1A3-OS-17b-2

OS-17 「ビッグデータと AI 技術」招待講演
Invited talk on OS-17 “Big Data and Artificial Intelligence”

非構造化データを扱う情報処理基盤の実現を支えるメディア処理技術

Media processing technologies for repurposing of large-scale unstructured data

池田 尚司　額賀 信尾
Hisashi Ikeda　Nobuo Nukaga

(株)日立製作所中央研究所
Hitachi Ltd., Central Research Laboratory

1. はじめに

ビッグデータという言葉が企業情報システムの分野で急速に浸透している．その定義は必ずしも明確ではないが，容量，種類，頻度 (リアルタイム性) といった観点で大規模であるデータを表すというという点についてはおおよそのコンセンサスは得られつつある．また，データそのものに加えて，こうしたデータを ICT によって利活用することで何らかの価値を生み出していくサービスまで含めてビッグデータという言葉で語られているのが現状である．

データの利活用により価値を生み出すための技術の歴史は古く，1960 年代の RDB とそれを操作するための SQL によって，蓄積したデータの分析が計算機上で行われたのに始まり，90 年代にはデータマイニングが登場し，データの間の相関の抽出が行われた．産業界においてもこうした技術をもとに，90 年代以前より小売業などで POS データを用いた CRM(Customer Relationship Management) が実現されている他，データウェアハウス，BI(ビジネスインテリジェンス) ツールなどがこれまでに実用化されてきた．

このように，必ずしも新規性のない大規模データの利活用が，ビッグデータというキーワードで取り上げられているのは，従来の容量という観点に加えて，種類，頻度という観点においても変化があったからだと考えられる．

例えば，IoT(Internet of Things) では，生産現場などの機器や商品につけられたセンサによる稼働状況や流通経路といったデータが時々刻々と情報システムに送信され，これをリアルタイムで解析することで最適な生産管理や供給制御が行えると期待されている．このような M2M(Machine to Machine) で高頻度にやりとりされる情報の量は，人間が介在してやり取りされる情報量を超えると言われている．

一方で，このような数値化あるいは定形化された構造化データだけではなく，インターネット上で流通する大量のマルチメディアデータや，日々の業務で発生する大量の音声データ，動画データが蓄積される企業内情報システム等，非構造化データの割合が増している．これらのマルチメディアデータを有効に活用するためには，効率的な検索手段が不可欠である．人の手により付与されたタグ情報を元に目的のデータを検索することは現実的ではなく，マルチメディアデータを利活用するための認識，検索技術が求められている．

日立ではこれまで，IT プラットフォームの観点から，知識化サービス基盤 KaaS の提案 [植田 10] や，大量・多種多様な非構造化データを扱う情報処理基盤 [児玉 11] に関して提案を行ってきた．

本講演では特に，データの種類という観点に着目し，多様なデータ，特に非構造化データを利活用するための技術としてメディア処理技術に着目し，これらの分野における研究の取り組みを述べる．

連絡先: 池田尚司，株式会社日立製作所中央研究所，東京都国分寺市東恋ヶ窪 1-280，hisashi.ikeda.vz@hitachi.com

2. 非構造化データ利活用ニーズ

近年，医療や金融，企業情報，政府機関，ビデオ監視分野など様々な分野において，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ビジネスに活用したいというニーズが高まっている．表 1 に分野毎の活用例をまとめた．

医療分野では ICT を段階的に導入してきたため，部門やアプリケーション毎にシステムが存在し，異なるベンダ間では相互接続性が低いという問題があった．近年では，DICOM 等データ形式の標準化が進みシステム間のデータ連携が進んでいるが，スキャンデータや医療画像，録音メモなど病院内に未だ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等で再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等の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 制度への対応が重要になっている．

上記以外にも映像監視，政府，金融といった分野で非構造化データの利活用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，様々なシステムに蓄積されたデータを集約管理し，再活用するための分野共通的な基盤が求められている．

3. メディア処理技術

日立ではこれまで，テキスト処理，音声認識，画像認識の研究を行ってきた．以下，画像処理と音声処理に関する最近の取り組みを紹介する．

3.1 事例ベースオブジェクト検出 [渡邉 (裕) 11]

画像中から人の顔や車などのオブジェクト領域を特定する技術はオブジェクト検出技術と呼ばれ，画像認識分野においては古典的な課題である．特定のカテゴリに対するオブジェクト検出，例えば，顔検出機能は，デジタルカメラ等の撮像機器の組み込み機能として搭載されている．メーカ提供の技術に関しては，内部技術の詳細は不明であるが，特定のオブジェクトに特化するアプローチと，汎用的なオブジェクト検出を用いるアプローチがあり得る．後者は，適切な学習データさえ用意で

