報道各位

2019年7月2日
株式会社インプレスR&D
https://nextpublishing.jp/

Elasticsearch 中級者のための実務活用事例集！

『Elasticsearch NEXT STEP』発行

技術の泉シリーズ、7月の新刊

インプレスグループで電子出版事業を手がける株式会社インプレス R&D は、『Elasticsearch NEXT STEP』(監修:アクロクエストテクノロジー株式会社、著者:樋口 慎、山本 大輝、佐々木 峻、束野 仁政)を発行いたします。

最新の知見を発信する『技術の泉シリーズ』は、「技術書典」をはじめとした各種即売会や、勉強会・LT 会などで頒布された技術同人誌を底本とした商業書籍を刊行し、技術同人誌の普及と発展に貢献することを目指します。

『Elasticsearch NEXT STEP』
https://nextpublishing.jp/isbn/9784844398981

監修:アクロクエストテクノロジー株式会社
著者:樋口 慎、山本 大輝、佐々木 峻、束野 仁政
小売希望価格:電子書籍版 1600 円(税別)／印刷書籍版 1800 円(税別)
電子書籍版フォーマット:EPUB3／Kindle Format8
印刷書籍版仕様:B5 判／カラー／本文 96 ページ
ISBN:978-4- 8443-9898-1
発行:インプレス R&D

＜＜発行主旨・内容紹介＞＞

「Elasticsearch NEXT STEP」は、入門書の次のステップ(NEXT STEP)に踏み出すための実務事例集です。

Elasticsearch は、ダウンロードやインストールが非常に簡単な製品で数コマンド実行すれば、簡単に操作することが可能ですが、環境を考慮した設定でデータ分析などへの活用を考えると、非常に難易度が上がり、次の壁を踏み越えるのが大変です。

