

CoreExplorer の大規模化対応と今後の展開

Introduction of CoreExplorer Supporting large-scale Data and Future Business Development

(株)日立ソリューションズ東日本(以下、HSEと記す)は、テキストマイニングシステムであるCoreExplorer 5.0の提供を2018年6月に開始した。本バージョンは、CoreExplorerのアーキテクチャを刷新し、オープンソース(以下、OSSと記す)のElasticsearchと連携することにより、数千万規模の大規模データを扱えるようになった。また、ユーザインタフェースも刷新し、ダッシュボード等の表現を可能とし、今までの分析に特化した画面構成から、より業務に活用しやすい形になっている。本稿では、CoreExplorer 5.0の紹介と、今後の事業展開の方向性について述べる。

宮林 靖典	Miyahayashi Yasunori
宮内 秀彰	Miyauchi Hideaki
塚原 朋哉	Tsukahara Tomoya
高橋 毅	Takahashi Tsuyoshi

1. はじめに

1.1 CoreExplorerの変遷

HSEは、2003年よりテキストマイニングシステムであるCoreExplorerを提供し、お客様が社内に蓄積している様々なテキストデータを活用するためのソリューションを展開している^{1,2,3)}。

1.2 ビジネスを取り巻く状況

1.2.1 データの大規模化

近年、メモリの大容量化、SSDなどのハードウェアの登場、安価で高速なクラウドサービスなど、インフラ環境の充実に伴い、大量のデータを扱うことが可能となり、より緻密な情報から迅速に分析／業務活用することが求められるようになってきた。

例えば、コールセンタでは音声データをテキスト化し、その後の履歴／分析に使いたいという引き合いが増えている。音声テキストは、オペレータがCRMに應對の履歴を入力する際に省かれる情報が含まれており、クレーンにつながるワードや、商品開発に有用なワードが潜んでいる可能性がある。そのため、音声テキストは分析の入力情報として有望視されているが、オペレータが登録した履歴データと比較して、データ量が膨大になる。

1.2.2 分析ニーズの多様化

近年、機械学習のライブラリ/フレームワークがOSSで多数公開され、企業での評価や導入が容易になってきた。例えば、自然言語処理の分野においても、単語の共起関係からニューラルネットワークによって単語の

特徴量を学習するword2vecの技術⁴⁾を用いることで、従来よりも高い精度のクラス分類や要約などを実現することが期待されている。これらの新技術を組み込んだ分析を行いたいという市場のニーズに応えていくために、CoreExplorerに分析機能を随時追加していく必要がある。

そこで、数千万件規模のデータ処理と分析機能追加の容易性を考慮したアーキテクチャに見直したCoreExplorer 5.0を2018年6月にリリースした。

2. CoreExplorerの課題と対応

2.1 現状のCoreExplorerの課題

前章で述べた大規模化、分析ニーズの多様化にCoreExplorerが対応するにあたってCoreExplorerでは以下の課題があった。

(1)更新処理時間

データベースの更新時に分析で使うデータを事前に作成しておくことで分析の高速化を実現しているが、データベース中の全件を対象とした処理が必要であり、文書数増加に伴い更新時間が増大してしまう。そのため、100万件オーダーのデータを夜間や週末にバッチ的に更新する運用となり、バッチ処理時間内に収めるために扱うデータ量を増やすことが難しい。また更新処理時間がかかるため、リアルタイムで蓄積されるデータの取り込みや分析への適用が難しい。

(2)分析ニーズ反映

CoreExplorerでは特徴語の検索・集計で分析を実施

している。ここでの特徴語とは自然文中の単語（名詞や係り受け情報など）や文書の属性情報（顧客名や受付日など）である。自然文の長さにより特徴語の数も変わり、RDBなどの表形式のデータ構造ではデータが扱いにくい。文書ごとに不定個の特徴語が扱える独自のデータ保持形式を採用している。このデータ形式を前提とした分析のロジックをデータベースの検索エンジンに組み込んでいる。これにより分析の高速化を実現しているが、新たな分析のロジックをCoreExplorerに組み込むためにはデータ形式を意識した検索エンジンへの処理の作りこみが都度発生するため、市場ニーズに合わせた分析の追加が難しい。

