検索技術による企業内外データの仮想統合

Virtual Integration of Data Inside and Outside an Enterprise using Search Engine

石　井　　愛

要　約　検索技術により，企業内外に散在するデータを仮想的に統合する「情報アクセス基盤」と，その上でデータを分析・活用するアプリケーション（Search Based Application：SBA）が注目を集めはじめている．SBA は意思決定や，マーケティング，リコールの予兆の検知等，複雑な要求に応えるアプリケーションである．

本論文では，オープンソースの検索エンジンである Apache Solr をはじめとした複数の OSS を活用して情報アクセス基盤を構築する意義と，構築にあたっての留意点を，SBA の構築事例を踏まえて報告する．

Abstract　An information access platform using search engine virtually integrates multiple disparate data inside and outside an enterprise. The applications based on such platform are called Search Based Applications (SBA), and have started gaining attention in the business world. SBAs are built as the tools for analysis and decision-making.

This paper reports the significance of the information access platform which uses some OSS(s) including Apache Solr, the OSS search engine. It also shows a case example of SBA construction to show some important points to be known for the building of such information access platform.

1. は　じ　め　に

IT 環境の進化によって，企業内に蓄積されるデータ量は爆発的に増大し続けている．それらの膨大なデータは，各種業務システムやグループウェア，ファイルサーバー等，企業内の各部門単位で存在しており，その中から必要な情報を即座に探し出すことは非常に困難である．一方で，急速に拡大しているソーシャルメディアや口コミ情報等のデータの中から有益な情報を抽出し，より良い意思決定に利用したいというニーズが高まっている．そのような課題とニーズに対し，検索技術により，企業内外の構造化・非構造化データを仮想的に統合することで情報アクセス基盤を構築し，その上にデータを分析・活用するアプリケーション（Search Based Application：SBA）を構築することが，解決策の一つとなる．

本論文では，オープンソースの検索エンジンである Apache Solr をはじめとした複数の OSS を活用して情報アクセス基盤を構築する意義と，構築にあたっての留意点を，SBA の構築事例を踏まえて報告する．

2. 検索技術による企業内外に散在するデータの仮想統合

本章では，企業内外に散在するデータを対象とした検索技術による情報アクセス基盤の構築の意義と，その位置づけについて記述する．

2.1　インターネット検索とエンタープライズサーチ

インターネット検索は1994年にはじまったとされており，インターネット検索エンジンの最大手であるGoogleの創業は1998年である．その技術を企業内で利用するための検索製品は2000年頃から徐々に登場し，本格的に活用されはじめたのは2005年頃からである．この企業内外のWebサイトを含め，企業内の業務システムやファイルサーバー等に格納されている情報を横断的に検索するコンセプトを「エンタープライズサーチ」という．

インターネット検索は従来，インターネット上のWebページをはじめとした文書内の文章の検索に用いられてきたが，エンタープライズサーチではインターネット上の文書だけでなく業務システムのデータベース内の情報も対象になることが大きな特徴と言える．

2.2　情報アクセス基盤によるデータの仮想統合

検索技術は情報を見つけ出す手段であるとともに，企業内外に散在する構造化データおよび非構造化データを，仮想的に統合し処理する基盤技術として利用されている．このような技術を基に，意思決定や，マーケティング，リコールの予兆の検知等，複雑な要求に応えるアプリケーションを，Search Based Application（以下，SBAと記述する）と呼び，それを可能とする基盤を「情報アクセス基盤」と呼ぶ．

2.3　検索技術による情報アクセス基盤の位置づけ

検索技術による情報アクセス基盤は，従来のリレーショナルデータベース（以下，RDBと記述する）の役割の代替とはならず，共存する関係となる．RDBは企業の基幹データを扱うトランザクション処理を担い，検索技術ではRDBのスナップショットを定期的に取得しデータを効率よく参照するための機能を提供する．RDBでも全文検索は可能だが，パフォーマンスが問題となる場合がある．またRDBは集合であるため，値によるソート順での表示となるが，検索技術では，表記ゆれへの対策や「スコア」によるランキングが行われるため，よりユーザニーズに合致したデータを，合致した度合いで提示することができる．

