

大量・多種多様な非構造化データを扱う
情報処理基盤

—データ再目的化時代の到来—

Information Processing Architecture of Mass and Various Unstructured Data

児玉 昇司
Kodama Shoji額賀 信尾
Nukaga Nobuo植田 良一
Ueda Ryoichi井口 慎也
Iguchi Shinya

爆発的に増加し、多様化するデータを分析し、主業務以外の別目的で活用したいというニーズがさまざまな分野で拡大している。従来のデータ分析ソフトウェアは主に構造化された数値データを扱ってきたが、近年は文書や画像など人が作成した非構造化データの扱いが求められている。

従来、非構造化データはデータサイズやデータ量が大きい処理に時間がかかり、また、数値データとは異なりそのままの形式では機械処理できないため、分析が困難であるという課題があった。

日立グループは、これまで、ストレージシステムなどインフラ技術や大量データ処理技術、テキスト・音声・画像などメディア処理技術の研究開発を進めてきた。今後はこれらの技術を発展・融合させ、非構造化データを統一的に管理・活用可能とするコンテンツ活用プラットフォームの研究を推進する予定である。

1. はじめに

「データは21世紀の石油」という言葉を裏付けるように、データ分析向けソフトウェア市場は2011年には100億ドルを超え、その後も高い成長率を維持するとの予想がある¹⁾。

これまで分析対象となるデータは、機械による計算処理が容易な売上額や在庫数など、数値化あるいは定形化された構造化データが主であった。しかし、企業が有する全データ量に対する構造化データの割合はわずか20%にとどまり、残りはテキストや画像、音声など人間が生成した

非構造化データが占めている²⁾。

近年、医療や金融、企業情報、政府機関、ビデオ監視分野などさまざまな分野において、従来は蓄積・保管・参照するだけであった非構造化データを分析し、学術研究やマーケティングなど主業務以外の別目的で有効活用したいという再目的化(Re-purposing)へのニーズが高まっている。しかし、非構造化データは機械による計算処理が難しいという課題があり、これまで十分活用されてこなかった。

このような中、日立グループは非構造化データの再目的化を支える技術として、大量データ処理技術と情報抽出技術に注目している。

ここでは、企業内に蓄積された多様な非構造化データの再目的化ニーズの動向と技術課題、これまでの研究の取り組み、および日立グループが考えるコンテンツ活用プラットフォームの概要について述べる。

2. ニーズ動向と技術課題

2.1 非構造化データ再目的化へのニーズ動向

現在、さまざまな分野で、大量の非構造化データを対象にした検索や分析などの再目的化ニーズが高まっている。分野ごとのニーズ動向を表1に示す。

医療分野ではICT (Information and Communication Technology) を段階的に導入してきたため、部門やアプリ

表1 | 非構造化データの再目的化ニーズ動向

文書や画像、音声など非構造化データを分析しビジネスに活用するニーズが、さまざまな分野で高まっている。

分野	ニーズ動向	再目的化例	非構造化データ
医療	ヘルスケア分野でのICT利用	教育・学術利用や診断根拠管理、DPC分析など、医療データの分析・活用 ³⁾	スキャンデータ、医療画像、録音メモ
企業情報	電子証拠開示 (e-Discovery)	メールなど電子的な証拠データから特定人物の過去の行動履歴・理由を推測	メール、書類、ログ
映像監視	監視映像のマーケティング適用	自動販売機が撮影した動画から人の視線を検知し、商品陳列順序を改善 ⁴⁾	監視映像
政府	Data.gov	政府が所有するさまざまなデータを定型化し、40万件近くWeb上で公開 ⁵⁾	アンケート結果、倒産銀行リスト
金融	株価予測へのソーシャルメディア活用	ある企業の株価とソーシャルメディアにおける人気度との関連性を研究 ⁶⁾	ユーザー生成コンテンツ

注：略語説明 ICT (Information and Communication Technology)、DPC (Diagnosis Procedure Combination)

ケーションごとにシステムが分断し、異なるベンダー間では相互接続性が低いという課題があった。近年はDICOM (Digital Imaging and Communications in Medicine) などデータ形式の標準化によってシステム間のデータ連携が進んでいるが、病院内にはいまだにスキャンデータや録音メモなど多くの非構造化データが存在する。こういったデータを学術目的や診断根拠管理目的などで利用したいというニーズがある。

