■ 論　文

企業におけるビッグデータの利活用とデータベース技術を用いた分析

葛 西 正 裕

目　次

概　要

情報通信技術の発達により，数値データ，テキスト，音声，画像，動画といった多様かつ膨大なデータが収集可能になっている。このようなデータはビッグデータと呼ばれ，競争力を高めようとする企業において，ビッグデータの利活用が進んでいる。そこで，本稿は，企業におけるビッグデータの利活用の現状を明らかにする．その上で，ビッグデータの蓄積・運用や分析に不可欠なデータベース技術に注目し，要素技術を踏まえながら，筆者の先行研究である階層的分類を用いた分析手法がデータの多様化に起因する諸課題に対して応用可能であることを示す。考察の結果，ビッグデータに関する記事検索数から，2011年中頃から急速にビッグデータに対する関心が高まっていることが明らかになった。同時に，ビッグデータの利活用の現状は，ビッグデータを対象にしたビジネスインテリジェンス，すなわち，データを分析し社内の意思決定支援に利活用するという狭義のビッグデータの利活用と，ビッグデータを自社の製品やサービスの高付加価値化に利活用するという広義のビッグデータの利活用に整理できる。先行研究のデータに対する階層的分類は，多様であるビッグデータの種類を問わず一元的に管理できるデータの構成法である。さらに，ビッグデータの多様性は分類時にデータの不均一性となって現れるため，先行研究の不均一なデータに対するデータの構成法や分析に供するデータの指定法は，ビッグデータの蓄積・運用に係るデータウェアハウスのデータモデルやビッグデータの分析工程の一部において応用可能である。

キーワード

階層的分類，多様なデータ，データウェアハウス，データベース，データマイニング，データモデル，ビジネスインテリジェンス，ビッグデータ，不均一なデータ

採択日：2012年10月17日

1 はじめに

今日における情報通信技術の発達は目覚ましく，インターネット関連技術を中心としたネットワーク環境の整備とタブレット型パーソナルコンピュータやスマートフォンに代表される新たなデバイスの普及が進むことにより，膨大なデータが日々生まれている。また，センサ技術の発達やSNS[1]などのWebサービスの普及について，数値データのみならず，テキスト，音声，画像，動画といった多様なデータを収集できるようになっている。このような膨大かつ多様なデータは，ビッグデータと呼ばれ，ビッグデータの利活用が注目されている[11]，[15]。

中でも企業におけるビッグデータの利活用に対して関心が高まっている[26]。その背景には，情報通信関連機器の高性能化や低価格化により，ビッグデータを比較的容易に収集し利活用できるようになってきたという背景がある。技術的背景に加えて，複雑さを増し変化の激しい経済環境下において企業は，情報通信技術による費用削減といった経営のスリム化の段階を超えて，情報通信技術を用いた経営の高度化を図ることで競争力を維持したいという背景がある。とりわけ企業運営に携わる経営層や管理者層など意思決定者は，恒常的に高度かつ迅速な意思決定を行う必要性に迫られている。意思決定においては，過去の企業活動や現在の経済状況など企業内外で得られたデータから知見を導き現在や将来の意思決定に結びつけることが求められ，そのためには，ビッグデータを分析することが重要になってくる[40]，[45]。文献［32］では，経済活動で得られたデータを分析し様々な意思決定に反映させることが企業の生産性の向上，ひいては企業業績の向上に寄与することが実証されており，ビッグデータを分析し意思決定を行う重要性は明らかである。

企業運営に関連する様々なデータを蓄積し分析する一連の工程に関するツールや手法は，ビジネスインテリジェンスと呼ばれ[43]，意思決定を支援する情報システムの中で用いられてきた[2]，[47]。今日における分析対象となるデータは，企業運営における財務や在庫等に関連する数値データのみならず，Webサイトのテキストやログデータ，コールセンターの音声データ，製品やサービスに関連する画像や動画，スマートフォンに搭載されたGPSセンサから取得した位置データなど質的に拡大しているのと同時に，その量も膨大になっている。すなわち，対象データの質的な多様化と量的な拡大に伴って得られる知見の新規性や高度化に対する期待が高まり，幅広い業種においてビッグデータを対象とした分析が企業における意思決定に応用されることが多くなってきている。さらに，一部企業においては，顧客が製品やサービスを使用している際に取得できるデータや顧客がSNS等のデジタルメディアを通じて発信するデータを収集し，それらのデータを分析・加工したものを再び顧客に対して製品やサービスの付加価値として還元するといった段階に来ており，ビッグデータの利活用はより高度化している[10]，[20]。

このように，ビッグデータを分析し意思決定に役立てたり，ビッグデータから製品やサービスの高付加価値化を図ったりする際，データを収集し分析・加工を行っていく必要があり，デー

タベース技術が基盤となる。対象となるデータがビッグデータになれば，データ量が増大するのと同時にデータの種類も多様化するため，データの構成・処理・分析に関する新たな要求やそれに伴う課題が生じており，解決手法が求められている。

よって，本稿では，ビッグデータに着目し，まず，ビッグデータの利活用の現状について整理する．その上で，経営の高度化を図るために必要な情報システムの中でも重要性を増しているデータベース技術に注目し，ビッグデータの分析を目的とした情報システムにおける要素技術を述べる。それらを踏まえて，筆者の先行研究である階層的分類を用いた分析手法がビッグデータの大容量化や多様化に起因する諸課題に対して応用可能であることを示す。すなわち，本稿は，ビッグデータに関する現状把握を踏まえた上で，データベース技術の視点からみたビッグデータに対する分析の技術体系の概説と先行研究の有用性について明らかにするものである。

本稿の構成は，以下の通りである。第２章は，ビッグデータの定義やビッグデータの利活用の種類について述べるのと同時に，それらに関する関心の程度を定量的に把握する。第３章は，ビッグデータの利活用に関する国内外の事例を紹介する。第４章は，ビッグデータの分析を目的とした情報システムの体系と主要技術をデータベース技術の視点から説明する。第５章は，階層的分類を用いた分析手法の概要とビッグデータに対する有用性を述べる。第６章は，本稿のまとめである。

２　ビッグデータの利活用の種類と関心の推移

本章は，ビッグデータの定義について述べた後，ビッグデータに関する関心の程度の推移をビジネスやコンピュータ関連の雑誌及び経済系新聞の記事検索件数を時系列でみることで定量的に把握する。その上で，ビッグデータの利活用は２種類に分類されることを示し，２種類の利活用に対する関心の推移について述べる。

２－１　ビッグデータの定義

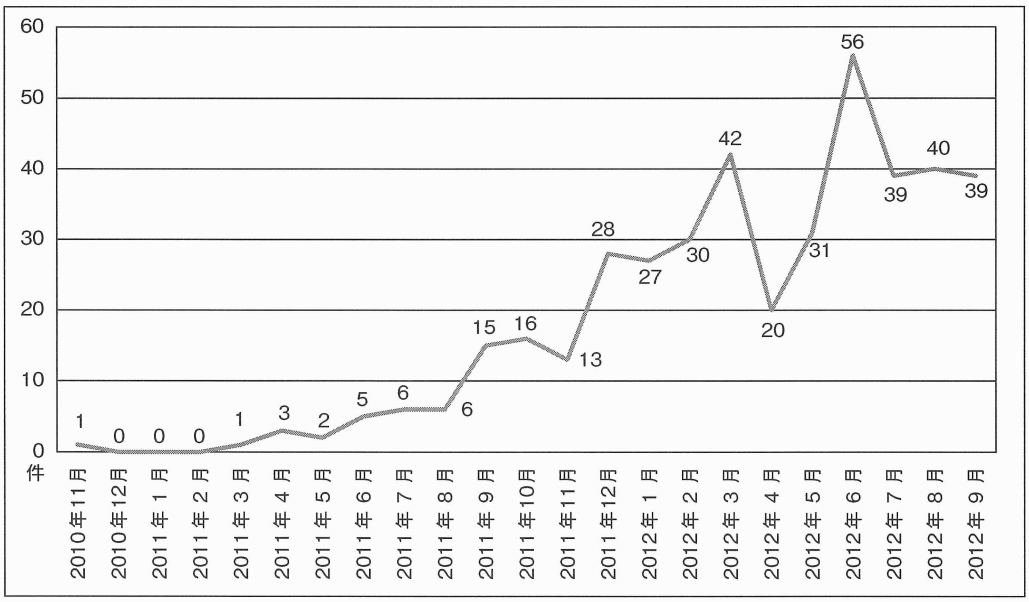
情報通信技術の発達により，数値データのみならず，テキスト，音声，画像，動画といった多様なデータを膨大に収集できるようになった。このような膨大かつ多様なデータを一般にビッグデータと呼ぶ[48],[52]。ビッグデータに関する明確な定義は存在しないが，文献［10］を参照すると，ビッグデータとは，既存の一般的な技術では管理するのが困難な大量のデータ群という量的側面に加えて，企業でいえば，販売や在庫に係る数値データ，Web サイトのテキストやログデータ，コールセンターの音声データ，製品やサービスに関連する画像や動画，スマートフォンに搭載された GPS センサから取得した位置データなど多様性を持つという質的側面，およびデータの発生や更新頻度が高いという速度的側面の３側面の特徴をもつデータと定義されている。一方，量的側面についてみても，文献［27］では，現段階では数百テラバイトから