データの統合と分析は，従来，データウェアハウス（DWH）が役割を担っており，高度な分析に使用される．検索技術による統計や分析は簡易的なものとなるが，比較的低コストでの

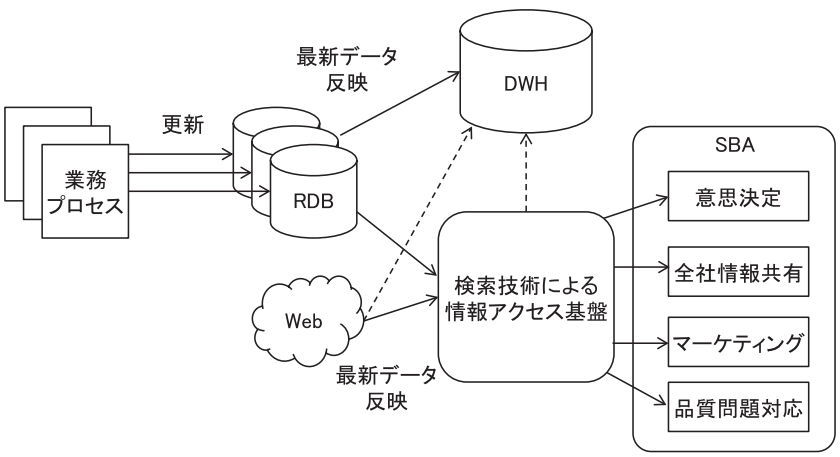

図1　情報アクセス基盤による企業内外データの仮想統合

導入が可能なため，データ活用のエントリーポイントとして考えることができる．

図 1 に，企業における検索技術を含めたシステム構成の例を示す．

3.　情報アクセス基盤を実現する検索技術の仕組み

本章ではオープンソースの全文検索エンジンサーバーである Apache Solr を例として，SBA の情報アクセス基盤となる検索技術の仕組みについて解説する．

3. 1　全文検索とは

全文検索とは，複数のドキュメント（文書ファイル・DB・Web ページなどの個々のテキストデータ）にまたがって，ドキュメントに含まれる全文を対象として検索するという意味である．全文検索の方式には大きく分けて順次検索（grep）方式とインデックス方式があるが，ほとんどの全文検索エンジンは高速な検索が可能となるインデックス方式を採用している．また，インデックスのデータ構造にはいくつか種類があるが，パフォーマンスの面で優れる「転置インデックス」[*1] が一般的に使用されている．

3. 2　全文検索の仕組み

全文検索は，事前準備としてクロール・インデクシングという処理が必要であり，検索実行時は，検索画面から入力された検索式をサーチャが解析し，事前作成していたインデックスを検索することにより，検索結果を利用者に返す（図 2）．

クローラにより，検索対象とするドキュメント（ファイル・DB・Web ページなど）を収集し（クロール），インデクサにより，収集したドキュメントを検索するための索引（インデックス）を作成する（インデクシング）．