表 1: 非構造化データの利活用ニーズ例

分野	利活用ニーズ例		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 [渡邉 (龍) 11]	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 [視線検知技術]	監視映像
政府	政府保有データの利活用	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 [Data.gov]	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 [Facecount]	ユーザー生成コンテンツ

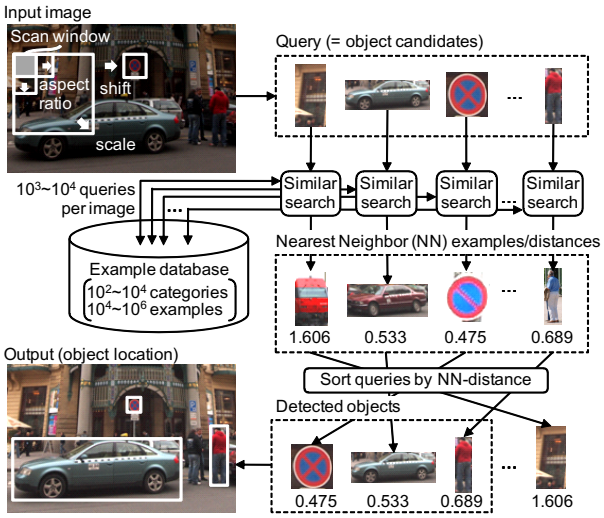

図 1: 類似画像検索を用いたオブジェクト検出

きれば，あらゆるオブジェクトの検出に応用可能である利点を持つ．

この問題に対し日立では，汎用的なオブジェクト検出を目的として，類似画像検索技術を応用した事例ベースのオブジェクト検出手法を提案した [渡邉 (裕) 11]．本手法は，入力画像の部分領域と検出対象の事例画像とを，類似画像検索用の特徴量ベースで照合することで，オブジェクト領域を検出する．照合処理は，多数の事例画像に対する最近傍探索によって行われるため，多様な事例を登録しておけば，アピアランスの変動の大きいカテゴリのオブジェクトも検出可能である．また，事例の登録のみで新しいカテゴリのオブジェクトを検出できるため，少量多種の一般的なオブジェクトに柔軟に適用可能になる．一方で，最近傍探索を用いた判別は計算コストがかかる処理である．特に，多数の候補領域を判別する必要のあるオブジェクト検出においては，実用面で大きな課題になる．これに対し，提案手法では，類似画像検索用のエッジパターン特徴量の特性を活かした粗密探索法と，高速類似ベクトル検索手法を用いることで，実用的な処理速度を達成した．

類似画像検索を用いたオブジェクト検出手法の概要を図 1 に示す．提案手法では，前処理として，検出対象のオブジェクトの事例を用意し，各事例から抽出した画像特徴量を事例データベース (Example database) に登録する．画像特徴量は，通常，高次のベクトルデータで与えられる．報告者らは，Web 画像を対象とした大規模類似画像検索システムを構築しており [廣池 11]，ここで開発したエッジパターン特徴量とデータベースシステムを利用した．事例は複数登録することが可能であり，事例同士は同一のカテゴリである必要はなく，たとえば，「車両」と「人の顔」のように，まったく異なるカテゴリを登録することができる．

検出処理においては，スキャンウィンドウを平行移動・拡大縮小させることにより選択された部分領域から，事例の特徴量と同一手法で特徴量を抽出する．次に，抽出した特徴量を用いて類似画像検索を行う．類似画像検索は，特徴量ベクトルの距離（特徴量距離）が小さい順に，データベースの要素を並び替えて出力する処理である．ここでは，報告者らの研究グループで開発した，特徴量ベクトルのクラスタリングに基づく高速類似ベクトル検索手法 [Matsubara 09] を用いた．本手法は，予め類似するサンプルをクラスタとしてまとめて保存しておき，検索時には，類似クラスタ探索とクラスタ内探索の 2 段階探索を行うことで，高速検索を実現する方法である．

この結果，特徴量がもっとも「近い」事例（最近傍事例）とその特徴量距離が得られる．最近傍事例との特徴量距離が閾値以下の場合に，部分領域がオブジェクトであると判定し，その領域情報（位置とサイズ）を出力する．