企業情報分野では、訴訟に関する電子証拠を社内に散在するメールやファイルなど非構造化データから検索、集約し、開示要否を分析するe-Discovery (電子証拠開示) 制度への対応が重要になっている。

これら以外にも映像監視、政府、金融といった分野で非構造化データの再目的化ニーズが高まっている。こういったニーズに対応するため、さまざまなシステムに蓄積されたデータを集約管理し、再目的化するための分野共通的な基盤が求められている (図1参照)。

2.2 データ再目的化時の課題

データベースを代表とする構造化データと比較し、非構造化データの再目的化には、データサイズやデータ量が大きいと処理時間がかかり、また、データ形式が数値データのように機械可読ではないため統計処理などデータ分析が難しいという課題がある。

例えば、大量の画像データから類似した画像を高速に検索するには、複数台のコンピュータによる類似度計算処理の並列化や、HDD (Hard Disk Drive) 上でのデータ配置

の工夫など大量データ処理に向けた基盤技術が必要となる⁷⁾。また、紙媒体の医療カルテをスキャンした画像を対象に類似症例検索や統計処理など分析処理を行うためには、スキャン画像から検査結果などの数値データとその意味、固有表現とその属性値などコンピュータが処理可能な情報を抽出し、構造化データとする技術が必要となる。

このように、非構造化データの再目的化を支援するコンテンツ活用基盤を実現するには、大量データ処理技術と情報抽出技術が重要になると考えている。

3. 従来技術と日立グループの取り組み

データ再目的化を支える大量データ処理技術と情報抽出技術の取り組み状況について以下に述べる。

3.1 大量データ処理技術

気象シミュレーションをはじめとする大量データ処理の歴史は、スーパーコンピュータの発展によって牽(けん)引されてきた。1990年代まで、専用のハードウェアで構成されてきたスーパーコンピュータにおいて、2000年以降、汎用CPU (Central Processing Unit) と汎用OS (Operating System) を活用したシステムが主流になると合わせて、汎用のPCサーバを多数接続して、大規模なクラスターシステムを構築し、大量データのリアルタイム処理を実現する技術が急速に発展した。この技術革新を牽引したのがインターネット検索サービスを提供するGoogle^{*1)}、オンライン販売サイトを運営するAmazon^{*2)}、SNS (Social Networking Service) サイトを運営するFacebook^{*3)}などいわゆる「ネット列強」である。

特に、2004年にGoogleが発表した、大規模クラスター上での大量データ並列分散処理向けフレームワークMapReduce⁸⁾に注目が集まっている。

MapReduceは、MapフェーズとReduceフェーズから成る。Mapフェーズでは入力データをより小さい単位に分割し、クラスター内の個々のマシンに分割された別々のデータを割り当てて並列処理する。Mapフェーズでの処理は、別マシンの処理と非依存になるように実装する。Reduceフェーズでは、個々のマシンからの個別の出力を束ねて、最終出力を生成する。最終出力の生成に必要なデータを互いに非依存な部分データ集合に分割できる場合、Reduceフェーズも並列実行可能となる。

MapReduceは、オープンソースソフトウェアHadoop^{9) ※4)}

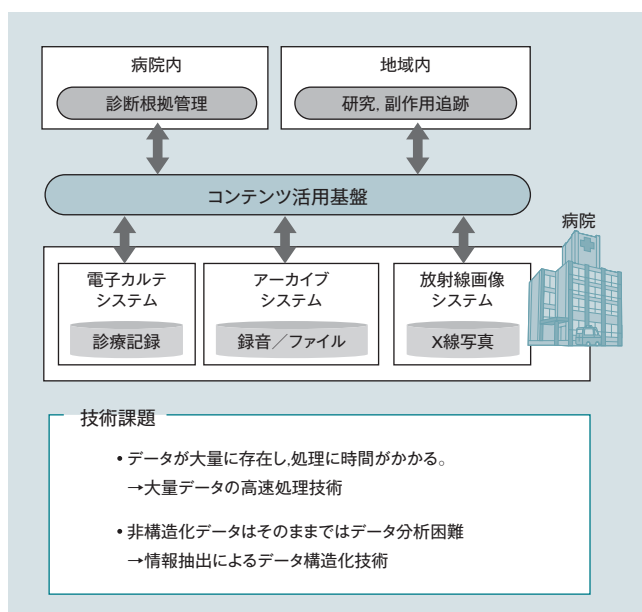


図1 | 医療におけるデータ再目的化の事例と技術課題

さまざまなシステムに格納されたデータを横断的に分析・活用するコンテンツ活用基盤を実現するには、大量データ高速処理技術と情報抽出によるデータ構造化技術が必要である。