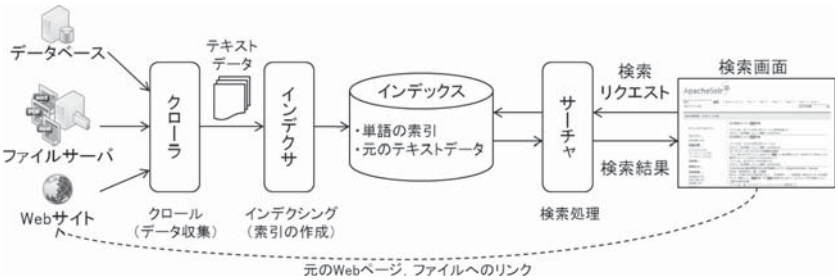

図 2　全文検索の仕組み

3. 3　OSS 全文検索エンジン Apache Solr

Apache Solr[*2]（アパッチ ソーラー．以下，Solr と記述する）はオープンソースの全文検索エンジンサーバー[*3] の一つである．Solr は，コミュニティーの活発さ（数千人の開発者），更新頻度（ほぼ毎日），ダウンロード数（1 日 6000 回～ 10000 回）から見て，2012 年現在，オープンソースでは最有力の検索エンジンと言える．Solr は必要十分な機能と高いスケーラビリティを持ち，Twitter 社等 Web 系企業を中心に超大規模事例が爆発的に増えており，国内でも適用事例が増加し続けている．また，動的な絞り込み検索（ファセット機能）等の検索機能やインデクシングにカスタマイズ処理を追加できる仕組みを持ち，商用製品と比べても機能的

に優位性があると言える．

Solr の最も大きな特徴はオープンソースソフトウェア（以下，OSS）であることである．一般的に OSS は，リリース管理，サポート等の課題があると言われているが，IDC，ガートナーによるとほぼ半数以上の企業で OSS が導入されており[5]，震災等の影響から更に OSS の採用率が加速するとの分析結果がある[6]．商用の検索製品はサーバー台数やデータ量によってライセンス費用が非常に高額になるものが多いため，特に大規模なデータ処理を必要とする案件では OSS の使用によるコストメリットが大きい．

3. 4　情報アクセス基盤を実現するために必要な機能

SBA の情報アクセス基盤を検索技術で実現するために必要な機能は，Solr を中心とした複数の OSS を活用して提供することができる．図 3 に，情報アクセス基盤における Solr の提供機能の範囲を示す．クローラや管理機能は他の OSS を組み合わせて構築する．グループウェア等のクロールには，商用のクローラや作り込みが必要となる場合がある．SBA の構築には，OSS の Web アプリケーションフレームワークが Solr 向けに用意されており，また Solr をコアとして利用可能な商品も存在するため，用途に合わせて構築方法を検討する．

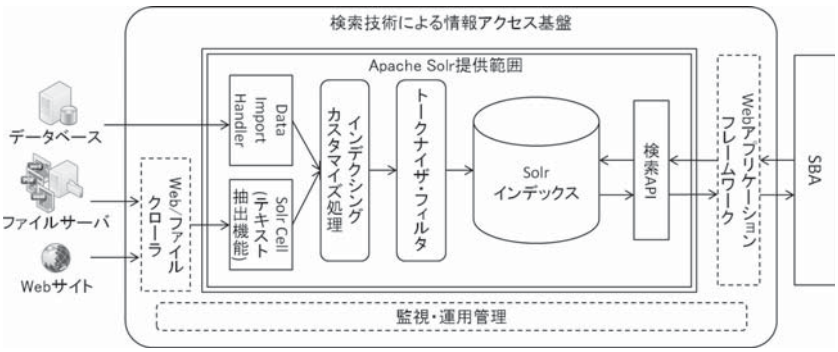

図 3　Solr の提供機能の範囲

3. 4. 1　データベースクロール

情報アクセス基盤として，データベース内のデータを収集して仮想的に統合するためには，データベースの複数のテーブルを関連付けた状態で取得し，インデクシングする機能が必要である．Solr では DataImportHandler（DIH）というツールが標準で提供されている．DIH は RDB や XML，ファイル等のデータソースから検索対象となるデータを取り出し，Solr に登録するツールである．DIH で取得したデータは項目ごとに，正規表現による文字列の置換や，日付，数値のフォーマット変換等を指定できる．

3. 4. 2　トークナイザとフィルタ

検索エンジンのインデックスを作成するには，文章を単語に分割するトークナイザと，表記ゆれへの対策や名寄せ等を行うフィルタを使用する．トークナイザとフィルタは情報アクセス基盤におけるデータクレンジング[*4]の機能の一部を担うものである．

Solr ではインデックスに格納する各データ項目に対し，トークナイザと複数のフィルタを組

み合わせて設定することができる．以下に主なトークナイザおよびフィルタを示す．

・　トークナイザ

日本語等のスペース区切りでない文章を単語に分割するためには，形態素解析方式またはN-Gram 方式のトークナイザを使用する．それぞれの特徴は表 1 のとおりである．