数ペタバイト程度としているが，今後の情報通信技術の進展を考慮するとより大容量のデータが対象になると解釈しており，明確な定義は存在しない。質的側面においても同様である。また，速度的側面については，本稿の議論とは関連性がないために取り上げない。ゆえに，本稿では，一般的に膨大かつ多様とされるデータをビッグデータとして定義し，ビッグデータにおける量的側面と質的側面の２側面に着目し議論を展開する。

２－２　ビッグデータに対する関心の推移

ビッグデータに対する関心の高まりについて述べる。図表１は，ビジネスやコンピュータ関連の雑誌及び経済系新聞において，“ビッグデータ”という用語を含む記事の件数を月次で示したものである。“ビッグデータ”という用語が初めて使用されたのは，2010年11月17日の日経産業新聞においてである。同記事は米国においてストレージ事業[2)]を展開するEMC社に関するものであり，同社CEOがビッグデータ時代の到来について言及している。2010年11月以降の記事件数の推移をみると，2011年中頃より急速に記事件数が伸びており，今日に至る約１年の間に関心が高まっている。

図表１　ビッグデータに対する関心の推移（雑誌・新聞における記事検索数）



資料）記事検索の対象とした雑誌および新聞は，日経ビジネス，日経コンピュータ，日経情報ストラテジー，日経ソリューションビジネス，日本経済新聞朝刊，日本経済新聞夕刊，日経産業新聞，日経MJ（流通新聞），日経金融新聞，日経地方経済面，日経プラスワン，日経マガジンである。同雑誌および新聞を対象にして，日経BP記事検索サービスおよび日経テレコン21におけるデータベース機能を活用し検索を行った。
注）図書紹介や広告に関する記事は除く。また，重複した記事は単数として扱う。

2－3　狭義と広義のビッグデータの利活用

ビッグデータに関する関心の高まりの背景には，膨大なデータを安価かつ容易に蓄積できるようになった技術的背景と同時に，世界的な経済環境の厳しさを受けて競争力を高めようとする企業の動きがある。これまでの情報通信技術に関する投資目的は，業務プロセスの自動化やインターネットを通じた電子商取引といった効率化や利便性に寄与するものであり，費用削減を主たる目的とするものであった。情報通信技術による経営のスリム化が進んだ今日では，情報通信技術の役割は，企業の意思決定支援，さらには製品やサービスの高付付加価値化や自社のブランド価値向上といった経営の高度化の段階にきている[12],[13]。文献［29］において，日本企業は社内業務や製品の受発注などを効率的に行うための情報システムについては効果をあげているのに対して，米国企業は経営戦略立案や製品やサービス開発など企業における意思決定に係る分野において情報通信技術を効果的に利活用していると指摘しており，日本企業においても経営の高度化を目的とした情報通信技術の利活用の必要性が指摘されている。

ビッグデータを用いて企業が経営の高度化を図る場合には，まず，ビッグデータを蓄積し分析を行って自社内の様々な意思決定支援に利活用する段階がある。これを狭義のビッグデータの利活用と考える。例えば，ある企業が自社のPOSデータを分析し，地域別に特定期間における売れ筋商品を抽出したり，同時に購入している商品群に関するルールを発見したりするなどして，発見した有益な情報を知識として販売促進や在庫管理等に反映させていく場合が該当する。狭義のビッグデータの利活用においては，ビッグデータの分析結果の利活用範囲は主として企業内であって，その利活用の主たる目的は製品やサービスの開発・生産・供給や企業戦略等に係る意思決定支援である。

狭義のビッグデータの利活用では，収集したデータの分析結果を社内における種々の意思決定支援に応用するというものであったが，今日の一部企業においては，顧客が製品やサービスを使用している際のデータやそれ以外の場面において顧客がSNS等を通じて発信するデータを一元的に収集し分析・加工したものを顧客に製品やサービスの付加価値として還元する取り組みがなされている。このようなビッグデータの利活用は，ビッグデータを蓄積し分析・加工することでコンテンツ[3)]として企業の製品やサービスの高付加価値化や自社ブランド価値の向上を図るものである。例えば，自動車業界であれば，自動車と人と社会インフラをネットワークで結ぶITS[4)]において，プローブカーと呼ばれるデータ収集・送信の機能を持つ自動車から得られるブレーキやアクセルといった運転情報や車両の位置情報などを一元的に集約し，膨大なデータを分析・加工した後，渋滞情報や交通安全情報としてカーナビゲーションに表示するなどして利用者に還元するシステムが提供されている[3]。同システムによって，精緻な渋滞情報が取得できたり，急ブレーキの頻発地帯を注意喚起したりと，快適で安全な運転を実現できるという効用を消費者は得ることができる。すなわち，消費者から収集したビッグデータを分析・加工し，渋滞や交通安全情報というコンテンツを生み出すことで，自動車という製品の高

付加価値化に寄与している。さらに，同事例では，自動車という製品に付随するサービスとしての価値を生み出し，新たなサービスとして企業に利益をもたらすシステムにまで到達している。ゆえに，広義のビッグデータの利活用は，狭義のビッグデータの利活用における次の段階として位置づけられる。類似例として，エスエス製薬は，Twitter[5]のツイートと呼ばれるコメントを分析し，風邪の種類や流行場所を表示するサイトを構築した[20]。同サイトは，風邪予防を目的としており，製薬企業という立場からは利益に反する事業になるが，企業ブランド価値の向上を図る目的があり長期的な視点から企業利益の向上につながることを期待している。成熟した日本や欧米等の市場では，製品自体の物的価値に加えてそれに付随するコンテンツといった新たな付加価値が求められるようになっており[9]，日本が競争力を有する製造業を中心とした既存産業がコンテンツを利活用して高付加価値化を図る必要性が迫られている現状において[6]，２つの事例は好例であると考えられる。このようなビッグデータの利活用は，狭義のビッグデータの利活用と比較して明確かつ厳密な分類は困難ではあるが，広義のビッグデータの利活用と考えられる。これまで述べてきた狭義のビッグデータの利活用と広義のビッグデータの利活用の相違点についてまとめたものが図表２である。

図表２　ビッグデータの利活用の種類と特徴

ビッグデータの利活用の種類	主な利活用の目的	主な利活用の範囲	利活用の事例	コンテンツとしての価値の有無
狭義のビッグデータの利活用	製品やサービスの開発・生産・供給や企業戦略等に係る意思決定支援	企業内	・小売業においてPOSデータの分析により関連売れ行き商品といったルールを発見し販売や在庫管理に利用	無し
広義のビッグデータの利活用	・製品やサービスの高付加価値化 ・自社ブランド価値の向上	企業内と消費者	・ITSによる渋滞情報や交通安全情報を作成しカーナビに提示 ・ツイートを収集し加工することで風邪流行を予測するサイトの展開	有り

資料）各種資料を参考に筆者作成。

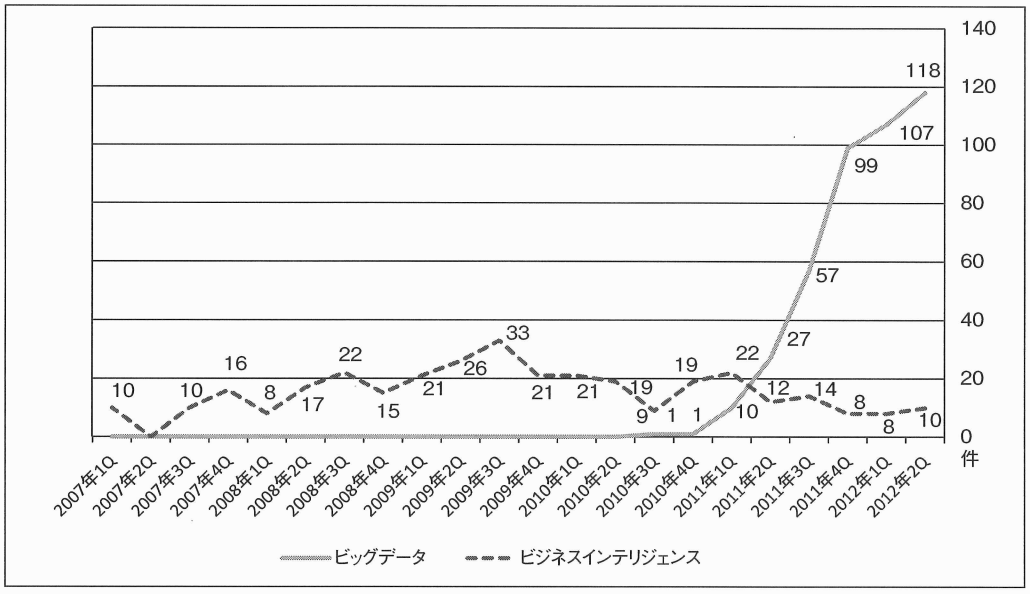
２－４　ビッグデータとビジネスインテリジェンスの関係と関心の推移

狭義のビッグデータの利活用については，データ量の多寡を問わず，従来からビジネスインテリジェンスとして，データを収集して蓄積し分析を行うことで意思決定支援に利用するという概念が存在する。文献［43］によると，ビジネスインテリジェンスは，企業におけるビジネスデータを意思決定に有用な情報に迅速かつ高度に変換するツールであり手法であると定義されている。すなわち，狭義のビッグデータの利活用とは，対象データをビッグデータとするビジネスインテリジェンスと考えることができる。