そこで実務での経験を事例集として掲載しました。読後には、より Elasticsearch が活用できるようになっています。

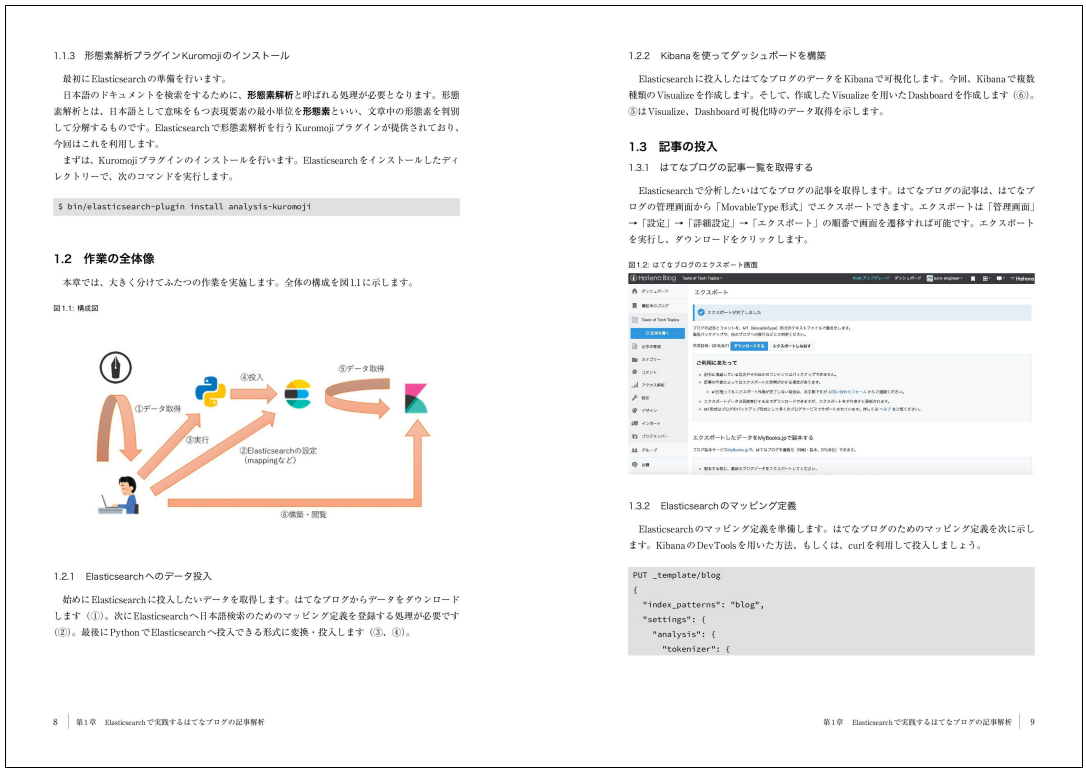
〈本書の対象読者〉

・Elasticsearch を多少触ったことがある方

・実践的な次の一歩を踏み出そうとしている方

（本書は、次世代出版メソッド「NextPublishing」を使用し、出版されています。）

ブログの記事解析を通じて Elasticsearch の各機能を使いこなします



1.1.3　形態素解析プラグインKuromojiのインストール

最初にElasticsearchの準備を行います。

日本語のドキュメントを検索をするために、**形態素解析**と呼ばれる処理が必要となります。形態素解析とは、日本語として意味をもつ表現要素の最小単位を**形態素**といい、文章中の形態素を判別して分解するものです。Elasticsearchで形態素解析を行うKuromojiプラグインが提供されており、今回はこれを利用します。

まずは、Kuromojiプラグインのインストールを行います。Elasticsearchをインストールしたディレクトリーで、次のコマンドを実行します。

```
$ bin/elasticsearch-plugin install analysis-kuromoji
```

1.2　作業の全体像

本章では、大きく分けてふたつの作業を実施します。全体の構成を図1.1に示します。

図1.1: 構成図

1.2.1　Elasticsearchへのデータ投入

始めにElasticsearchに投入したいデータを取得します。はてなブログからデータをダウンロードします（①）。次にElasticsearchへ日本語検索のためのマッピング定義を登録する処理が必要です（②）。最後にPythonでElasticsearchへ投入できる形式に変換・投入します（③、④）。

1.2.2　Kibanaを使ってダッシュボードを構築

Elasticsearchに投入したはてなブログのデータをKibanaで可視化します。今回、Kibanaで複数種類のVisualizeを作成します。そして、作成したVisualizeを用いたDashboardを作成します（⑥）。⑤はVisualize、Dashboard可視化時のデータ取得を示します。

1.3　記事の投入

1.3.1　はてなブログの記事一覧を取得する

Elasticsearchで分析したいはてなブログの記事を取得します。はてなブログの記事は、はてなブログの管理画面から「MovableType形式」でエクスポートできます。エクスポートは「管理画面」→「設定」→「詳細設定」→「エクスポート」の順番で画面を遷移すれば可能です。エクスポートを実行し、ダウンロードをクリックします。

