学術情報処理研究 No.20
2016年9月

徳島大学情報センターにおける全文検索システムの導入検討

Considering adopting a full-text search on the Center for Administration of Information Technology, Tokushima University

谷岡広樹 †, 板東孝文 †, 松浦健二 †

Hiroki Tanioka†, Takafumi Bando†, Kenji Matsuura†

tanioka.hiroki@tokushima-u.ac.jp, bandou.takafumi@tokushima-u.ac.jp, ma2@tokushima-u.ac.jp

徳島大学情報センター †
The Center for Administration of Information Technology, Tokushima University†

概要

徳島大学情報センターは, ISMS に基づく情報セキュリティポリシーに則り, 教員及び職員が作成した ISMS 文書をファイルサーバで管理している. ISMS 文書以外の本センターが関わる業務文書, 契約書, マニュアル, ログ等といった業務運用系文書についても, 同一のファイルサーバで管理している状況である. ISMS 文書については, ディレクトリ構造やファイル名に運用規定を設けることによって, 必要な人が必要なときに使用できる状態を維持している. しかしながら, 教員及び職員全員が, ファイルサーバのディレクトリの最新状況を常に把握することは困難なため, ISMS 文書やその他の必要書類を即座に使用できない場面があるのも事実である. この状況を改善するため, 我々は, 本センター内で利用するファイルサーバに全文検索システムを導入することを検討した. 本論文では, 全文検索システムの要件定義, システム設計, 導入範囲の見積もり, 性能評価を行い, その導入効果を検証した結果を報告する.

キーワード

文書管理, ISMS, 情報検索, Elasticsearch

1 はじめに

徳島大学情報センター (平成 26 年度より情報化推進センターから改組) では, 情報セキュリティ対策や情報セキュリティポリシーの構築や運用に関する仕組みとして, 情報セキュリティマネジメントシステム (以下 ISMS) [*1] として知られている国際標準の規格 ISO/IEC 27001 に基づき, ISMS を確立, 導入, 運用, 監視, レビュー, 維持及び改善を実施している [1].

本センターでは, 文書管理システムとしてファイルサーバ [*2] を利用しており, ISMS 文書と関連文書の文書群を, ファイル名の命名規則を策定して管理している. ISMS 文書以外の関連文書には, 業務の属人化を避けるための手順書, 台帳, 様式などがある [2] が, 同一のファイルサーバ上に, 教員及び職員 (以下, スタッフ) が自由に利用可能な領域や, 一時的に情報共有する目的で利用可能な領域がある.

1.1 業務上の課題

スタッフが業務で必要となった ISMS 文書, 手順書, 台帳, 様式などを参照する場合, ディレクトリ構造とファイル名の命名規則について把握しているスタッフは, ISMS 文書及び関連文書についてファイル名の命名規則が策定されているので, 目的の文書を比較的容易に参照可能である. しかしながら, 配属間もないのスタッフは, ディ

[*1] Information Security Management System
[*2] Windows Server® 2012 R2: Disk 1.4 TB, NIC 1.0 Gpbs

レクトリ構造とファイル名の命名規則について把握していないため, 自力でファイルサーバ上の文書群から見つけ出すことが困難であり, ファイルの探索に多大なコストを割くか, 他のスタッフに協力を仰ぐ必要がある. このような事例は, ベテランスタッフであっても, 担当業務以外の関連文書を参照する必要がある場合に起こりうる.

前述の事例の 1 つ 1 つは軽微な事案であるが, 年間を通して行われる日常の業務負荷の総量は, 1 人のスタッフの業務負荷が 1 日あたり平均 5 分としても, 年間 240 日で 1,200 分 (20 時間) の作業時間増となるため, 看過できない. また, 本センター内で突発的に発生する機器の障害, 情報セキュリティインシデント, 学生や教職員からの要望やクレームといった各種インシデントに対して, 素早く対応するためには, 当該インシデントの関連文書を素早く参照する必要がある. これらの事案から, 本センターのファイルサーバを用いた文書管理には, 次の 2 つの業務上の課題があると考えられる.

- 定常的な業務負荷の低減
- 突発的なインシデント対応時間の短縮

1.2 全文検索システムの導入検討

業務のオペレーションの中で文書管理に関連したオペレーションの内, 必要となる文書を参照するオペレーションを素早く行えるようにする手段の 1 つとして, 全文検索システムが考えられる. 本センターが文書管理のために利用しているファイルサーバに対して, 全文検索システムを導入することで, スタッフが必要とする文書に素早く参照できるようになれば, 文書管理に関連した業務のオペレーションを簡略化することが可能となる. その結果, 本センターの 2 つの業務課題「定常的な業務負荷の低減」と「突発的なインシデント対応時間の短縮」の解決につながると考えられる.