ビジネスインテリジェンスの関心の傾向について，ビッグデータと同様にビジネスやコン

ピュータ関連の雑誌及び経済系新聞において，“ビジネスインテリジェンス”もしくは“BI[6)]”という用語を含む記事の累積件数を四半期別に図表３の通り示した。図表３によると，ビッグデータに関しては，最近急速に注目を浴びる概念であるのに対して，ビジネスインテリジェンスは５年前から恒常的に関心がある。ビッグデータに関する記事件数が急速に伸びる中，ビジネスインテリジェンスに関する記事件数はそれに伴って増えていない状況を踏まえると，図表２で示したビッグデータの狭義の利活用，すなわち，ビッグデータを対象にしたビジネスインテリジェンス以上に，ビッグデータを用いてコンテンツ化をして製品やサービスの高付加価値化や自社ブランドの価値向につなげるようなより高度な利活用，すなわち，広義のビッグデータの利活用が増えている，あるいはビジネスインテリジェンスに関する記事がビッグデータに関する記事に包含されているとも考えられる。

図表３　ビッグデータとビジネスインテリジェンスに対する関心の推移



資料）記事検索の対象とした雑誌および新聞は，日経ビジネス，日経コンピュータ，日経情報ストラテジー，日経ソリューションビジネス，日本経済新聞朝刊，日本経済新聞夕刊，日経産業新聞，日経MJ（流通新聞），日経金融新聞，日経地方経済面，日経プラスワン，日経マガジンである。同雑誌および新聞を対象にして，日経BP記事検索サービスおよび日経テレコン21におけるデータベース機能を活用し検索を行った。
注１）図書紹介や広告に関する記事は除く。また，重複した記事は単数として扱う。ただし，“ビッグデータ”と“ビジネスインテリジェンス”（“BI”）の両方で検索される記事に関しては両者に含める。
注２）１Qは当該年度の４月から６月，２Qは当該年度の７月から９月，３Qは当該年度の10月から12月，４Qは当該年度の１月から３月であり，図表は累積件数を示している。

３　ビッグデータの利活用に関する国内外の企業の事例

本章は，ビッグデータの利活用に関して，狭義のビッグデータの利活用と広義のビッグデータの利活用に区分しながら，２種類のビッグデータの利活用に関して，代表的な事例を国内外

の企業別に紹介する。

3－1　国内企業の事例

狭義のビッグデータの利活用に該当する事例と広義のビッグデータの利活用に該当する事例を抽出し，国内企業ないしは日系企業に関してまとめたものが図表4である。ALSOK（綜合警備保障株式会社）は，警備関連サービスを展開する企業であり，コンビニエンスストアに配置しているATMへの現金補充業務を行っている。ATMには，出金と入金の両方の用途があり，現金が多すぎることによって入金ができない障害や反対に少なすぎることで出金ができない障害があり，いずれの場合においてもATMを停止させることになる。したがって，ATMの稼働率の向上には，現金補充量の調整が不可欠である。現金補充量の決定に関して，これまでは情報システムを介さず人的な経験や勘によって行っていたが，ATMの設置台数と利用者が増えたために，膨大な出入金データを人間の経験や勘のみに頼った分析では適切な意思決定ができなくなってきた。そこで，同社は，ATMの現金需要の把握を目的とする情報システムを構築し，現金補充量の意思決定に過去のデータ等を分析した結果を適応することでATMの稼働率の向上を図っている[24]。同様に，狭義のビッグデータの利活用の事例として，駅構内において飲料水の販売事業を展開するJR東日本ウォーターは，Suica[7]決済端末を自動販売機に導入し，消費者の行動パターンと消費者の属性を組合せて分析することで，消費者のニーズにあった製品開発を行ってヒット製品を生み出した[21]。さらに，空運業を展開する全日本空輸は，部品数が300万程度あるといわれる航空機の保守管理に対応した整備資材管理システムを2011年秋に構築し，部品等の在庫管理の効率化を図っている[17]。これまでの事例は，収集したビッグデータを在庫管理や製品開発の意思決定支援に用いるものであり，狭義のビッグデータの利活用の事例といえる。

一方，ビッグデータの利活用に関して，自社における意思決定支援という利活用にとどまらず，製品やサービスの高付加価値化や自社ブランドの価値向上を目的とした利活用を展開している企業の代表例をあげる。まず，気象情報を提供する事業を展開するウェザーニューズの事例である。同社は，約30万人の有料会員“ウェザーリポーター”が所持する携帯電話を活用して，会員から提供されるGPSによる位置情報とカメラによる天気画像を収集し分析・加工することで，従来の天気予報では予報が困難であるゲリラ雷雨の事前予測を高精度に実施し，顧客に注意喚起や情報提供を行っている[19]。また，建設用機器・車両の製造・販売を展開するコマツは，同社製品に情報通信機器を搭載し，世界中にある自社の機器・車両に関する使用状況・時間・場所を一元管理し分析するシステム“KOMTARX”を構築している。同システムにより，機器・車両の盗難防止や障害監視といった利活用以外にも燃料の顧客差を分析することで効率的な運転方法の助言といった多岐に及ぶサービスを行っている[10],[18]。具体的には，トルコにおいてSNSの代表的なサイトであるフェイスブックにKOMTRAXによって得られた知見とし

て顧客の運転の仕方や燃費動向を掲載し高評価を得ている[23]。さらに，化粧品販売を展開するファンケルでは，過去の販売履歴や他店舗・複数チャネルによる販売履歴を一元的に管理する顧客管理システムを構築し，顧客にあった情報を抽出しタブレット端末を利用して顧客に提示しており，今後はSNSを通じた情報提供も行う予定である[25]。これらの事例は，ビッグデータを一元的に収集した後，ビッグデータを分析・加工しコンテンツ化することで，製品やサー

図表4　国内企業におけるビッグデータの利活用事例

会社名	業種	事業内容	概要	利活用の分類	利活用の目的
ALSOK（綜合警備保障株式会社）	サービス業	セキュリティ事業および総合管理・防災事業	同社は，セブン銀行におけるATMの現金補充業務に携わっており，入出金があるATMの現金需要の把握を目的とするシステムを構築した。2009年にSASインスティチュートジャパンの需要予測ソフトウェアを導入し，ATMの現金量データを解析し予測精度を向上させている。	狭義のビッグデータの利活用	需要予測（補充量の調整）に係る意思決定支援
JR東日本ウォータービジネス	小売業	JRの駅構内を中心とした飲料品の販売	Suica決済端末を自動販売機に導入することで，消費者がどの時間にどの製品をどこで買ったかという販売データとその消費者の性別や年齢といった属性を組み合わせて分析を行うシステムを構築した。結果として，男性の30～40代は夕刻に甘い飲料を好んで買うことが判明し製品開発を行った結果，ヒット商品となった。	狭義のビッグデータの利活用	製品開発に係る意思決定支援
全日本空輸	空運業	国内外における航空機による旅客や貨物の運送	2011年秋を目途に，部品点数が極めて多い航空機に対応した整備資材管理システムを構築する。同システムは，航空機体の修理に使う整備用部品の在庫や購買履歴，部品ごとの交換時期などのデータを一元管理する。同システムでは，整備用品の需要予測機能を強化しているため，在庫の効率的な管理に有用である。	狭義のビッグデータの利活用	在庫管理や品質維持（保守メンテナンス）に係る意思決定支援
ウェザーニューズ	サービス業	天気に関する情報提供サービス	同社の約30万人の有料会員「ウェザーリポーター」が所持する携帯電話を活用して，GPSによる位置情報とカメラによる天気画像を収集し分析・加工することで，従来の天気予報では予報が困難であるゲリラ雷雨の事前予約を高精度に実施し顧客に注意喚起や情報提供を行っている。	広義のビッグデータの利活用	自社サービスの高付加価値化
コマツ	機械製造業	建設用機器・車両の製造・販売	KOMTRAXというシステムを構築している。同社の建設機器や車両に情報通信機器を搭載し，使用状況・時間・場所を一元管理し分析を行うことで，燃料の顧客差を分析などし効率的な運転方法の助言を行っている。	広義のビッグデータの利活用	自社製品及びサービスの高付加価値化
ファンケル	サービス業	化粧品販売	過去の販売履歴や他店舗・複数チャネルによる販売履歴を一元的に管理する顧客管理システムを構築し，顧客にあった情報を抽出しタブレット端末を利用して顧客に提示している。今後は，SNSを利用して販売促進につながる情報を提供していく。	広義のビッグデータの利活用	自社製品及びサービスの高付加価値化