本手法を，CMU+MIT の顔データの検出に適用した場合，100 枚程度の少量の事例を使ったときでも約 80%の高い検出率を実現できることを示した．また，エッジパターン特徴量の特性を活かした粗密探索手法や，クラスタリングによる高速類似ベクトル検索手法を用いることで，10000 事例を登録した場合の検出処理を単純な実装と比べて約 30 倍高速化し，実用的な時間（平均 0.55 秒）で動作可能であることを確認した．

3.2 Web 画像データベースを用いた画像アノテーション [渡邉 12]

制約のない実世界の画像中の物体やシーンを計算機に認識させ，画像中の位置や物体名称などを一般的な表現で記述させる技術は一般物体認識と呼ばれ，画像認識の研究において最も困難な課題の一つとされている．一般物体認識の要素課題である画像アノテーションは，画像が表す内容に対応するメタデータを自動的に付与する技術である．近年では，インターネットの急速な発展を背景として，Web 上のデータを用いた画像認識の研究が発展しており，ノイズを含む低品質なデータを大量に集め，それらを直接的に使用することで学習なしの画像認識

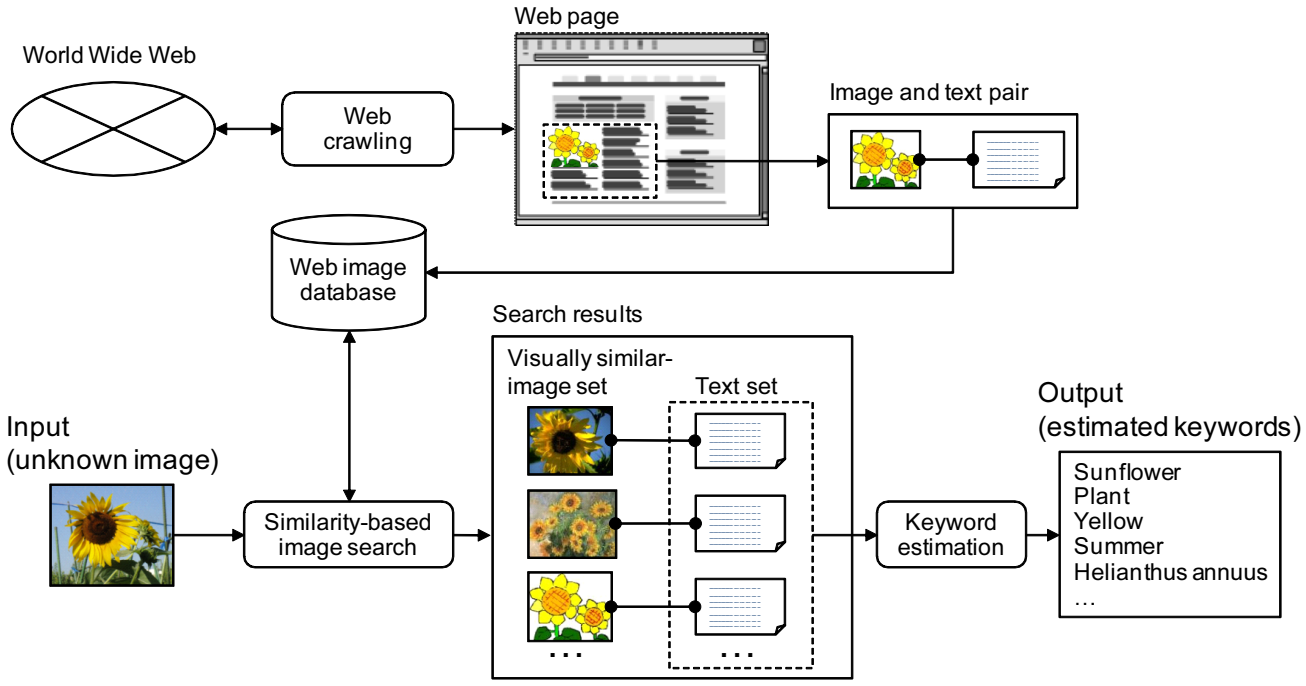

図 2: Web 画像データベースを用いた画像アノテーションシステム

を行う，事例ベースの手法が提案されている．

この問題に対し日立では，大規模 Web 画像データベースと類似画像検索技術を用いた画像アノテーションシステムを提案した [渡邉 12]．本システムは，与えられた画像をクエリとして類似画像検索を行い，検索結果の画像に付随するテキスト中の単語を確率的指標により評価することで，特別な事前学習なしに画像を意味付けるキーワードを推定可能である．