2.2 CoreExplorerのアーキテクチャ見直し

2.1の課題の解決のため CoreExplorerのアーキテクチャを見直した。

- (1)更新処理時間短縮のための高速データベース適用
- (2)分析ニーズ反映のための検索と分析ロジックの分離

2.2.1 更新時間短縮のための高速データベース適用

2.1の(1)の課題解決のため、バッチ的に全件を対象とした更新処理ではなく新規文書のみ追加できる差分更新機能を検討した。開発期間の短縮と製品維持コスト低減のため OSSのデータベース適用を検討した。検討の際には以下の項目を実現できることとした。

- (1)文書ごとに不定個の特徴語を持つデータが扱える
- (2)更新処理・検索処理が高速
- (3)スケールアウトが容易

上記の条件を満たすデータベースとして、OSSのデータベースであるElasticsearchを採用した。これにより、差分更新が可能となり今まで数時間かかっていた更新処理が数分で終了するようになった。

2.2.2 分析ニーズ反映のための検索と分析ロジックの分離

2.1の(2)の課題解決のため、データベースの検索エンジンの役割を検索と分析で分離した。分析処理を検索条件の組み合わせや複数回の検索リクエストで実現することにより、分析ニーズの変化を検索条件・リクエストの変更で対応可能とする。これにより、ニーズに合わせて分析を追加することが容易となる。

2.3 システム構成

アーキテクチャ見直し結果を元に、システム構成を検討した（図1）。

クライアント画面には、Elasticsearchの可視化ツールである Kibanaを利用し、可視化・分析UIを Kibanaのプラグインとして開発している。プラグイン開発においては Kibanaのフレームワークに従い、通信機能・画面初期化機能等を利用している。

プラグインでは分析ロジックの実装として分析データ生成も行っている。Kibanaクライアントから送信された分析クエリを解釈し、複数の検索クエリとしてデータを検索・集約し分析データを生成しクライアントに返している。これらプラグインの修正・変更によりニーズに合わせた分析ロジックの実装を可能としている。

Elasticsearchに投入するデータは、分析対象文書である文書情報と、文書の特徴語をひとまとめでしたJSONデータである。文書情報はCSVの1レコードのようにカラムに分けられた情報であり、特徴語は自然文カラムから抽出した名詞・動詞や独自の日本語自然文の解析により生成した複合語・係り受け情報と、属性カラムから抽出した属性情報である。

検索処理はスケールアウト可能なElasticsearchの並列・分散処理を利用している。



図1 システム構成

3. CoreExplorer 5.0 の特徴

2018 年 6 月に本アーキテクチャに基づく CoreExplorer 5.0 をリリースしたことにより、以前のバージョンと比較し、以下の特徴を持つツールとなった。

3.1 大規模データの業務システムへ適用できるツール

大規模データが扱えるようになったことにより、今まで対応が難しかった1月あたり100万件の応対履歴データを蓄積するような大規模のコールセンタであっても、複数月を比較した分析などに対応が可能となった。また、それ以上のデータ量を持つ音声データにもスケールアウトすることにより、対応することができる。

3.2 容易に業務活用できる分析ツールへの変革

CoreExplorer 5.0では、Elasticsearchによる大規模データ対応だけでなく、新たな業務利用シーンを想定した機能をそろえている。

3.2.1 分析条件保存機能

分析担当者が作成した「分析条件」を呼び出すことにより、ワンクリックで同じ分析の切り口で結果を閲覧することが可能となった。これにより、非分析者でも分析者と同じ観点で結果を得ることができるため、容易に分析が可能となる。業務活用の例として、コールセンタのスーパーバイザの場合、日々報告のために集計している問合せ傾向の分析条件をテンプレートとして保存し、呼び出すことで日報作成などの定型的に行っている分析・報告業務の効率化が可能となる。