資料）日経ビジネス，日経コンピュータ，日経情報ストラテジーの各種記事を参考に筆者作成。

ビスの高付加価値化や自社のブランド価値の向上を図っているという点で広義のビッグデータの利活用の代表例といえる。

３－２　国外企業の事例

海外企業ないしは外資系企業に関してまとめたものが図表５である。国外企業において，ビッグデータの利活用で代表的な事例は，日本においてもサービスを展開しているアマゾン・ドット・コムである[15]。同社は，インターネット上で書籍を中心とした電子商取引を展開するアメリカ発祥の企業であり，創業時より顧客第一の理念の下，データ分析によって顧客が必要としている情報を提供することで成長している企業である。同社は，電子商取引を行っているため，顧客の購買履歴は当然のこと商品に対する顧客の評価記事といったものまで膨大なデータを収集できる。同社は，そのようなビッグデータを分析対象にして，顧客が購入したい商品を検索すると過去の膨大なデータから関連商品や同時購入商品の情報を提供することで，顧客単価の向上や顧客満足度を高め継続的な購買につなげている。同時に，顧客に応じた広告配信も行っており，販売促進や広告配信に関する意思決定支援にビッグデータを利活用している狭義のビッグデータの利活用の代表的な事例といえる。同じくアメリカに拠点を置くVISAは，世界最大級のクレジットカード会社であり，世界中にいる顧客のクレジットカード利用に関連する様々なトランザクション処理を実行する傍ら，過去の履歴を含めた数百テラバイトに及ぶ履歴データというビッグデータと顧客属性などのデータを関連付けて，不正と思われる利用の検出を随時行っている[16]。本事例も，膨大な顧客の履歴データなどを分析し不正利用の検出を行うことで顧客管理に係る意思決定を支援する取り組みであることから，狭義のビッグデータの利活用の好例であるといえる。

一方，広義のビッグデータの利活用の代表的な事例としては，グーグルがあげられる[1]。同社は，世界中の情報を整理し世界中の人々がアクセスし使用できることを企業理念として，Webサイトの検索サイトの運営から始まり，検索サイトにおける検索順位に応じた広告収入を得るビジネスモデルを展開している。また，地図情報などのコンテンツ作成からメールサービス，OS[8]の提供といった多岐に及ぶ事業を展開している。グーグルが展開する事業において，地図情報に関する取り組みをみても，単なる地図情報にとどまらず自社の撮影車で撮影した画像を関連付けることで臨場感のあるサービスや自社で収集したWebサイトの検索と関連付けたサービスを提供しており，単なるポータルサイトというよりはコンテンツとしての価値を有している。こうしたコンテンツの提供が自社のサービスの高付加価値化に寄与していることはもちろんのこと，ひいてはグーグル社のブランド価値の向上につながっており，それが広告収入の増加や優秀なエンジニアの採用といった好循環を実現している。

図表５　国外企業におけるビッグデータの利活用事例

会社名	業種	事業内容	概要	利活用の分類	利活用の目的
アマゾン・ドット・コム	小売業	インターネットによる書籍等の販売	電子商取引によって発生した顧客の行動履歴のデータを分析することで，関連商品の紹介を行う販売促進や広告配信を効果的に展開している。	狭義のビッグデータの利活用	販売促進や広告配信に係る意思決定支援
VISA	金融業	クレジットカード事業	同社は，クレジットカード決済に関連する事業を展開しており，世界中にいる顧客のクレジットカードによる決済業務を行っている。それらの決済業務で得られた膨大なデータと顧客データを関連させ，不正利用と考えられる履歴を検出するシステムを構築し運用している。	狭義のビッグデータの利活用	顧客管理（不正利用の発見）に係る意思決定支援
グーグル	情報サービス業	情報検索サイトの運営に付随する広告業	Webサイトの検索サイトを運営し，検索順位に応じた広告収入を得るビジネスモデルを展開している。同時に，地図情報などのコンテンツ作成からメールサービスやOSの提供といった多岐に及ぶ事業を展開している。	広義のビッグデータの利活用	自社サービスの高付加価値化及び自社ブランド価値の向上

資料）各種文献を参考に筆者作成。

４　ビッグデータの分析に対するデータベース技術

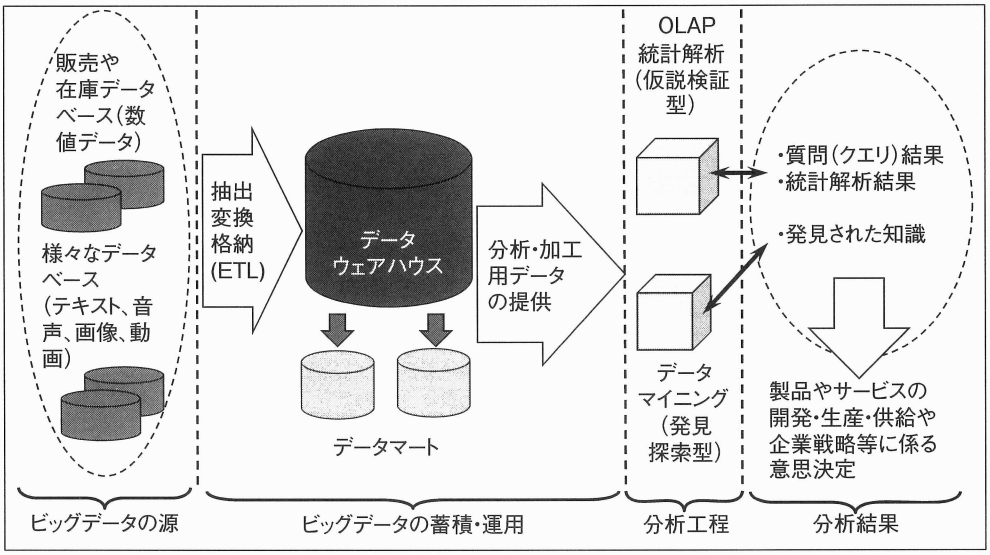
ビッグデータの利活用において，データを収集する段階においてはインターネットやセンサといったネットワーク関連技術が重要である一方，ビッグデータを蓄積し分析・加工する段階では，ビッグデータに対するデータの構成や分析に係る技術が必要となり，データベース技術が大きな役割を果たす。本章は，ビッグデータの分析を目的とする情報システムについて，主要技術となるデータベース技術の視点から情報システムの全体像を述べた後，分析の段階別に詳細を述べる。

４－１　ビッグデータの利活用を目的とした情報システム

ビッグデータの利活用において，広義のビッグデータの利活用に関連する情報システムは，利活用の目的が自社で扱う製品やサービスに大きく依存することから，汎用的な技術体系が存在するというよりは，各々の製品やサービスに特化した情報システムが構築されている場合が多い。また，製品やサービスの高付加価値化という企業の競争力の源泉に直結するものであり，詳細な情報システムは秘匿で不明な点も多く詳細な解説が困難である。したがって，汎用的な情報システムの開発が進んでおり，一定程度共通した情報システムが普及している狭義のビッグデータの利活用，すなわち，ビジネスインテリジェンスに係る情報システムについて述べていく。ただし，本章と第５章で述べるデータベース技術を活用したビッグデータの分析手法は，広義のビッグデータの利活用におけるデータの構成・処理・分析・加工の一部分に応用可能である。

ビッグデータを分析し意思決定支援を行っていく情報システムにおける要素技術の体系について，文献［８］，［33］，［46］，［50］を参照しまとめたものが図表６である。ビッグデータの分析に要する段階は以下の通りとなる。第一に，分析対象となるビッグデータの源となるデータベース[9]がある。第二に，それらを蓄積・運用する段階に移行し，データの構成法としてデータウェアハウスというデータベース技術がある。第三に，分析工程では，大別すると仮説検証型と発見探索型の分析があり，それらを実現するためのデータベース技術がある。最終的に，分析結果として仮説検証型の分析で得られた質問（クエリ[10]）に対する結果や統計解析の結果，発見探索型の分析で得られたデータマニングによる知識[11]が，製品やサービスの開発・生産・供給や企業戦略等に係る意思決定に反映される。これより，３つの段階別に詳細を述べる。

図表６　ビッグデータを分析し意思決定支援を行う情報システムの要素技術の体系



資料）文献［８］，［33］，［46］，［50］をもとに筆者作成。

４－２　分析対象となるビッグデータの源

企業において分析対象となるビッグデータには，日々の企業活動において発生する販売や在庫に関するデータベースがあり，それらは主に数値データである。これらは，一般に，表（リレーション）形式でデータを表現し演算を行うリレーショナルデータモデルをデータモデル[12]とするデータベース管理システム[13]で管理・運用されているトランザクションデータである。リレーショナルデータモデルを採用したリレーショナルデータベースでは，トランザクションデータ中における１件のトランザクションは，リレーションにおける各属性に対応する属性値の集合として表現され，属性値の対応関係を利用して処理や分析が行われる。このような従来のデータモデルに基づいて形式化されているデータは，構造化データと呼ばれる[35]。ビッグ

データは，大量の構造化データのみで構成される場合もあるが，それ以外にも企業内外における Web サイトのテキスト，コールセンターの音声データ，製品やサービスに関連する画像や動画も分析対象となる場合も多く，質的な多様性を有する。テキストや音声などのデータは，データ自身の内容を表現するタグをデータに付すことで一部を構造化して半構造化データとして扱う場合がある [35]。さらに，構造化が困難であるようなデータ，あるいは音声や動画などに対して構造化をせずに特定の目的の下で分析・加工を行うために構造化を行わないデータは，非構造化データとして扱われる。したがって，ビッグデータの分析対象には，構造化データ，半構造化データ，非構造化データのいずれも対象となる。