表 1　日本語で主に利用するトークナイザ

種類	説明	特徴
形態素解析	文章を形態素（意味を持つ最小の単位）に分割し，品詞を判別する技術	検索ノイズを少なくできることが長所であるが，辞書が必須でメンテナンスコストがかかる
N-Gram	検索対象を単語単位ではなくN文字の文字列で分割する技術	検索漏れが発生しにくく，辞書を使用しないところが長所であるが，検索ノイズが多く発生する

・　フィルタ

フィルタは文字を置換する文字フィルタと，トークナイザによる単語分割後に処理をするトークンフィルタの 2 種類があり，それぞれの特徴は表 2 のとおりである．トークンフィルタの一部はトークナイザが形態素解析である必要がある．

表 2　日本語で主に利用するフィルタ

種類	用途	説明
文字フィルタ	表記ゆれ対応（半角カタカナ・全角英数字・漢字異体字等）	変換前文字・変換後文字マッピング表を用いて，解析対象を変換する
トークンフィルタ	類義語検索（同義語，類義語，関連語）	類義語辞書を用いて，片方向（名寄せ）または双方向に対象の単語を展開する
	大文字・小文字の変換	解析対象の英字の大文字を小文字に変換する
	ストップワードの除外	解析対象から，検索に使用されない単語（「する」「ます」「です」など）を除去する
トークンフィルタ※形態素解析器が必要	単語の基本形出力	「楽しく」を「楽しい」などと基本形に変換する
	長音（ー）の削除	解析対象から4文字以上のカタカナの最後の長音（ー）を除去する
	品詞を指定した単語の除外	解析対象から指定した品詞（「助詞」「句点」「読点」など）をフィルタリングする

3. 4. 3　インデクシングへのカスタマイズ処理の追加

フィルタによる表記ゆれへの対策や，名寄せでは対応できないマスターテーブルの参照，プログラムによるデータの変換や補完が必要となる場合，Solr ではインデクシングにカスタマイズ処理を追加して対応することができる．例として，以下のようなものがある．

・　ファイルパスから独自のルールでタグや分類を付加する
・　口コミの文章から特徴語を抽出して属性として付加する
・　テキストマイニング技術で対象データに自動的にカテゴリを設定する

3. 4. 4　ダイナミックドリルダウン

インデクシングされたデータを検索する機能として，キーワードによる検索以外で重要なのがダイナミックドリルダウンと呼ばれる機能であり，Solr では「ファセット機能」として提供されている（図 4）．ダイナミックドリルダウンは，検索結果全体をカテゴリや日付などで絞り込み，利用者をナビゲートする仕組みであり，オンラインストア等の検索機能において必須の機能となりつつある．また，この機能は検索結果の中から指定した属性を持つ件数を項目ごとに把握できるため，SBA では分析・統計の目的に使用される．

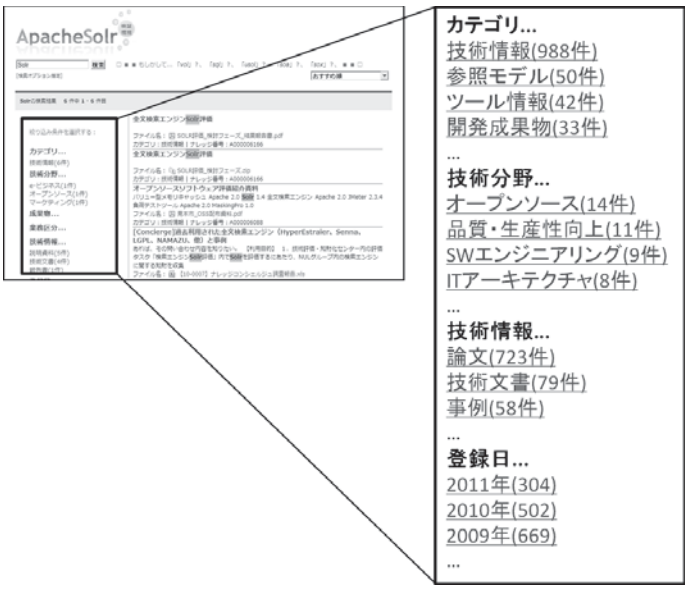

図 4　Solr によるダイナミックドリルダウンの例

4.　仮想データ統合を実現するデータ設計

本章では，SBA の構築事例である品質管理に対する課題解決ソリューションをもとに，効果的な SBA を構築するためのポイントとなる情報アクセス基盤のデータ設計について記述する．