3.2.2 急上昇ワード抽出機能

テキストデータに含まれる大量の単語から、急激に増加した単語のみを抽出することで、データのトレンドを把握できる。例えば、製造業企画部門の場合、Tweet等のSNSのデータの急上昇ワードを調査することにより、実施したキャンペーンの効果や、製品に向けたクレームの予兆に気付くことに役立つ。

4. 今後の事業展開

4.1 CoreExplorerの事業拡大に向けた施策

CoreExplorerは企画部門や品質保証部門、コールセンタのスーパーバイザなど、限られた分析担当者に向けた商品開発を行ってきた。最近では分析用途の要望に加え、CoreExplorerの機能とword2vec等の自然言語処理や他製品を組み合わせ、ライン部門の業務効率化に直接繋がるソリューションが求められている。そこで、HSEは、コールセンタを対象としたソリューションと、研究報告等の知財データ可視化のソリューション展開を進めている。

4.2 今後の展開予定のソリューション

4.2.1 コールセンタ向けソリューション

HSEは、日立製作所の「音声データ利活用ソリューション⁵⁾」と連携し、コールセンタ向けソリューションを提案している。

弊社が考えるコールセンタソリューションの全体像を図2に示す。その中で、赤枠で囲まれた部分についてHSEで独自に研究、マイクロサービス化、製品化を順次進めている。各ソリューションの概要を以下に示す。

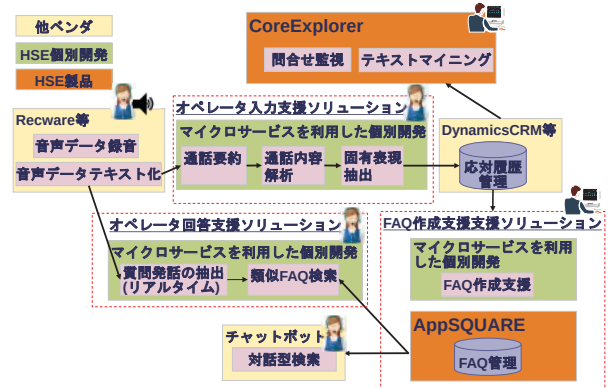


図2 コールセンタ向けソリューション

4.2.1.1 オペレータ入力支援

コールセンタでは、オペレータが対応履歴をCRMシステムに登録するアフターコールワーク（以下、ACWと記す）業務の効率化を求める声が多い。その課題に対して、音声テキストデータ要約を利用した、オペレータACW入力支援ソリューションの検討を進めている。オペレータは通話終了後、要約された対応内容を確認、修正するだけで対応内容の登録ができる。これにより、正確な対応内容を残しつつACW時間の短縮が図れる。

4.2.1.2 オペレータ回答支援

コールセンタでは、平均通話時間（以下、ATTと記す）の短縮が重要なKPIの一つとなっており、オペレータが素早く回答に辿り着くことは重要である。

この課題に対し、リアルタイム音声テキスト化のデータを利用し、お客様の質問を自動判定する技術と、自然文検索技術を連携することで、リアルタイムでのFAQ検索を実現し、回答候補をオペレータに提示することが可能となりATT短縮に繋がると考えている。

4.2.1.3 FAQ作成支援

前項のオペレータ回答支援ソリューションの効果をより上げるためには、FAQを充実させることが重要である。FAQ検索ヒット率を向上させるためには、問合せの多い質問に対する回答を準備することが重要で、どのような質問が多いかを把握するためには、過去の対応履歴が活用できる。過去の対応履歴から、類似の問合せをグループ化し、グループを最も特徴付ける質問の一つ抽出することで、FAQの候補とすることが可能となり、FAQの作成支援につながると考えている。

4.2.2 技術情報管理可視化のソリューション

企業は社内に技術文書や保守情報といった知財データを管理しているが、関連情報の検索に留まっていることも多く、更なる活用を期待されている。HSEではCoreExplorerとAppSQUAREを組合せ、データの蓄積／管理から分析結果の可視化まで一貫したソリューションの提供が可能である。