４－３　ビッグデータの蓄積・運用に関する段階

ビッグデータの蓄積･運用に係る段階について述べる。様々なデータベースを一元的に蓄積･運用していく工程における要素技術としてデータウェアハウスがある。データウェアハウスは，企業における経営層，管理者層，アナリストなどがより効果的かつ迅速に意思決定を行うのを支援するための技術であり，複数のデータベースを一元的に蓄積･運用し，分析に供されるデータを提供する役割を担う [41]。データウェアハウスは一元的なデータの蓄積･運用に関する技術の総称ではあるが，分析を目的としたデータベース管理システムとそれに格納されたデータベースを合わせたデータベースシステムであるともいえる。

データウェアハウスを構築するためには，複数のデータベースから，データを抽出 (Extract) し，統一した形式に変換 (Transform) し，データウェアハウスに格納 (Load) する必要がある。こうしたビッグデータの源となるデータベースからデータウェアハウスを構築するまでの段階は，各工程の頭文字をとって ETL と呼ばれ，ETL の自動化ツールなどが開発されている。

データウェアハウスは企業内外のデータを一元的に管理しているため，開発，生産，供給といった業務別や製品・サービスの種類別に分析用のデータを供給する場合，分析目的に対して無関係なデータも混在するため処理が煩雑かつ多大な負荷を要してしまう場合がある。したがって，データウェアハウスを業務別や種類別など特定の利用目的の下で分割したものをデータマートと呼び，データウェアハウスを分割したシステムが構成される場合もある。

分析に供されるデータは，格納されているデータ量や分析の目的に応じて，直接データウェアハウスから供給される場合やデータマートから供給される場合があるが，データの構成法や処理方法を規定するデータモデルに大きな差は生じないため，データマートを包含するデータウェアハウスに用いられているデータモデルについてのみ述べる。データウェアハウスに用いられるデータモデルの種類は格納するデータの質的側面による。ビッグデータが数値データのみで構成されていて構造化データとして処理できる，あるいはテキストや画像などであっても半構造化データとして処理すれば，半構造化によるタグはメタデータと呼ばれ構造化データとして処理できるので，主としてリレーショナルデータモデルを用いればよい。または，メタデー

タに対してキーワードによって検索を行うことで，半構造化データは蓄積・運用することができる[34]。ただし，質的な多様性の問題が解決されたといえども量的な問題が依然として残るため，重複データの効率的な処理といった手法は必要となる。また，データモデルというソフト面についてのみ述べているが，同時にハード面での高速処理技術なども重要である。

一方，テキスト，音声，動画といった多様なデータの多様性を維持したまま，あるいは多様性を生かした分析を行うのであれば，非構造化データとして運用していく。非構造化データを運用していくためのデータモデルは，NoSQL[14]のデータモデルと呼ばれ，リレーショナルデータモデル以外のデータモデルのことを示す[28]。NoSQLのデータモデルの特徴は，明示的なスキーマ[15]を必要としない，ACID特性[16]を満たさない場合も可とするといった特徴があり，言い換えれば，従来のデータモデルにおける厳密性や正確性を犠牲にする代わりに，ビッグデータの量的課題に対する高速処理や質的課題に対する柔軟な処理を可能にしたデータモデルといえる。これは，データウェアハウスが分析を目的としているため基幹業務系のデータベースシステムとは独立しており，基幹業務系におけるトランザクション処理においては許されない厳密性や正確性を放棄しても，基幹業務に影響を与えずに分析という目的を遂行できるからである。

非構造化データに関する代表的なNoSQLのデータモデルは以下の通りである[22]。まず，キーバリューストア型のデータモデルがあり，キー(Key)[17]とバリュー(Value)の組でデータを保持するデータモデルである。非常に単純な構造ゆえに，データの追加といった拡張性や障害による停止も少なく可用性にも富んでいる。カラム（列）指向型のデータモデルは，リレーショナルデータモデルでは，１件のデータを１行に対応させる行単位なのに対して，列単位でデータの処理を行う構造をとるデータモデルである。ゆえに，列単位による集計といった演算が高速に実現できる。ドキュメント指向型のデータモデルは，１件のデータをドキュメント形式で保存し，スキーマを設定しない構造をとる。ゆえに，ドキュメント形式であるブログなどのデータに対する蓄積・運用に適している。

分析対象となるビッグデータの質的な特徴や分析目的に応じて，これらのデータモデルを選択し，データウェアハウスを構築する。また，非構造化データに適したデータモデルは，データウェアハウスのデータモデルに限らず，ビッグデータの源となるデータベースにおいても非構造化データを蓄積・運用するデータモデルとして採用されている。

４－４　ビッグデータの分析に関する段階

ビッグデータに対する分析を大別すると，仮説検証型と発見探索型の２通りがある。仮説検証型の主な手法としては，OLAP (OnLine Analytical Processing) があり，構造化データや半構造化データを中心とした数値化されたデータに対して，特定期間における製品の売上といった全体の傾向や支店ごとの売上といった部分間の比較などの分析を行うためにデータの集約等の質問（クエリ）に対する答えを迅速に提供する[46]。OLAPを実現するために，複数の属性を持

つ数値化されたデータを各属性に次元を対応させ，多次元空間にデータを配置して処理を行う多次元データベースと呼ばれるデータモデルが用いられる[30],[44]。さらに，利用者の操作性を考慮してデータキューブと呼ばれる3次元のデータモデルが採用され，1断面，すなわち2次元において，例えば，A店舗における月次（時間という属性に対する次元）での，各商品（商品という属性に対する次元）の売上に関する集計表といった形式で提供される。データキューブには，様々な演算があり，それらを組合せることで集計が可能になる。例えば，断面を変える演算（スライシング）によってB店舗の集計結果が得られたり，次元を変える演算（ダイシング）によって店舗ごとの月次の集計結果が得られたり，次元の階層レベルを下げる演算（ドリルダウン）によって週次の集計結果が得られたり，次元の階層レベルを上げる演算（ドリルアップ）によって年次の集計結果が得られたりする。

OLAPによって，仮説の検証にある程度目途がついた場合，さらにその検証を厳密に行う際には検証内容に応じた統計解析手法を用いて，仮説検証を精緻化させていく。また，仮説の下で，分析目的が明確化している場合には，OLAPを介さず目的に応じた統計分析手法を適用していく場合もある。OLAPや統計解析は，利用者が分析を行う際に仮説を設定しそれを検証する形で行う。例えば，販売促進キャンペーンによって，ある地域においてある銘柄のビールがある期間においてどれだけ通常よりも多く販売されたかといった仮説が立てられる。続いて，仮説を検証するために必要なデータの集約結果が利用者に戻され，キャンペーンによる販売効果の検証がなされる。したがって，OLAPや統計解析など仮説検証型の分析を用いる際には，分析を行う利用者側において仮説が分析時に想定できている，あるいは自ら仮説を設定できるという状況や能力が必要になる。

一方，ビッグデータという膨大なデータゆえに仮説を設定することが困難な場合が多々ある。そうした場合には，仮説という明確な目標を設定せずに，膨大なデータから有用な情報，すなわち，知識を半自動的に抽出する技術がデータマイニングである。データマイニングの主な分析手法には，関連性を分析するアソシエーション，分類を行うクラスタリング，分類された集団の特徴づけを行うクラシフィケーションがある[14]。まず，アソシエーションとは，相関分析とも呼ばれ，商品群から同時購入商品といったルール[18]を導くものである。本手法では，例えば，有名な事例でいえば，紙おむつを買う客は缶ビールも買うことが多いといった，通常であれば想定しない仮説，すなわち，仮説設定が困難であるようなルールを抽出することができる。第3章で紹介したアマゾン・ドット・コムの販売促進の事例で用いられているのはアソシエーションである。次に，クラスタリングとは，分類基準が分かっていないデータをデータの類似性からいくつかの集団に自動的に分類する手法である。これにより，例えば，自動車の顧客プロフィールから，第一集団は若者向き，第二集団は家族向きの車種といった知識を抽出することができる。さらに，クラスタリングによって分類された集団の特徴づけを行うのがクラシフィケーションである。例えば，クラスタリングによって分類された若者向きの集団を，スポーツ

カーを購入することが多い顧客層は年収500万円以上，30代以上の男性であるといったようなより精緻な知識を抽出する。データマイニングの分析手法にはこれら以外にも，ビッグデータの利活用において多用される手法として例外値検出がある。例外値検出は，膨大なデータ中から外れ値とよばれる特定の基準から大きく外れる，もしくは明らかに誤りである値を発見できるので，例えば，クレジットカードの利用に関するトランザクションデータから不正と考えられるトランザクションを抽出し，クレジットカードの不正利用といった注意喚起を行うなどの顧客管理に応用されている。

5　階層的分類を用いた分析手法とビッグデータに対する有用性

本章は，著者の先行研究である階層的分類を用いた分析手法の概要について述べながら，階層的分類を用いた分析手法がビッグデータのデータ量の増大や多様性から生じる諸課題に対して応用可能であることを示す。