4. 1　品質管理に対する SBA 構築事例の概要

製品の事故・トラブル，ひいては製品回収といった事態が発生し，その対応に多大な人的・経済的コストを費やすことがある．そういった事態への予防策として，Web 上に公開されているエンドユーザーの口コミ情報と社内の製品情報を検索技術によって掛け合わせ，視覚的に見せる SBA を構築した．この SBA にて，ある製品に関するクレームが増えていることを検知することにより，リコールにつながる不具合の予兆や，問題のある製品と同じ部品を使用している他の製品を調査し，更なる事故を未然に防ぐことを目標とした．

4. 2　企業内外のデータを繋ぐ情報アクセス基盤のデータ設計

企業内外のデータを繋ぎ，効果的な SBA を構築するためには，複数のデータソースから必要な情報を選定し，用途を踏まえて適切に関連付けることがポイントとなる．

データ設計では，まず Web 上の口コミ情報と社内の製品データから，必要とされるデータを入力として選定する．また，データ同士の関連付けが意味のある情報連鎖となる場合，どのようにデータを関連付けるかを定義する．データの出力としては，どの項目をどのように表示したいか，画面構成を検討する．最後にデータの入力と出力を合わせ，検索システムのインデックス構成を設計する．

4. 2. 1　入力データの選定

社内外のデータから入力とするデータを選定し，データの関連付けを整理する．検索システムによって情報を統合する場合，複数のデータソースを関連付け，一つのビッグテーブルを作成するように設計する．データソース同士が共通の項目を持たず，データの関連付けに必要な情報が存在しない場合は，何らかの規則をもとにデータを変換するプログラムを作成したり，マッピングテーブルを作成したりすることで，関連付けを実現する．

図5に，Web 上の口コミ情報と，社内のデータベース内の製品情報および部品情報を選定し，関連付ける例を示す．この例では，インデクシングに以下のカスタマイズ処理を追加してデータを関連付けている．

- 口コミ情報に含まれる製品の通称を，正式名称にマッピングし，社内で管理されている製品情報に関連付ける
- 部品 ID が拠点間で異なるため，マスターテーブルを参照して統一した ID に置き換える