例として、AppSQUAREで管理した研究報告書に書かれた文章の解析結果と、書誌情報などフォームに登録される属性情報を組み合わせて、「研究者と技術」や「研究者同士」の繋がりをCoreExplorerで可視化することにより、類似研究をしている有識者を検索するなど、保有している技術情報を多面的に活用することが可能となる。図3に可視化のイメージを示す。

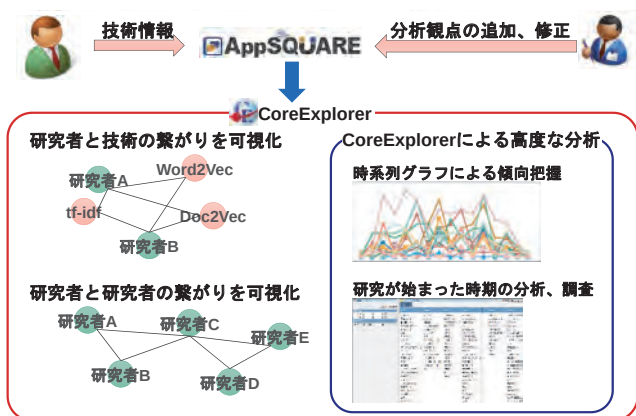


図3 技術情報の可視化イメージ

4.3 テキスト分析マイクロサービスの展開

現在、上記の業務効率化に向けたソリューションを提案・運用が出来るように、固有表現抽出や要約・自動分類などの自然言語処理の各機能を再利用しやすくする必要がある。そのため、HSEでは各機能のマイクロサービスの開発を進め、テキスト分析基盤として整備中である。本マイクロサービスは、RESTfulなインタフェースで構成されており、Webサービスとして各種テキスト分析機能の利用が可能である。

2018年10月時点では、テキスト要約の機能を実装予定としており、本マイクロサービスを利用して、PoCや顧客先への基盤構築を予定している。

5. おわりに

昨今、音声のテキスト化をはじめとするデータの大容量化が益々進み、分析対象の増加への対応は命題である。また、経営判断の迅速化に伴い、分析結果を有効活用す

るために、分析部門だけではなく、他部門や経営層とスピーディーに情報共有できる機能が必要である。

それらに対応するため、スピードやスケーラビリティに優れているOSSの検索エンジン「Elasticsearch」と連携できるCoreExplorer 5.0をリリースし、大規模化データへの対応と、業務利用に繋がる機能向上を行った。今後は機械学習や深層学習などの自然言語処理や、他製品との連携を用いて更に業務を見据えたソリューション開発を進め、テキスト情報の活用ニーズに応えていく。

参考文献

- 1) 宮林靖典他, HSE 技報第 20 号「CoreExplorer Ver.3.0 のリリースと今後の事業展開」
- 2) 宮内秀彰他, 日立 TO 技報第 18 号「テキストマイニングシステム CoreExplorer の今後の事業展開」
- 3) 宮内秀彰他, HSE 技報第 22 号「CoreExplorer Web API オプションのリリースと今後の事業展開」
- 4) Tomas Mikolov他, “Distributed representations of words and phrases and their compositionality,” Advances in neural information processing systems, pp. 3111-3119 (2013)
- 5) 日立製作所, ニュースリリース「音声データ利活用ソリューション」
<http://www.hitachi.co.jp/New/cnews/month/2016/03/0323.html> (2016 年 9 月閲覧)



宮林 靖典 2003 年入社
Viz ソリューション部
CoreExplorer の提案・事業推進
yasunori.miyahayashi.zu
@hitachi-solutions.com



宮内 秀彰 2008 年入社
Viz ソリューション部
CoreExplorer の提案・構築
hideaki.miyauchi.zd
@hitachi-solutions.com



塚原 朋哉 1997 年入社
第一パッケージ開発部
CoreExplorer の製品開発
tomoya.tsukahara.uh
@hitachi-solutions.com



高橋 毅 2013 年入社
Viz ソリューション部
CoreExplorer の提案・構築
tsuyoshi.takahashi.rd
@hitachi-solutions.com