5－1　多様なデータに対する階層的分類の有用性

ビッグデータは，これまでの述べてきたように数値データ，テキスト，音声，画像，動画といった質的に多様なデータである。データの構造面に着目すると，構造化データ，半構造化データ，非構造化データに分類される。こうした生データ[19]としての多様性を持ち，構造面からも不均一なデータを一元的に分析に供するためには，多様なデータに対応するデータの構成法が必要になる。データの構成法として，非構造化データを構造化データに変換し，リレーショナルデータベースによって蓄積・運用するといった手法があることについて第４章で述べた。一方，生データを一元的に蓄積・運用する手法として，データの持つ意味に応じて分類階層[20]に対応するラベルを付すような階層的分類が有用である[49]。階層的分類は，データの構成の容易性や操作性に優れており，例えば，検索サイトにおけるカテゴリ検索に用いられている[51]。

階層的分類において，１件のデータが，分類に用いる属性に対して，単一の属性を持ち，ただ１つの終端クラス[21]に均質に分類されるデータであれば，データの構成法は単純である。同時に，それらのデータに対する分析手法も明快である。しかし，ビッグデータは質的な多様性を有していることにより，データの構成やデータを分析に供する際に解決すべき問題が生じる。

5－2　多様なデータが有する４種類の不均一性

多様なデータに対して階層的分類を行うと，以下に述べるような４種類の不均一性が生じる場合が多い。第一に，概念レベルの異なるデータであり，それらデータは終端クラスに分類されず，データには終端クラス以外のクラスのラベル[22]が付される。例えば，地域という属性における分類階層において，終端クラスが市町村レベルであったとすると，日本という意味のデータや愛知県という意味のデータなど概念レベルの異なるデータが混在している場合は，終端ク

ラスには分類されず，日本や愛知県のクラスにとどまり，各々に“日本”や“愛知”といったラベルが付される，すなわち，概念レベルの異なるラベルが付されることになる。

第二に，複数の意味を持つデータであり，そのデータは１つのクラスのみに分類されず，データには複数のクラスのラベルが付される。例えば，愛知県と福岡県の両方の意味を持つデータであれば，愛知県と福岡県の両方のクラスに分類され，２つのクラスのラベルを要素とする集合のラベル，｛愛知，福岡｝が付される。

第三に，クラスとデータの関係の強さが異なるデータであり，データには関係の強さを示すランクを有するラベルが付される。例えば，愛知県について主に関係するデータであるが福岡県についても若干関係のあるような意味のデータは，関係の強さを示すランクを主と副という２段階で設定した場合では，｛愛知（主），福岡（副）｝といったラベルが付されることになる。

第四に，複数の属性に対して意味を持つデータであり，データには，１つの属性に対するラベルの集合を要素とする集合のラベルが多重属性のラベルとして付される。例えば，産業と地域という２つの分類階層で分類した場合，自動車産業における愛知県と福岡県に関する意味のデータは，｛｛自動車｝，｛愛知，福岡｝｝といったラベルが付されることになる。

５－３　概念レベルが異なり複数の意味を持つデータの構成と解釈

データの不均一性の程度に応じて，４種類の不均一性に対し，すべての性質を持つデータから，そのうちの数種類が組み合わされたものまである。これらの４種類の不均一性は，相互に関連しつつ複合的にデータの不均一性の程度として現れる。４種類のすべての不均一性を同時に議論していくと複雑な議論になるので，概念レベルの異なるデータと複数の意味を持つデータの２種類の不均一性について着目し，これらのデータに対するデータの構成上の問題，ならびに不均一性を反映させた分析に供する際のデータの解釈上の問題を整理し，解決策を示す。

データの構成上，均一なデータであれば，親クラスのデータはいずれかの子クラスに属するという性質（充足性）と親クラスのデータは２つ以上の子クラスに属することはないという性質（排他性）を利用することで効率的なデータの蓄積・運用が可能になる。充足性により，非終端クラスのデータは終端クラスの和集合で求められるので，データを記録しておく必要がなく，データの挿入や削除といった更新が効率的に行われる．排他性により，データに重複がないため和集合の演算の負担も小さい。しかし，概念レベルの異なるラベルのデータは充足性を満たさず，複数のラベルを持つデータは排他性を満たさない。

次に，分類したデータを分析に供する際，“クラスのデータ”の解釈が問題になる。概念レベルの異なるデータでは，中部というクラスのデータは，中部，愛知など中部またはその下位概念のラベルを持つデータを対象にする場合と中部というクラスのラベルと同一ラベルを持つデータを対象とする２通りの解釈がある。また，複数の意味を持つデータでは，中部というクラスのデータに，中部以外に関するラベルを持つ｛愛知，福岡｝というラベルのデータを含め

るかどうかで２通りの解釈が生じ，２通りのデータを指定する手法が求められる。

まず，概念レベルの異なるデータで生じたデータの構成と解釈に対する解決手法を述べる[38]。データの構成に関しては，充足性を満たさないデータを格納する仮想クラスを不分類クラスとして，子クラスに分類できないデータを格納しているクラスに設置することで対応する。さらに，その不分類クラスにおいて，クラスのラベルと同一ラベルを持つデータを格納する仮想クラスとして固有クラスを設置することで２通りのデータの指定が可能になる。

次に，複数の意味を持つデータで生じたデータの構成と解釈に対する解決手法を述べる[7],[42]。データの重複が生じるという問題は，データが複数の子クラスに分類された際，そのうちの１つを選び，親クラスのデータとして代表させることで対応できる手法をＬ値として開発している。また，クラスにはデータの意味とは無関係なデータが分類されており，それらを含めるかで２通りのデータの指定が必要になる。さらに，１つのクラスではなく複数のクラスにまたがる問合せ，すなわち，ラベル集合による問合せも含めて考えると，複数のクラスに対するデータの問合せは，データのラベルが問合せに用いるラベル集合の範囲内であるようなデータのみを求める手法が必要とされる。その際，一般には，データが問合せのラベル集合に関係しないクラスに分類されていないかを他のすべてのクラスに対して調べる必要があり，多大な負担を強いる。そこで，先のＬ値と組合せて利用するＲ値を開発したことにより，そうした負担なく問合せに対応可能である。

５－４　分析に供するデータの指定

複数のラベルを持つデータのラベルはラベル集合なので，データのラベルが複数であることに基づいたデータの利用を想定すると，ラベル集合による問合せが必要になる。しかし，与えられたラベル集合に対応するデータの解釈は様々である。例えば，中部と九州というラベル集合に対応するデータの解釈には，“中部と九州のみに関するデータ”とし，{愛知，福岡} といったラベルが付されたデータを対象とする一方で，“中部と九州に関するデータ”として，{愛知，福岡，東京} のように中部や九州に無関係な東京を含むデータも対象としたい場合などもある。そこで，ラベル集合によって表現されるデータの解釈を整理した後，ラベル集合間の順序によってそれらが表現できることを理論的に明らかにしている[39]。同時に，ラベル集合によって表現されるデータの解釈は，複数のラベルを持つデータに対する分析対象の考え方と一致することが導かれたため，複数のラベルを持つデータを分析に用いる際の指定法として有用であることが示されている。

これまでの議論を，クラスとデータの関係の強さが異なるデータに拡張している。文献［５］では，関連性の強弱を示すランクを有するラベルをデータに付し，ランク付ラベル集合を用いて分析対象となるデータの指定法を構築している。さらに，ランク付ラベルを持つデータに対する分析対象の考え方とランクの強弱に起因する分析対象となるデータ集合の包含関係を明ら

かにしたことで，分析対象となるデータ集合の拡張や限定といったことが精緻に行えるのと同時に，分析対象の増減といった変化を利用した分析手法が可能であることも示している。また，文献［37］において，ランク付ラベルのデータの構成法として，ランク別にクラスのデータを生成し求める手法，ランクの特性を意識したクラス間の集合演算の定義，分類作業支援機能を持つデータの構成法も提案している。

最後に，複数の属性において意味を持つデータについて述べる。データの分類は，一般に分類に用いる属性ごとに分類され，分類階層において該当するカテゴリのラベルが付された後，複数の分類階層を複合的に用いることによって，分析対象となるデータを指定することになる[36]。文献［4］では，単一属性で用いた議論を複数の属性に拡張することで，多重属性を持つラベル集合による複数の属性を持つデータに対する記述法を明らかにし，それが十分な記述能力を有することを示している。ただし，複数の属性を持つデータに対する分析対象の考え方については言及していないため，多重属性を持つラベル集合によって記述されるデータが分析