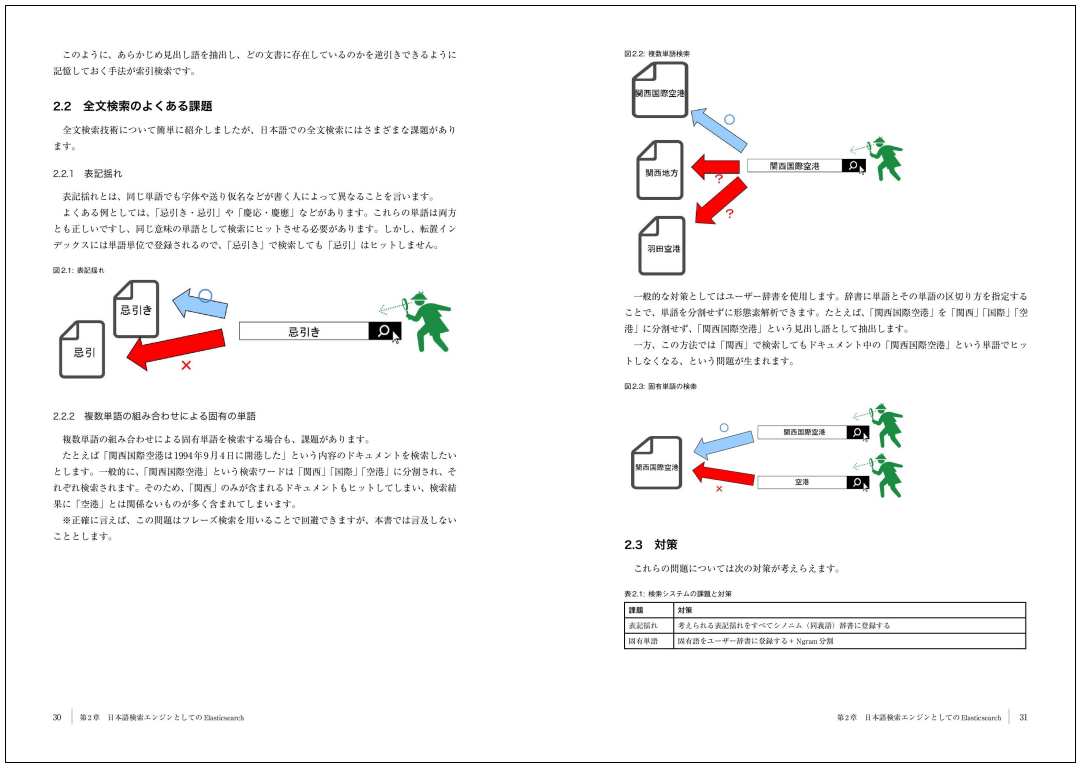
図1.2: はてなブログのエクスポート画面

1.3.2　Elasticsearchのマッピング定義

Elasticsearchのマッピング定義を準備します。はてなブログのためのマッピング定義を次に示します。KibanaのDevToolsを用いた方法、もしくは、curlを利用して投入しましょう。

```
PUT _template/blog
{
  "index_patterns": "blog",
  "settings": {
    "analysis": {
      "tokenizer": {
```


日本語検索エンジンとして Elasticsearch を使うための方法を紹介



このように、あらかじめ見出し語を抽出し、どの文書に存在しているのかを逆引きできるように記憶しておく手法が索引検索です。

2.2　全文検索のよくある課題

全文検索技術について簡単に紹介しましたが、日本語での全文検索にはさまざまな課題があります。

2.2.1　表記揺れ

表記揺れとは、同じ単語でも字体や送り仮名などが書く人によって異なることを言います。

よくある例としては、「忌引き・忌引」や「慶応・慶應」などがあります。これらの単語は両方とも正しいですし、同じ意味の単語として検索にヒットさせる必要があります。しかし、転置インデックスには単語単位で登録されるので、「忌引き」で検索しても「忌引」はヒットしません。

図2.1: 表記揺れ

2.2.2　複数単語の組み合わせによる固有の単語

複数単語の組み合わせによる固有単語を検索する場合も、課題があります。

たとえば「関西国際空港は1994年9月4日に開港した」という内容のドキュメントを検索したいとします。一般的に、「関西国際空港」という検索ワードは「関西」「国際」「空港」に分割され、それぞれ検索されます。そのため、「関西」のみが含まれるドキュメントもヒットしてしまい、検索結果に「空港」とは関係ないものが多く含まれてしまいます。

※正確に言えば、この問題はフレーズ検索を用いることで回避できますが、本書では言及しないこととします。

図2.2: 複数単語検索

一般的な対策としてはユーザー辞書を使用します。辞書に単語とその単語の区切り方を指定することで、単語を分割せずに形態素解析できます。たとえば、「関西国際空港」を「関西」「国際」「空港」に分割せず、「関西国際空港」という見出し語として抽出します。