Web 画像データベースを用いた画像アノテーションシステムを図 2 に示す．本システムにおける処理は，クローリングによってデータベースを構築する前処理と，データベースを用いて画像認識する処理に分けられる．

前処理においては，Web クローリングによって自動的に取得した Web ページから，画像とその周辺テキストを抽出し，それらを関連付けて画像データベースに保存しておく．このようにして機械的に抽出されたテキストは，必ずしも画像を説明するものではないが，関連する単語が含まれる可能性は高い．

認識処理においては，まず，ユーザから入力された未知の画像をクエリとして類似画像検索を行う．類似画像検索は，画像そのものが持つ色や形状などの特徴による検索であり，この結果，入力画像と「見た目」の類似した画像が得られる．また，画像データベースには，画像とテキストが関連付けて保存されているため，検索結果からテキストの集合が得られる．

次に，得られたテキスト集合を 1 つの文書とみなし，この文書を特徴付ける重要語を抽出する．重要語の抽出では，文書に含まれるすべての単語に対して，重み付け指標によるスコアでソーティングし，その上位またはスコアが閾値以上の単語を出力する．

以上の処理により，未知の画像に対して自動的にタグを付与し，データ解析や検索に利用したり，ユーザが詳細なタグ付作業を行う際の補助ツールとして利用したりできる．本手法は，事前に認識対象を定義する必要がないため，膨大な概念が認識対象となる一般物体認識における学習コストの問題を解消することができる．また，継続したクローリングにより，時代の流れと共に生まれる新たな概念の画像に自動的に対応することができる．

Web 画像検索サービス「GazoPa」[廣池 11] で収集した約 1 億件の Web 画像データベースを用いて，画像アノテーションシステムを構築した．本システムは，一般的な PC サーバ（CPU 2.40 GHz，メモリ 4GB）を 13 台使用しており，各サーバに約 400 万件の画像データを管理する DB サーバプロセスが 2 つずつ存在する．それらに対して並列に検索処理を行うことで大規模類似画像検索を実現した．

本システムを用いて 5 カテゴリ 30 概念の画像に対するアノテーションを行った結果，10 位内正解率がカテゴリ平均で 43〜75%，全概念の平均で約 59.1%であった．また，処理速度は，データベース構築時の画像とテキストの解析に 1 画像あたり 90ms，画像アノテーション処理に 643ms を要した．

3.3 音声ドキュメントからの検索語検出 [神田 12]

日々の業務で発生する大量の音声録音データの中から所望の音声データを検索する技術は，テキスト検索との対比から，音声ドキュメント検索と呼ばれている．その中でも鍵となる技術が，音声データ中から所望のキーワードが含まれている音声区間を検出する「検索語検出」技術である．検索語検出の最も単純な実現方法は，大語彙連続音声認識技術を用いて音声をテキストに変換した後，テキストに基づくマッチングを行うというものである．しかしこの方法では音声認識の語彙に含まれない単語はテキストとして書き起こされえず，従ってそのような単語を検出することもできない（未知検索語問題）．新語や固有名詞などは検索語として重要であるもののその出現頻度の低さから未知語となりやすいため，未知検索語問題は検索語検出技術における主要課題となっている．これまで，単語よりも短い「サブワード」と呼ばれる単位で検索を行うことにより任意検索語の検出を可能とした技術が開発されていたが，処理速度が遅く，検索精度も低いという問題があった．