図表7　不均一なデータに対する階層的分類を用いたデータの構成と指定

データの不均一性の種類	データの例	不均一性から生じるラベルの特徴	ラベルの例	データの構成		データの指定	
				問題点	解決手法	問題点	解決手法
概念レベルの異なるデータ	日本という意味のデータや愛知県という意味のデータが混在する場合	概念レベルの異なるラベル	“日本”といったラベルと“愛知”といったラベルが混在する場合	充足性を満たさない(子クラスに分類できない)	不分類クラスを設置	“クラスのデータ”の解釈が2通り	固有クラスにより2通りのデータを指定
複数の意味を持つデータ	愛知県と福岡県の両方の意味を持つデータ	複数のラベル	{愛知，福岡} といったラベル	排他性を満たさない(データが重複する)	L値の開発(重複データの効率的な処理手法)	クラスや複数クラスに対する問合せには無関係なラベルを含むデータが混在する	R値を開発しL値と組合せることで効率的に無関係のデータを排除したデータを指定
クラスとデータの関係の強さが異なるデータ	愛知県について主に関係するデータであるが福岡県についても若干関係のあるような意味のデータ	ランク付ラベル(主と副といった2段階を設定した場合)	{愛知(主)，福岡(副)} といったラベル	ランクを記憶したデータの構成	ランク別にクラスのデータを生成するデータの構成法	ランク付ラベルのデータに対する分析対象の考え方が不明瞭で指定法がない	分析対象となるデータの指定法を体系化
複数の属性に対して意味を持つデータ	自動車産業における愛知県と福岡県に関する意味のデータ	1つの属性に対するラベルの集合を要素とする集合のラベル(多重属性ラベル)	{{自動車}，{愛知，福岡}} といったラベル		関連研究において属性ごとに階層を用意し組合せて利用する構成が提案	多重属性ラベルのデータに対する分析対象の考え方が不明瞭で指定法がない	多重属性ラベルによって記述されるデータを体系化(※分析対象の考え方の議論に至っておらず記述法が指定法として十分か検証が必要)

資料）筆者らの先行研究をもとに筆者作成。

を目的とした場合の分析対象となるデータを包含できているかについては今後の研究課題である。以上，これまで述べてきたものについてまとめたものが図表７である。

６ おわりに

本稿は，ビッグデータに着目し，ビッグデータに関する現状を把握した上で，データベース技術の視点からみたビッグデータに対する分析の技術体系の概説とビッグデータの分析に対する階層的分類の有用性を明らかにした。

ビッグデータの利活用に対する関心は，ビッグデータに関する記事検索数からみると，2011年中頃から急速に記事件数が伸びていることから，直近の１年間において急速に関心が高まっていることが定量的に判明した。同時に，ビッグデータの利活用の現状は，ビッグデータを対象にしたビジネスインテリジェンス，すなわち，データを分析し社内の意思決定支援に利活用するという狭義のビッグデータの利活用と，ビッグデータを自社の製品やサービスの高付加価値化に利活用する広義のビッグデータの利活用に分類されることを示した。さらに，２通りの利活用について，国内外の代表的な事例を例示した。

後半では，データベース技術を用いたビッグデータの分析を目的とした情報システムについてまとめた。それらを踏まえて，ビッグデータのような多様なデータを一元的に分析に供する有用な手法として，データに対する階層的分類について述べた。先行研究は，以下の点でビッグデータに対する分析に対して有用である。第一に，階層的分類は多様なデータをデータの種類を問わず一元的に管理できるデータの構成法であり，ビッグデータは数値のみならず，テキスト，画像，動画といった多様なデータから構成されるため，ビッグデータに対するデータの構成法として応用可能である。第二に，ビッグデータの質的な多様性は分類時にデータの不均一性となって現れる。先行研究は，データの不均一性に着目した研究であり，不均一なデータに適したデータの構成法と不均一なデータを分析に供するための指定法を提案しているため，ビッグデータの蓄積・運用に係るデータウェアハウスのデータモデル，ならびにビッグデータの分析に係る OLAP における問合せ等に応用可能である。

ビッグデータの質的側面から生じる課題に対する先行研究の有用性を具体的に述べると，データの多様性から生じる不均一性は，概念レベルの異なるデータ，複数の意味を持つデータ，クラスとデータの関係の強さが異なるデータ，複数の属性における意味を持つデータという４種類の不均一性であると明らかにしたのと同時に，不均一なデータに対する構成法を提案している。これにより，ビッグデータの不均一性を維持したままデータの蓄積･運用が可能になり，ビッグデータが有する質的な多様性を分析に反映できる。さらに，不均一なデータに対する分析対象の考え方とそれに対応する指定法が構築されているので，ビッグデータを分析する段階になった際のデータ供給に応用ができ，例えば，OLAP における問合せの精緻化などに有用である。さらに，多重属性についても言及しているので多次元データベースにおけるデータモデ

ルの一部にも応用可能である。

先行研究は，ビッグデータの質的側面のみならず量的側面から生じる課題に対しても有用である。クラスのデータは，不分類クラスによって関連するクラスのデータから生成できるのでデータの更新時の負担が少ない。さらに，L値により，重複データを排除した演算が可能なのでクラスのデータを生成時に効率的に処理できる。また，複数の意味を持つデータに対するラベル集合を用いた問合せでは，問合せに関連するクラスのデータ数を n，全データを N とすると，一般的に全データを検証する場合と比較して，L値とR値により n/N の処理で済むことになり，ビッグデータのような N が大きい場合において本手法の有用性が高まる。

本稿で述べてきた先行研究は，分析を目的としたデータの蓄積・運用に関連したものが多い。今後は，分析工程に関連する研究として，例えば，OLAPにおける問合せに対して，不均一性を利用した分析を体系化するなどし，データの蓄積・運用から分析まで一貫した理論的枠組みを構築していく予定である。

謝　辞

磯村孝志先生には，筆者が大学教員となり右も左もわからない中で，研究，教育，校務において様々なお気遣いを頂き，ここに感謝申し上げます。先生は，社会科学の分野における情報通信技術の重要性に早くから注目をし，研究や教育，さらには学科設立において多大なるご貢献をされてきました。筆者も未熟ながら文理融合分野の経済工学を専攻しており，先生の理念に共感するのと同時に，先生からご指導・ご鞭撻を受けることで今日に至ることができました。先生のご健康とご活躍をお祈りし，これからも先生の理念を引き継ぎ，研究，教育，さらには校務と日々精進していくことを誓い，拙稿を捧げます。

注

1）SNSとは，Social Networking Serviceの略称であり，知人間における交流を意図したコミュニティサイトである。
2）ストレージとは，データを記憶しておく装置のことであり，ストレージ関連事業とは，データ記憶装置およびその周辺装置の製造やストレージに関連する種々のサービスを展開する事業のことである。
3）コンテンツの定義は文献により異なるが，文献［9］によると，一般の鑑賞や使用に耐えうる，あるいはひとまとまりの形式になっており，娯楽や教養といった目的にそれ自体がかなう特徴をもつ情報とされる。本稿におけるコンテンツも同様のものとする。ただし，コンテンツに該当するのか，単体の情報でありコンテンツではないのかといった明確な基準は存在せず，本稿における議論においても同様であり，コンテンツか否かについては議論の余地が残る場合がある。
4）ITSとは，Intelligent Transport Systemsの略称であり，高度道路交通システムと訳される。その定義は，国土交通省交通局ITSホームページによると，『道路交通の安全性，輸送効率，快適性の向上等を目的に，最先端の情報通信技術等を用いて，人と道路と車両とを一体のシステムとして構築する新しい道路交通システムの総称。』とされる。
5）Twitterとは，主要なSNSの1つであり，文字制限が設けられている中でコメントを打つという特徴がある。
6）BIとは，Business Intelligenceの略称であり，ビジネスインテリジェンスを示す用語として多用されている。

7) Suicaとは，鉄道やバスの運賃や商品の購買時の決済に利用できるJR東日本が発行するICカードのことである。
8) グーグルは，スマートフォン向けにアンドロイド，パーソナルコンピュータ向けにクロームといったOperating Systemを開発し提供している。
9) データベースとは，体系的に格納されたデータ集合のことである。ただし，データベースという用語は，データ集合という狭義の意味以外にも，データベースを管理・運用するためのソフトウェア（データベース管理システム）や狭義のデータベースをデータベース管理システムによって管理・運用するシステム（データベースシステム）を表す場合がある。本稿では，データベースという用語は狭義の定義で用いているが，文脈から明確な場合には，データベース管理システムないしはデータベースシステムに該当する場合においてもデータベースという用語を用いている。
10) クエリとは，Queryのことであり，データベースにおける問合せ処理のことである。
11) 知識とは，本稿では，情報やデータが体系的にまとめられたもので，将来の判断や思考に活用できるものと考える。
12) データモデルとは，データベースモデルとも呼ばれ，データ構造やデータないしデータ間における操作を規定する枠組みといえる。主なデータモデルには，リレーショナルデータモデル以外にもネットワークデータモデルや階層型データモデルなどがある。
13) データベース管理システムとは，データベースを管理・運用するためのソフトウェアのことである。
14) NoSQLとは，非SQLという意味ではなく，“Not Only SQL”という意味であって，SQLに限定しない処理を併せ持つデータモデルと考えられる。SQLとは，Structured Query Languageのことであり，リレーショナルデータモデルにおける種々のクエリを操作する言語である。
15) スキーマとは，特定のデータモデルに基づき，データベース中のデータに対する構造，関連，制約などを具体的に記述したものである。
16) ACID特性とは，Atomicity（原子性），Consistency（一貫性），Isolation（隔離性），Durability（永続性）の４つの性質のことである。原子性とは，トランザクションで行うすべての変更がなされるか，すべて変更しないかのいずれかでなくてはならないことを保証するという性質である。一貫性は，トランザクションによって矛盾した状態にならないことを保証する性質である。隔離性とは，複数のトランザクションが平行に実行されていても互いに干渉しないことを保証する性質である。永続性とは，トランザクションが成功して完了した場合，その変更が保たれることを保証する性質である。
17) キーとは，属性の集合の値が決まれば，一件のデータを一意に決定することができる（一意識別能力を有する）属性の集合のことである。
18) 同時購入に関するルールは，“商品Aを購入した顧客のうち，70%が商品Bも一緒に購入している”といった情報であり，“相関ルール (Association Rule)”と呼ばれる。相関ルールは確信度と支持度という基準があり，それらの閾値を設定した上で，閾値を超えるようなルールを有益な情報とする。詳細は，文献［31］による。
19) 生データとは，データの発生ないしは制作時のままの状態のデータを示し，未加工なデータのことである。例えば，画像であれば，デジタルカメラ等で撮影した際に発生したデータであり，それが何を意味するかといったメタデータ等の付加などの処理を加えていない状態のデータである。
20) 分類階層とは，特定の属性下において，概念が体系的にまとめられたもので，シソーラスと呼ばれるものが該当する。例えば，地域という属性であれば，アジア，東アジア，日本，中部，愛知といった階層的な概念体系が示されたものが分類階層である。
21) 終端クラスとは，分類階層における概念レベルが最も下位であり，子となるクラスを持たないクラスのことである。
22) ラベルとは，階層的分類の際，所与の分類階層の下でデータを分類時にデータが分類される最下層のクラスを示しているものであり，データの意味を示すためのものとする。