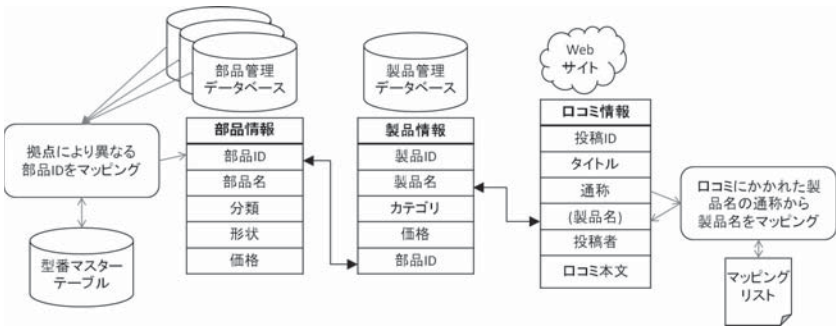

図5　データ関連図の例

4. 2. 2　画面構成の検討

検索結果を表示する画面の設計では，何のために，どのように検索（データを取得）し，どのような結果が得られることで，何が利点となるかを検討する．

本章の事例では，リコールにつながる不具合の予兆を検知するため，製品ごとに，クレームおよび不具合件数の時系列推移を視覚的に見たいという要望があった．そのため，通常の検索画面の他に，選択した製品のクレーム件数を縦軸，年月を横軸とする折れ線グラフの画面等を取り入れた．図6に，リコールの予兆を検知する例を示す．クレーム件数の製品別の時系列推移から，クレーム件数が増えている製品を選択すると，その製品で使われている部品別の時系列推移に画面遷移し，特定の部品を使用している製品全体のクレーム件数の傾向が見られる．グラフ等の画面は簡易的なビジネス・インテリジェンス（BI）[*5] として利用可能であり，また BI と異なり数値は検索結果数であるため，因子となる検索結果を動的に表示できる．

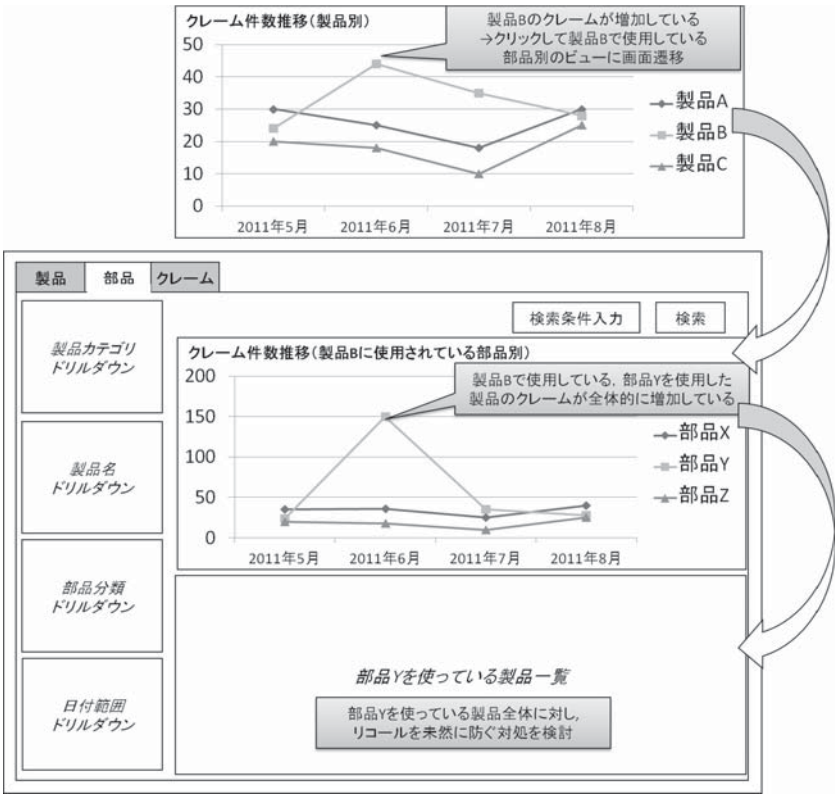

図 6　クレーム件数の時系列推移からリコールの予兆を検知する例

5.　OSS 検索エンジンを活用した情報アクセス基盤構築における考察

5. 1　要求状況

日本ユニシスでは，製造業や公共・医療分野等の案件に Solr を適用した実績がある．その中で，顧客の現行システムや業務における情報共有等の顕在的，潜在的な課題を引き出すため，顧客へのヒアリングを継続している．表 3 に，OSS の検索技術および SBA に対する要求状況をまとめる．表中の「仮想データ統合・分析に関するニーズ」が SBA による解決範囲である．

5. 2　OSS 検索エンジン適用における留意点

プロジェクトでの Solr の適用を経て洗い出した OSS 検索エンジン適用の要件定義における検討事項を表 4 に記述する．OSS に限らず，一般的な検索システムにおいても共通する検討事項がほとんどであるが，Solr は複数の OSS を組み合わせて使用するため，一部その点における検討事項を含んでいる．