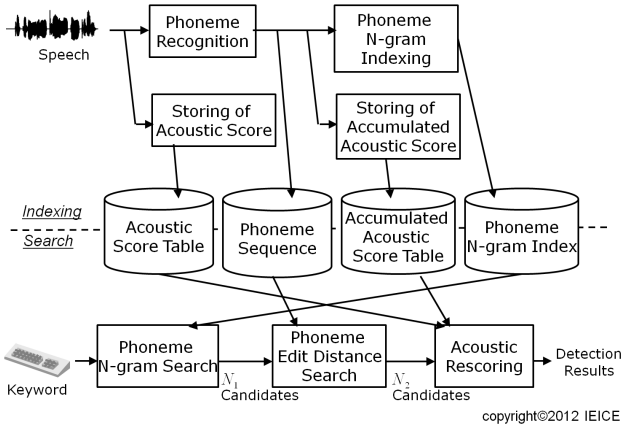

図 3: 音声中の検索語検出システムの構成

この問題に対し日立では，処理速度と検索精度の異なる 3 種類の手法を縦列接続することによる，高速かつ高精度な任意検索語の検出法を提案した [神田 12]．図 3 に，音声中の検索語検出システムの構成を示す．システムは，検索対象の音声データを前処理して音声用のインデックス (Index) を生成するインデキシング部 (Indexing) と，ユーザが指定した検索語の発話箇所を高速に検出する検索部 (Search) に分かれる．検索部は音素 N-gram 探索法 (Phoneme N-gram Search)，編集距離に基づく音素マッチング (Phoneme Edit Distance Search)，音響リスコアリング法 (Acoustic Rescoring) という 3 種類のモジュールを用いる．これらはいずれもサブワードに基づく任意検索語の検出システムである．この中で，音素 N-gram 探索法が最も高速に動作する反面，最も低精度である．音響リスコアリング法は最も高精度である反面，最も低速である．そして編集距離に基づく音素マッチングは，上記 2 つの中間の性能を持つ．開発したシステムでは，上記の検索モジュールを縦列接続することで，段階的に精度を向上させながら検索結果の候補点を絞り込み，検索精度の高さと検索の高速性を両立させた．

提案した手法は，604 時間の音声データ中で平均 5.7 回しか出現しない単語を，約 1.4 秒で F 値 67.8%の精度で検出可能であることを確認した．また，提案法は 0.11xRT という高速なインデキシング処理も特長として持つ．提案法は，検索精度・検索速度・任意検索語の検索・インデキシング速度といった要素を考慮する必要のある大規模な音声データベースの検索において，特に有効であると考えられる．

4. まとめ

ビッグデータという言葉によって期待される情報システム，あるいはサービスを実現するためには，収集，蓄積される大量かつ多様な情報の再活用を実現することが必要である．本稿では，そのための要素技術として，画像データからのオブジェクト検出技術，画像アノテーション技術，検索語検出技術を述べた．画像，音声，テキストといった各メディア情報からの情報抽出技術は，非構造化データの収集時の他，再活用時に文脈に応じて適用されると想定されるため，情報抽出の精度に加えて，処理速度，スケーラビリティが課題となる．今後はこれらの課題に取組むとともに，適用対象のメディア情報を拡大し，大量の非構造化データの再活用を可能にするプラットフォームの実現を目指す．

参考文献

[植田 10] 植田他: 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, Vol.92, No.5, pp.36-39, 2010.

[児玉 11] 児玉他: 大量・多種多様な非構造化データを扱う情報処理基盤, 日立評論, Vol.93, No.7, pp.56-59, 2011.

[渡邉 (龍) 11] 渡邉他: 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，Vol.93，No.3，pp.292-297, 2011.

[視線検知技術] 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/

[Data.gov] Data.gov: http://www.data.gov

[Facecount] Facecount：New study finds link between social media popularity and stock prices, http://www.facecount.com/news/new-study-finds-link-between-socialmedia-popularity-and-stock-prices-242652

[Matsubara 09] Matsubara, et.al: High-speed Similarity-based image retrieval with Data-alignment optimization using Self-organization algorithm, ISM2009.

[廣池 11] 廣池他: 類似画像検索のこれまでとこれから〜 Web 画像検索サービス「GazoPa」の経験を踏まえて〜, 信学技報, PRMU2011-91, pp.21-23, 2011.

[渡邉 (裕) 11] 渡邉他: 類似画像検索に基づく事例ベース一般オブジェクト検出, 信学技報, PRMU2011-142, pp.101-106, 2011.

[渡邉 12] 渡邉他: 大規模 Web 画像データベースを用いた画像アノテーションシステムの構築, 情報処理学会研究報告, Vol.2012-CVIM-181, pp.1-6, 2012.

[神田 12] 神田他: 多段リスコアリングに基づく大規模音声中の任意検索語検出, 信学論, vol.J95-D, no.4, pp.969-981, 2012.

講演者略歴

池田尚司

平 3 京大・工・情報卒．平 5 同大大学院修士課程了．同年 (株) 日立製作所入社．以来，同社中央研究所にて，手話通訳システム，文書解析，ヒューマンインターフェースの研究に従事．途中，平成 12 ドイツ人工知能センタ (DFKI) 客員研究員．平成 22 同社公共システム事業部勤務を経て．平成 23 より中央研究所にてメディア情報処理の研究マネージメントに従事．人工知能学会，情報処理学会，電子情報通信学会，日本認知科学会，IEEE，ACM 各会員．