一方、この方法では「関西」で検索してもドキュメント中の「関西国際空港」という単語でヒットしなくなる、という問題が生まれます。

図2.3: 固有単語の検索

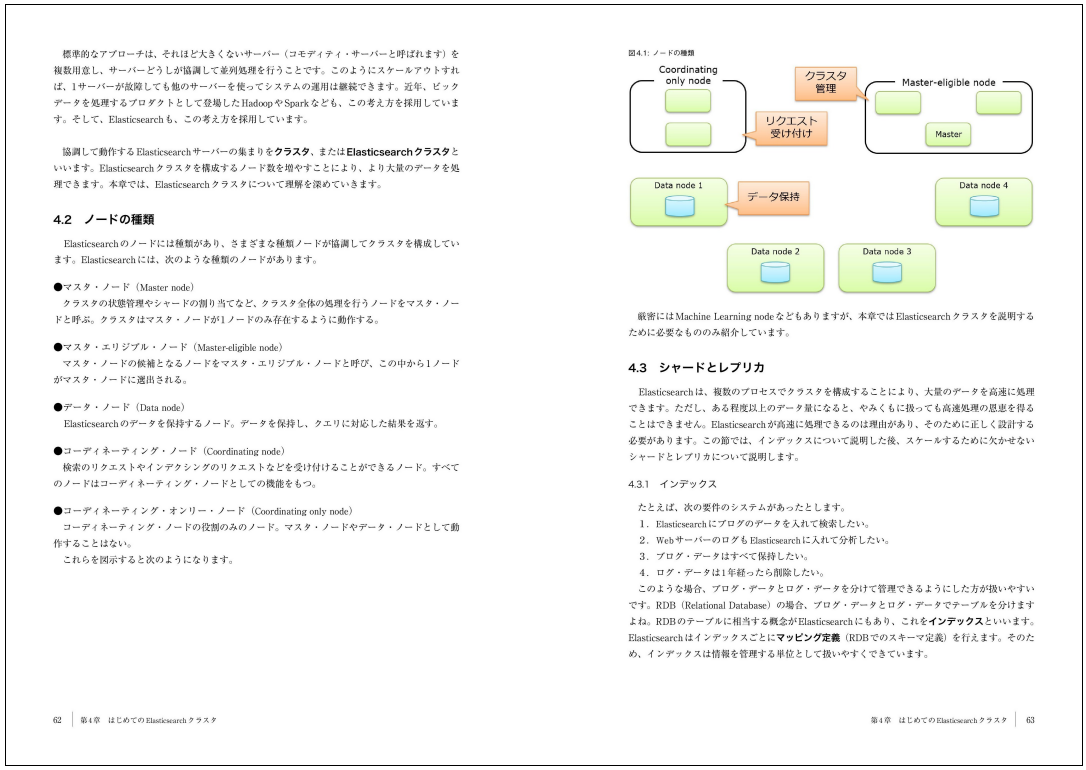
2.3　対策

これらの問題については次の対策が考えらえます。

表2.1: 検索システムの課題と対策

課題	対策
表記揺れ	考えられる表記揺れをすべてシノニム（同義語）辞書に登録する
固有単語	固有語をユーザー辞書に登録する＋Ngram分割

安全に Elasticsearch をクラスタ化して使うためのノウハウを紹介



標準的なアプローチは、それほど大きくないサーバー（コモディティ・サーバーと呼ばれます）を複数用意し、サーバーどうしが協調して並列処理を行うことです。このようにスケールアウトすれば、1サーバーが故障しても他のサーバーを使ってシステムの運用は継続できます。近年、ビックデータを処理するプロダクトとして登場したHadoopやSparkなども、この考え方を採用しています。そして、Elasticsearchも、この考え方を採用しています。

協調して動作するElasticsearchサーバーの集まりを**クラスタ**、または**Elasticsearchクラスタ**といいます。Elasticsearchクラスタを構成するノード数を増やすことにより、より大量のデータを処理できます。本章では、Elasticsearchクラスタについて理解を深めていきます。

4.2 ノードの種類

Elasticsearchのノードには種類があり、さまざまな種類ノードが協調してクラスタを構成しています。Elasticsearchには、次のような種類のノードがあります。

●マスタ・ノード（Master node）
クラスタの状態管理やシャードの割り当てなど、クラスタ全体の処理を行うノードをマスタ・ノードと呼ぶ。クラスタはマスタ・ノードが1ノードのみ存在するように動作する。

●マスタ・エリジブル・ノード（Master-eligible node）
マスタ・ノードの候補となるノードをマスタ・エリジブル・ノードと呼び、この中から1ノードがマスタ・ノードに選出される。

●データ・ノード（Data node）
Elasticsearchのデータを保持するノード。データを保持し、クエリに対応した結果を返す。

●コーディネーティング・ノード（Coordinating node）
検索のリクエストやインデクシングのリクエストなどを受け付けることができるノード。すべてのノードはコーディネーティング・ノードとしての機能をもつ。

●コーディネーティング・オンリー・ノード（Coordinating only node）
コーディネーティング・ノードの役割のみのノード。マスタ・ノードやデータ・ノードとして動作することはない。
これらを図示すると次のようになります。

図4.1: ノードの種類

厳密にはMachine Learning nodeなどもありますが、本章ではElasticsearchクラスタを説明するために必要なもののみ紹介しています。

4.3 シャードとレプリカ

Elasticsearchは、複数のプロセスでクラスタを構成することにより、大量のデータを高速に処理できます。ただし、ある程度以上のデータ量になると、やみくもに扱っても高速処理の恩恵を得ることはできません。Elasticsearchが高速に処理できるのは理由があり、そのために正しく設計する必要があります。この節では、インデックスについて説明した後、スケールするために欠かせないシャードとレプリカについて説明します。

4.3.1 インデックス

たとえば、次の要件のシステムがあったとします。
1．Elasticsearchにブログのデータを入れて検索したい。
2．Webサーバーのログも Elasticsearchに入れて分析したい。
3．ブログ・データはすべて保持したい。
4．ログ・データは1年経ったら削除したい。
このような場合、ブログ・データとログ・データを分けて管理できるようにした方が扱いやすいです。RDB（Relational Database）の場合、ブログ・データとログ・データでテーブルを分けますよね。RDBのテーブルに相当する概念がElasticsearchにもあり、これを**インデックス**といいます。Elasticsearchはインデックスごとに**マッピング定義**（RDBでのスキーマ定義）を行えます。そのため、インデックスは情報を管理する単位として扱いやすくできています。