参考文献

［1］雨宮寛二：アップル，アマゾン，グーグルの競争戦略，NTT出版，(2012).
［2］エヌティティデータ技術開発本部ビジネスインテリジェンス推進センタ：BI革命，NTT出版，(2009).
［3］尾代智之，梅津高朗：センサ，デバイスによる新たな情報と高度交通システム，情報処理（通巻571号），pp.1040-1046，(2012).

［4］葛西正裕：多重属性を持つラベル集合を用いたデータの記述，愛知学院大学産業研究所所報「地域分析」，Vol.49，No.1，pp.43-65，(2010).
［5］葛西正裕：ランク付ラベル集合による分析を目的としたデータの指定と性質，愛知学院大学論叢「商学研究」，Vol.52，No.3，pp.1-17，(2012).
［6］葛西正裕，藤井学：デジタルコンテンツを活用した産業の高付加価値化の現状と課題，九州経済調査月報（通巻767号），pp.3-12，(2010).
［7］葛西正裕，古川哲也：階層的分類における複数の意味を持つデータの利用，情報処理学会論文誌：データベース，Vol.47，No.SIG8 (TOD30)，pp.1-10，(2006).
［8］定道宏：ビジネス情報学概論，オーム社，(2009).
［9］佐藤典司：モノから情報へ　価値大転換社会の到来，財団法人経済産業調査会，(2012).
［10］城田真琴：ビッグデータの衝撃，東洋経済新報社，(2012).
［11］鈴木良介：ビッグデータビジネスの時代，翔泳社，(2011).
［12］トーマス・H・ダベンポート，ジェーン・G・ハリス：分析力を武器とする企業，日経BP社，(2008).
［13］トーマス・H・ダベンポート，ジェーン・G・ハリス：分析力を駆使する企業，日経BP社，(2011).
［14］徳山豪，森本康彦，福田剛志：データサイエンス・シリーズ③　データマイニング，共立出版，(2001).
［15］長橋賢吾：ビッグデータ戦略，秀和システム，(2012).
［16］日経コンピュータ2009年11月25日号，日経BP社，pp.106-109，(2009).
［17］日経コンピュータ2009年12月23日号，日経BP社，p18，(2009).
［18］日経コンピュータ2012年１月５日号，日経BP社，pp.28-29，(2012).
［19］日経コンピュータ2012年２月２日号，日経BP社，pp.34-35，(2012).
［20］日経コンピュータ2012年２月２日号，日経BP社，pp.37-38，(2012).
［21］日経コンピュータ2012年２月２日号，日経BP社，pp.41-42，(2012).
［22］日経コンピュータ2012年８月30日号，日経BP社，pp.56-65，(2012).
［23］日経産業新聞2012年８月14日号，日本経済新聞社，p12，(2012).
［24］日経情報ストラテジー2011年４月号，日経BP社，pp.44-46，(2011).
［25］日経ビジネス2012年５月７日号，日経BP社，pp.47-48，(2012).
［26］野村総合研究所ICT・メディア産業コンサルティング部：ITナビゲーター2012年版，東洋経済新報社，(2011).
［27］野村総合研究所イノベーション開発部：ITロードマップ2012年版，東洋経済新報社，(2012).
［28］松本雅和：NoSQLの世界，情報処理（通巻548号），pp.1327-1331，(2010).
［29］元橋一之：日米韓企業のIT経営に関する比較分析，*RIETI Discussion Paper Series 07-J-029*，独立行政法人経済産業研究所，(2007).
［30］Agrawal, R., Gupta, A., and Sarawagi, S. : Modeling Multidimensional Databases, *Proc. the 13th Int'l Conf. on Data Engineering*, pp.234-243, (1997).
［31］Agrawal, R., Imielinski, T., and Swami, A. : Mining Association Rules between Sets of Items in Large Databases, *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pp. 207-216, (1993).
［32］Brynjolfsson, E., Hitt, L., and Kim, H. : Strength in Numbers: How Does Data-Driven Decision-Making Affect Firm Performance?, *Proc. the 32nd Int'l Conf. Information Systems, Economics and Value of IS*, pp.1-18, (2011).
［33］Chaudhuri, S., Dayal, U., and Narasayya, V. : An Overview of Business Intelligence Technology, *Communications of the ACM*, Vol.54, No.8, pp.88-98, (2011).
［34］Chen, Y., Wang, W., Liu, Z., and Lin, X. : Keyword Search on Structured and Semi-Structured Data, *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pp. 1005-1009, (2009).
［35］*Data Modeling : Database Normalization, Data Model, Conceptual Schema, Key Field, Weak Entity, Meta-Object Facility, Tuple Relational Calculus*, Books LLC, (2010).
［36］Furukawa, T. : Multiple Classification Hierarchies in Cooperative Databases, *Advanced Database Syst. for Integration of Media and User Environments'98 Advanced Database Research and Development Ser.* World Scientific, Vol.9, pp.309-314, (1998).
［37］Furukawa, T. and Kuzunishi, M. : Classification and Utilization of Data Belonging to Multiple Classes, *Proc. The 8th World Multiconference on Systemics, Cybernetics and Informatics*, Vol.2, pp.289-294, (2004).

[38] Furukawa, T. and Kuzunishi, M. : Hierarchical Classification of Heterogeneous Data, *Proc. the 23rd IASTED International Conference on Databases and Applications*, pp.252-257, (2005).
[39] Furukawa, T. and Kuzunishi, M. : Multi-Labeled Data Expressed by a Set of Labels, *Proc. World Academy of Science, Engineering and Technology*, Vol.65, pp.857-863, (2010).
[40] Greengard, S. : Big Data Unlooks Business Value, *Baseline January 2012*, pp.21-23, (2012).
[41] Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P. : *Fundamentals of Data Warehouses*, Springer, (2010).
[42] Kuzunishi, M. and Furukawa, T. : Representation for Multiple Classified Data, *Proc. the 24th IASTED International Conference on Databases and Applications*, pp.135-142, (2006).
[43] Liu, L. and Oezsu, T. : *Encyclopedia of Database Systems Volume 1*, Springer, pp.287-288, (2009).
[44] Liu, L. and Oezsu, T. : *Encyclopedia of Database Systems Volume 3*, Springer, pp.1777-1784, (2009).
[45] Mikroyannidis, A. and Theodoulidis, B. : Ontology Management and Evolution for Business Intelligence, *International Journal of Information Management*, Vol.30, pp.559-566, (2010).
[46] Mosley, M., Brackett, M., Earley, S., and Henderson, D. : *The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide)*, DAMA International, (2010).
[47] Negash, S. : Business Intelligence, *Communications of the Association for Information Systems*, Vol. 13, No. 1, pp.177-195, (2004).
[48] Rogers, S. : BIG DATA is Scaling BI and Analytics, *Information Management*, pp.15-18, (2011).
[49] Theodoratos, D., Dalamagas, T., Koufopoulos A., and Gehani, N. : Semantic Querying of Tree-Structured Data Sources Using Partially Specified Tree Patterns, *Proc. Int'l Conf. on Information and Knowledge Management*, pp.712-719, (2005).
[50] Turban, E., Sharda, R. and Delen, D. : *Decision Support and Business Intelligence Systems*, PEARSON, (2011).
[51] Wang, Y. and Oyama, K. : Web Page Classification Based on Surrounding Page Model Representing Connection Type and Directory Hierarchy, *IPSJ Transacitons on Databases*, Vol.2, No.2, pp.29-43, (2009).
[52] White, M. : Big Data-Big Challenges, *EContent November 2011*, p21, (2011).