※1) Google, Google Squaredは、Google Inc.の登録商標である。

※2) AmazonおよびAmazonのロゴは、Amazon.com, Inc.またはその関連会社の商標である。

※3) Facebook, Facebookロゴは、Facebook, Inc.の登録商標である。

※4) Hadoop, Apacheは、Apache Software Foundationの商標である。

としてApache^{※4)}プロジェクトにおいて実装され、改良が進められている。

Hadoopを活用することで容易かつ安価に大量データ処理システムを構築できるため、多くの分野で適用が進められている。

一方、データをいったんRDB (Relational Database) に蓄積し、必要に応じて読み出して処理する従来のストック型ではなく、発生するデータの流れに対して処理を行うストリーム型処理にも注目が集まっている。ストリーム型処理では、データが到着した時点であらかじめ登録された処理を即座に実行／結果を出力することで、大量データのリアルタイム処理を実現する。

日立グループは、2008年からMapReduce上に機械学習や頻出パターン抽出など、高度なデータ分析アルゴリズムを実装し、大量データから有益な情報を抽出／提供するサービス (KaaS: Knowledge as a Service) の研究を開始し、保守分野などへの適用を進めている¹⁰⁾。また、SQL (Structured Query Language) 言語を拡張したCQL (Continuous Query Language) で処理内容を簡単に記述できるストリームデータ処理基盤を2009年に製品化した。

3.2 情報抽出技術

インターネットの世界では、大量のテキストデータを処理して有用なアプリケーションを構築する動きが広がっている。Googleは、検索結果を表形式で提示するGoogle Squared¹¹⁾ ^{※1)}を実験公開している。Google Squaredでは、通常のキーワード検索とは異なり、検索キーワードの下位に属する名称と属性を表形式で表示する。例えば、「cat」というクエリに対して、「American Shorthair」、「Persian」などの名称が、画像や説明とともに表示される。画像や説明は、インターネット上のリソースへのリンクとなっている。表示する名称や属性はユーザー自身で拡張できるインタフェースを備える点も特徴の一つとなっている。また、IBM^{※5)}は「Watson^{※5)}」という質問応答 (QA) システムを構築し、米国の人気クイズ番組に挑戦して最高金額を獲得した¹²⁾。自然言語で表現されるバラエティに富んだ複雑なクエリから解答を瞬時に求める。

これらの技術に共通するのは、異なるリソースからの情報を集約し、有用な知識を獲得して活用している点である。これまでは、Webページを高速に検索することが検索エンジンの主たる目的であったが、今後は、大量の文書からいかに有用な情報を抜き出すかという「情報抽出」の技術が鍵になると考えられる。また、映像データや音声

