用の可能性とスモールデータ化の際のランダム性の保証と精度、スモールデータ化の一例として系統抽出法を応用したsd方式の提案をした。系統抽出法を用いることで従来研究の問題点である抽出するデータに偏りが生じる問題を解決し、ランダムにデータを抽出することができる。また、コストの問題の解決にも期待できる。

ビッグデータをスモールデータ化することで、ビッグデータにおける処理時間の長期化、処理のオーバーフロー、サーバーの台数の問題、コストの問題を解決でき、スモールデータであっても多くの洞察を得られる効果が期待できる。

8. 今後

スモールデータ化するビッグデータのさらなる選定、既存の他のサンプリング手法の適用を行う予定である。

参考文献

- [1] Jeffrey Dean, Sanjay Ghemawat
"MapReduce:Simplified data Processing on Large Clusters", 2004
- [2] M.Henzinger, In Proceedings of the 9th International World Wide Web Conference, 2001, pp.295-308
- [3] Jure Leskovec, Christos Faloutsos, "Sampling from Large Graphs", 2006
- [4] Jeffrey Scott Vitter, "Random Sampling with a Reservoir", 1985
- [5] 日経ものづくり 2013年7月号 「攻めのビッグデータ活用」
- [6] 藤平芳紀, "視聴率の正しい使い方", 朝日新聞社 2007
- [7] 松井博, "標本調査法入門—基礎から学ぶ, 標本調査の理論と実際", 2005/9/30
- [8] 小島寛之, "完全独習 統計学入門", 2006/9/28

Fig.1 ビッグデータ活用パターンの分類

<p>個別最適・バッチ型</p> <p>特定の個人や物のデータを広範囲に収集, 分析 ここに最適な商品やサービスの推奨/最適な処置</p> <ul style="list-style-type: none"> ●ワン・トゥー・ワン・マーケティング ●顧客離反分析 ●危機の故障予測 など 	<p>個別最適・リアルタイム型</p> <p>特定の個人や物のデータを広範囲に収集, 分析 個々に最適な商品やサービスの推奨/最適な処置をリアルタイムに実施</p> <ul style="list-style-type: none"> ●行動ターゲティング広告 ●行動に応じたクーポン配信 ●リアルタイムの商品レコメンデーション など
<p>全体最適・バッチ型</p> <p>多数の個人や物が発する情報を収集・分析 コミュニティ全体に役立つ情報をフィードバック/最適な処置を実施</p> <ul style="list-style-type: none"> ●検索エンジンや翻訳エンジンの精度改善 ●Twitter のつぶやきをベースに株価予測 ●ウェブサイトのユーザビリティ改善 など 	<p>全体最適・リアルタイム型</p> <p>多数の個人や物が発する情報を収集・分析 コミュニティ全体に役立つ情報をリアルタイムにフィードバック/最適な処置を実施</p> <ul style="list-style-type: none"> ●ブローブ情報の分析による渋滞予測 ●スマートメーターの利用による電力需要予測 など