Solr と組み合わせる OSS には，Web・ファイルクローラ，形態素解析器などがあり，要件によって選定する．対応するデータソースの種別や，アクセス権限情報取得の要否により，採用するクローラが異なり，インデクシングに必要な時間も異なる．また，検索においても，全体および 1 ファイルあたりのファイルサイズや使用する検索機能によって検索応答時間が大きく変動する．そのため，実際のデータを使用して事前にインデクシングおよび検索性能を検証することが望ましい．

表3　顧客課題・ニーズと解決策

分類	課題・ニーズ	解決策
情報共有	関連部署で同じような文書を作成しているが共有できていない	部門毎のファイルサーバーを横断した情報共有
	共有フォルダの中にある必要なファイルの所在がわからない	
	ファイル名では内容がわからず，一つ一つファイルを開いて確認する作業が発生する	ファイル内容を示す属性の抽出と検索結果画面への表示
	海外支部との情報共有が難しい	多言語への対応と，支部間でのデータの関連付け
	商用製品のライセンス費用が高額である	他の商用検索製品からOSSへのリプレース
	社内で使用している検索システムでは欲しい情報がみつからない	
仮想データ統合・分析	インターネット上のエンドユーザーのクレーム情報などから，リコールの予兆を検知したい	ソーシャルメディア等社外のデータと社内のデータの仮想統合と分析
	口コミ情報を可視化し，マーケティングに利用したい	
	情報を得るには複数のシステムにログインして検索する必要があり，効率が悪い	社内に乱立した情報蓄積システムの仮想統合
	社内システムのデータベース統合をするにはコストが見合わず実現できない	
	SFA[※6]などの社内システムは独立しており，営業状況等の必要な情報をタイムリーに入手できない	SFAと必要な外部データの連携によるセールスプロセスの改善
	アクセスログや購買履歴ログが活用できていない	ログの可視化によるユーザ行動の把握・分析
	コンプライアンスの違反を検知するために，多大な人的コストを費やしている	メールや文書内にある違反となり得る単語の使用状況を，視覚化しアラート

表4　要件定義における検討事項

分類	項目	検討内容
対象とするデータの取り扱い	ファイル形式，データソース	・採用するクローラ・対象外とするファイル（ファイルサイズの上限の設定など）
	対象言語の種類，言語処理	・トークナイザ，フィルタの選定 ・類義語辞書の必要性
	検索対象へのアクセス方法 　必要なアクセス権限 　検索対象サーバーのリソース，運用状況	・クロール時の認証方法の検討 ・検索対象サーバーの計画的再起動等の運用状況に合わせたクロールのスケジュール
	アクセス権限への対応 　セキュリティポリシーの確認 　対応するユーザ認証の種類 　アクセス権限の更新頻度	・採用するクローラ ・クライアントアプリケーションでの認証と検索結果のフィルタリング方針
スケーラビリティ	対象データの量 　合計ファイル数/サイズ， 　１ファイルあたりのサイズ， 　テキスト量	・クロール/インデクシングに必要な時間とスケジュール ・分散検索の必要性
	対象データの更新頻度と更新量	・差分クロールの必要性 ・差分クロールの頻度，タイミング
	ユーザ数と同時アクセス数	・負荷分散の必要性
	性能 　検索応答時間（QPS） 　インデクシング時間	・キャッシュなどのチューニング ・サーバー構成の検討およびH/Wのサイジング
	拡張性	・ユーザ数の増加，データの増加に対する方針
運用	監視項目，運用方法	・各種チューニング項目，辞書などの運用方針 ・障害時の通知などの方針 ・保守，サポートの方針と範囲
追加開発	アプリケーションとの連携やカスタマイズ	・使用するS/Wの検討 ・インデクシング等へのカスタマイズ処理追加の必要性 ・追加開発の要求

6. お わ り に

従来のデータベース統合と比較し，検索技術による情報アクセス基盤の利点として，データベース内のデータだけではなく，インターネット上の口コミ情報や，社内のファイルサーバーのファイル等の非構造化データも横断的に扱うことができる点と，スケーラビリティが高い点があげられる．2012年現在，爆発的に利用が拡大したソーシャルメディア内の情報や，e-コマースでのユーザ行動を示すクリック履歴や購買履歴のログ等を分析したいというニーズが高まっている．しかし，それらの大規模な非構造化データを，どのように扱うかが企業においての課題となっており，検索技術による情報アクセス基盤はそのような課題に対する解決策の一つとなり得る．ログ分析のような大規模データ処理において，インデクシングの前にデータの加工が必要となる場合は，近年多数実例が報告されている Hadoop[*7] 等の分散処理技術との組み合わせが有効である．