＜＜目次＞＞

＜＜監修者紹介＞＞

Acroquest Technology 株式会社
1991年3月に創業。UNIXをいち早く採用した集中監視制御システムの開発を中心にミッションクリティカルな分野で事業を展開。Java 誕生直後から Java／オブジェクト指向を開発現場に導入し、分散システムを数多く開発した。
2016年4月に Elasticsearch 株式会社と OEM 契約を締結し、Elastic Stack をベースにしたデータ分析ソリューション「ENdoSnipe」を開発・販売している。
2018年7月に国内初の Advanced Reseller Partner となり、Elastic Stack の販売代理店として事業展開している。
2015年に、人を大切にする経営学会による「日本でいちばん大切にしたい会社大賞」の審査委員会特別賞を受賞。
2015年、2016年、2018年に、Great Place to Work が実施する「働きがいのある会社」ランキングの第1位（従業員数99名以下の部）に選出。

＜＜著者紹介＞＞

山本 大輝（やまもと ひろき）
第1章 Elasticsearch で実践するはてなブログの記事解析 執筆
データサイエンス、分析業務に従事、専門は画像処理。機械学習、Deep Learning を利用したソリューションの開発・提案を中心に行い、Elasticsearch による分析業務も並行して行っている。趣味は Kaggle で、日夜コンペティションに参加している。7198 チーム参加した過去最大（2018 年 9 月時点）のコンペティション"Home Credit Default Risk"で Kaggle 仲間と共に2位を獲得。現在、Kaggle Master。

佐々木 峻（ささき たかし）
第2章 日本語検索エンジンとしての Elasticsearch 執筆
自然言語処理、全文検索、IoT 分析関連のプロジェクトを中心に活動している。言語処理学会第 24 回年次大会ワークショップ「形態素解析の今とこれから」にて、「検索サービスに Sudachi を適用して運用コストを削減した話」というタイトルで発表した。

樋口 慎（ひぐち しん）
第3章 Elasticsearch SQL 執筆
データ分析・ElasticStack コンサルティング業務に携わる。Elasticsearch 社公認のテクニカルワークショップで講師などを務めているほか、世界でも数少ない Elastic Certified Engineer 資格を保有している。

束野 仁政（つかの さとゆき）
第4章 はじめての Elasticsearch クラスタ 執筆
分散システム、ビッグデータ、機械学習関連のプロジェクトを中心に従事している。ここ数年は、Elasticsearch を使ったシステムのコンサルティング・設計・開発を行っており、100TB の Elasticsearch クラスタの運用経験を持つ。日本

Java ユーザーグループ主催のセミナー「Elasticsearch 特集」や、Elastic Tokyo User Group 主催の「Elasticsearch 勉強会」での発表経験あり。趣味は数学、量子コンピュータ、スペイン語。社内では数学部のメンバーとして活動している。

＜＜販売ストア＞＞

電子書籍:

Amazon Kindle ストア、楽天 kobo イーブックストア、Apple Books、紀伊國屋書店 Kinoppy、Google Play Store、honto 電子書籍ストア、Sony Reader Store、BookLive!、BOOK☆WALKER

印刷書籍:

Amazon.co.jp、三省堂書店オンデマンド、honto ネットストア、楽天ブックス

※　各ストアでの販売は準備が整いしだい開始されます。

※　全国の一般書店からもご注文いただけます。

【インプレス R&D】 https://nextpublishing.jp/

株式会社インプレスR&D（本社：東京都千代田区、代表取締役社長：井芹昌信）は、デジタルファーストの次世代型電子出版プラットフォーム「NextPublishing」を運営する企業です。また自らも、NextPublishing を使った「インターネット白書」の出版など IT 関連メディア事業を展開しています。

※NextPublishing は、インプレス R&D が開発した電子出版プラットフォーム（またはメソッド）の名称です。電子書籍と印刷書籍の同時制作、プリント・オンデマンド（POD）による品切れ解消などの伝統的出版の課題を解決しています。これにより、伝統的出版では経済的に困難な多品種少部数の出版を可能にし、優秀な個人や組織が持つ多様な知の流通を目指しています。

【インプレスグループ】 https://www.impressholdings.com/

株式会社インプレスホールディングス（本社：東京都千代田区、代表取締役：唐島夏生、証券コード：東証 1 部 9479）を持株会社とするメディアグループ。「IT」「音楽」「デザイン」「山岳・自然」「旅・鉄道」「学術・理工学」を主要テーマに専門性の高いメデイア&サービスおよびソリューション事業を展開しています。さらに、コンテンツビジネスのプラットフォーム開発・運営も手がけています。

【お問い合わせ先】

株式会社インプレス R&D NextPublishing センター
TEL 03-6837-4820
電子メール: np-info@impress.co.jp