※5) IBM, Watsonは、米国およびその他の国における米国International Business Machines Corp.の登録商標である。

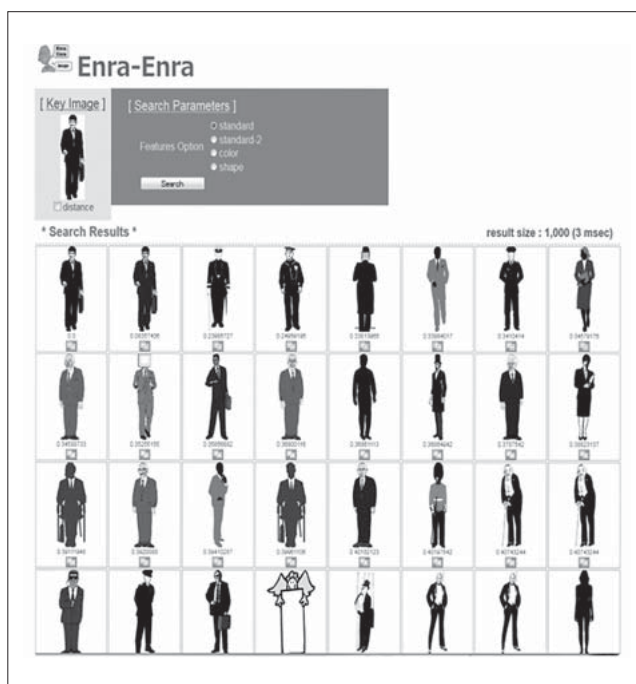


図2 | 類似画像検索プラットフォーム「EnraEnra」

画像から類似情報を抽出し、検索に適用している。これは、画像ごとに特徴量を計算し、多数の画像間の類似度を決定することで可能となる。

データなど、いわゆるマルチメディア情報からの情報抽出技術には課題が山積している。

日立グループはこれまで、同義語抽出¹³⁾、書誌情報の抽出¹⁴⁾など、文書からの情報抽出技術に取り組んできた。同義語抽出は、従来、人手によって編纂 (さん) されてきた同義語辞書を自動的に作成するための技術である。既存の同義語辞書を教師データとして用いることで抽出精度を向上させた¹³⁾。書誌情報の抽出は、PDF (Portable Document Format) 文書の検索容易性を高めるため、タイトルや著者などのメタ情報を文書の中から自動的に抽出する技術である¹⁴⁾。また、高速類似画像検索⁷⁾、音声検索¹⁵⁾などメディアデータに対する検索処理の開発も行っている。

今後は、あらゆるデータの再目的化が可能なソリューションの実現に向けて、これらの技術を活用したマルチモーダル情報からの情報抽出の研究開発を進めていく (図2参照)。

4. コンテンツ活用プラットフォーム

非構造化データの容易な再目的化の実現には、多様な形式の大量データを高速に処理し、計算機が分析可能な構造化された形式の情報を抽出できる、分野共通的に適用可能なコンテンツ活用プラットフォームが求められる。

また、このプラットフォームには、組織や地域に分散した多様な形式の非構造化データを統合的に管理し、既存システムとも連携してさまざまな目的で容易に活用できる環境の提供も求められる。

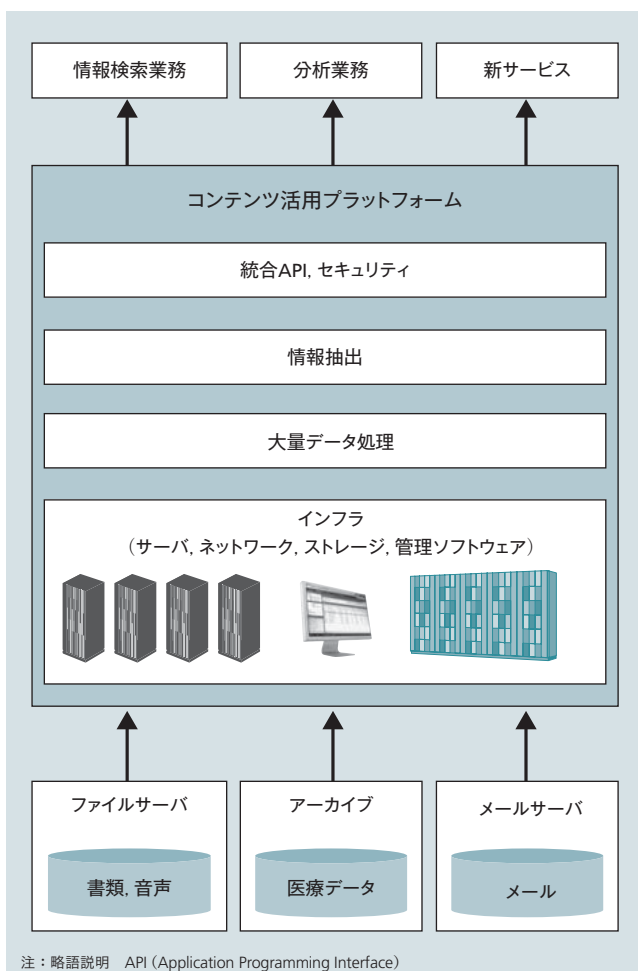


図3 | コンテンツ活用プラットフォーム

複数システムに散在する非構造化データの統合的な管理・活用を可能とする。

日立グループは、長年培ってきたストレージシステムなどプラットフォーム技術や大量データ処理技術、メディア処理技術をさらに発展させ、このコンテンツ活用プラットフォームの実現に向けた研究を推進する予定である（図3参照）。