今後も企業内外に存在する分析・活用対象となるデータは更に大量化・多様化すると考えられる．そのようなデータの取り扱いにおける課題の解決策の一つとして，本論文が参考になれば幸いである．

＊1　転置インデックスとは，全文検索を高速に行うことを目的とした，単語をキーとしその単語を含むドキュメントの ID のリストからなるテーブルのことである．

＊2　Apache Solr は Apache Lucene（Java を基盤とした検索ライブラリー）上に構築されたウェブ・サービス層である．Lucene を直接利用するには情報検索分野での経験および大掛かりなプログラム構築が必要となるが，Solr を利用することでコードを最初から作成するコストとリスクなしで，拡張性の高い検索機能が得られる．

＊3　Apache Solr は HTTP を介して呼び出し可能な REST 類似のウェブ・サービス・インターフェースを持ち，XML ベースの構成ファイルを使用することで機能の設定が可能．Java をはじめとし，C#，Perl，Ruby 等の API も備える．

＊4　データクレンジングとは，データの統計をとる目的で，データ形式の統一や，欠損値の補完，データの正規化等を行うことである．

＊5　ビジネスインテリジェンス（BI）とは企業に蓄積されたデータを集約・整理・分析し，経営上の意思決定に役立てる手法である．企業内に散在する情報をデータウェアハウスに蓄積し，レポーティングや，データマイニングを通じて，高度に活用することを目的とする．

＊6　SFA（営業支援システム，sales force automation）は営業を支援するシステムである．データベースに顧客情報，コンタクト履歴や商談のプロセス，営業スケジュールを蓄積し，営業案件の進捗状況や案件成立の見込みをチーム内で共有する．

＊7　Apache Hadoop はオープンソースの大規模データの分散処理を支える Java ソフトウェアフレームワークである．

参考文献　［1］ Gregory Grefenstette and Laura Wilber, “Search-Based Applications - At the Confluence of Search and Database Technologies”, Morgan & Claypool Publishers, 2011

［2］ Leslie Owens, “Tapping The Power Of Search-Based Applications”, Morgan & Claypool Publishers, March 2011

［3］ 清兼義弘．関口宏司．田澤孝之．松野良蔵，「エンタープライズサーチ 技術と導入」，アスキー・メディアワークス，2008 年 9 月

［4］ 吉川日出行，「サーチアーキテクチャ「さがす」の情報科学」，みずほ情報総研株式会社，ソフトバンククリエイティブ，2007 年 10 月

［5］ Laurie F. Wurster, Bob Igou, Zeynep Babat, “Survey Analysis: Overview of Preferences and Practices in the Adoption and Usage of Open-Source Software”, Gartner, Inc., January 2011

［6］「国内ソフトウェア市場予測を発表」，IDC Japan 株式会社，2011 年 5 月 24 日，http://www.idcjapan.co.jp/Press/Current/20110524Apr.html

［7］「2010年度オープンソースソフトウェア活用動向調査」，The Linux Foundation Japan，2011年7月，
http://www.linuxfoundation.jp/jp_uploads/SI_Forum_OSS_Survey_2010.pdf
［8］ Aaron Wall, “History of Search Engines: From 1945 to Google Today”,
http://www.searchenginehistory.com/

※参考文献［6］～［8］に含まれるURLは，2012年2月時点での存在を確認

執筆者紹介　石　井　　愛（Ai Ishii）
2003年日本ユニシス(株)入社．オープンミドルウェア製品主管部にて保守・開発に従事．2009年より総合技術研究所に移籍し，検索技術領域の評価と研究を主体とした活動をする．