5. おわりに

ここでは、企業内に蓄積された多様な非構造化データの再目的化ニーズの動向と技術課題、これまでの研究の取り組み、および日立グループが考えるコンテンツ活用プラットフォームの概要について述べた。

マルチメディアからセンサー情報に至るまで多種多様な非構造化データが増加し、活用ニーズも高まる中、人に代わってこれらの情報を解析して扱う技術が進化することは、人間と同等な認識能力を持つシステムの構築につながり、人と情報システムがより自然な形でコミュニケーションできる環境の実現を促す。その結果、人々はあたかも人間の同僚のような感覚で情報システムを使うことで、より

創造的で情緒豊かな関係を情報システムと築くことができ、人間味と優しさにあふれた知識創造社会の実現を加速すると日立グループは考えている。このような社会の実現に向けた研究開発を今後とも推進していく。

参考文献など

- 1) Gartner : Gartner Forecasts Global Business Intelligence Market to Grow 9.7 Percent in 2011, <http://www.gartner.com/it/page.jsp?id=1553215>
- 2) Datacentrix : 5th SARMAF Seminar Non Proprietary (Open Source) VS Proprietary Software, July 2009
- 3) 渡邊, 外:ヘルスケア分野のICT利活用と日立グループのソリューション, 日立評論, 93, 3, 292~297 (2009.12)
- 4) 視線検知技術をたばこ自動販売機マーケティングへ活用するための実証実験, <http://www.hitachi.co.jp/Div/jkk/research/jt/>
- 5) Data.gov, <http://data.gov>
- 6) Facecount : New study finds link between social media popularity and stock prices, <http://www.facecount.com/news/new-study-finds-link-between-social-media-popularity-and-stock-prices-242652>
- 7) D. Matsubara, et al. : High-Speed Similarity-Based Image Retrieval with Data-Alignment Optimization Using Self-Organization Algorithm, ISM2009
- 8) J. Dean, et al. : MapReduce: Simplified Data Processing on Large Clusters, OSDI 2004
- 9) Apache Hadoop Project, <http://hadoop.apache.org/>
- 10) 植田, 外: 社会インフラの革新に貢献する知識化サービス基盤KaaS, 日立評論, 92, 5, 362~365 (2010.5)
- 11) Google Squared, <http://www.google.com/squared>
- 12) IBM質問応答システム“ワトソン”がクイズ番組に挑戦!, <http://www-06.ibm.com/ibm/jp/lead/ideasfromibm/watson/>
- 13) 森本, 外: 文脈類似度と表記類似度を用いた教師あり同義語抽出, 言語処理学会年次大会 (2010.3)
- 14) 藤尾, 外: レイアウト解析による書誌情報の抽出, 情報処理学会全国大会 (2010.3)
- 15) N. Kanda, et al. : Open-vocabulary keyword detection from super-large scale speech database, MMSP, 2008

執筆者紹介



児玉 昇司

1998年日立製作所入社, 横浜研究所 情報プラットフォーム研究センター ソフトウェアプラットフォーム研究部 所属
現在, 非構造化データ活用システムの研究開発に従事
情報処理学会会員



額賀 信尾

1994年日立製作所入社, 中央研究所 情報システム研究センター 知能システム研究部 所属
現在, 音声処理システムの研究開発に従事
人工知能学会会員, 情報処理学会会員, 電子情報通信学会会員



植田 良一

1994年日立製作所入社, 横浜研究所 情報プラットフォーム研究センター ソフトウェアプラットフォーム研究部 所属
現在, 大量データ処理システムの研究開発に従事



井口 慎也

1998年日立製作所入社, 横浜研究所 情報プラットフォーム研究センター ソフトウェアプラットフォーム研究部 所属
現在, 非構造化データ活用システムの研究開発に従事
情報処理学会会員