本研究では, 文書管理に関連した業務のオペレーションを簡略化することを目的とし, まず, 本センターのファイルサーバ上に格納された文書のうち, 本センターのスタッフが必要とする文書にはどのようなものがあるか, どのような手がかりを用いて参照しようとするかといった観点でアンケートをとる. このアンケートの分析結果に基づいてテストセットを作成し, 実験することで, 本センター内のファイルサーバを対象として全文検索システムを導入することで, どの程度の業務改善が可能かという観点で, 全文検索システムの導入効果を検証する.

本論文では, 国立大学法人におけるファイルサーバに対する全文検索システムの導入効果について, 導入前の実験結果から見積もった結果を報告する. 続く 2 章では, 本センターのファイルサーバの利用状況, 3 章では, 全文検索システムの要件と設計, 4 章では, 全文検索システムの導入範囲と実験方法, 5 章では, 実験結果と考察, 6 章では, まとめとして今後の課題と展望について述べる.


ファイルサーバ利用者アンケート

図- 1: アンケートの内容

表- 1: 文書を参照する 1 日平均の作業時間の統計

	最大	75%	平均	中央	25%	最小
作業時間	12.00	8.00	8.05	7.88	7.19	6.00
端末作業時間	10.00	8.00	7.13	7.38	6.25	3.50
文書作成・閲覧時間	7.00	2.75	2.20	1.75	1.00	0.00

* 数値はすべて自己申告, 単位は時間 [h], 有効桁数 2 桁.

2 ファイルサーバの利用状況

本センターで利用しているファイルサーバの利用状況を明らかにするために, スタッフ全 19 名 (教員 5 名, 技術系職員 8 名, 事務系職員 6 名) を対象に, 以下のような内容のアンケートを行い, その利用状況について調査した.

1. 業務内容
2. 作業時間
3. 端末での作業時間
4. 文書作成・閲覧時間
5. 業務で参照する文書

業務で参照する文書の項目については, 図-1 のように, 思いつく限りの文書のタイトル, ファイル名, 参照日時, ファイルの内容を列挙する形式とした.

2.1 作業時間の分析

スタッフの業務時間と, 業務として文書を参照する 1 日平均の作業時間について, 集計した結果を, 表-1 および図-2 に示す. この結果から, 情報センターのスタッフの日常業務において, 平均して 2.2 時間 (約 132 分), 文書を閲覧または編集していることがわかる. スタッフ 1

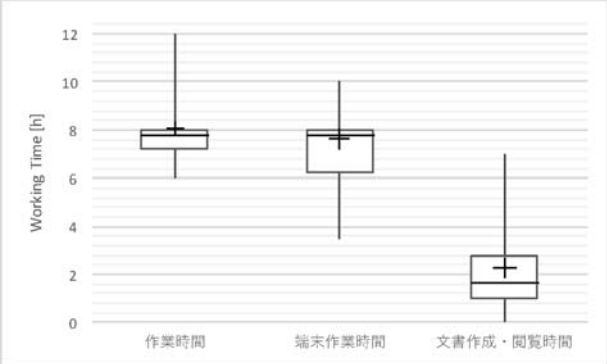

図- 2: 文書を参照する作業時間の分析

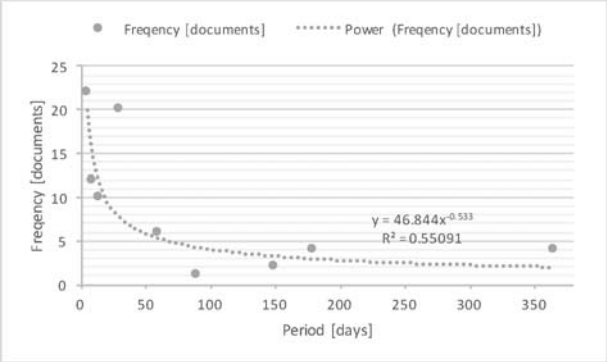

図- 3: 文書を参照する期間の分析

表- 2: アンケート結果の一部

タイトル	ファイル名 (ファイルパス)	参照日時	ファイルの内容
鍵管理台帳	【AIT2011-01-F32-0001】鍵管理台帳20160701.xlsx	7/14/2016	ISMS のセキュリティ目的
係・講座変換テーブル	a_職員証発行担当マスター_rev5.xlsx	7/14/16	職員証発行関連のマスター参照
受付履歴管理台帳	受付履歴管理台帳_pg	毎日	対応履歴管理
情報センター朝会議事録	【AIT2013-00-H96 - 0005】情報センター 朝会議事録 H2806.txt (... 情報センター朝会 \H28)	月次作業	朝会議事録（月毎のファイル）
内部監査報告書	\\192.168.**.**\... \ISMSdocs\0data\COM-B04-D05 内部監査	1ヶ月以内	ISMS 内部監査資料
脆弱性診断関連ファイル		2-3 月	学内設置端末の脆弱性診断用の告知用・診断結果等のファイル

* 前提として，スタッフが記憶している限りでの回答を求めたため，業務で参照する文書について，タイトル，ファイル名（ファイルパス），参照日時，ファイルの内容は，一部の情報が空欄の場合がある．

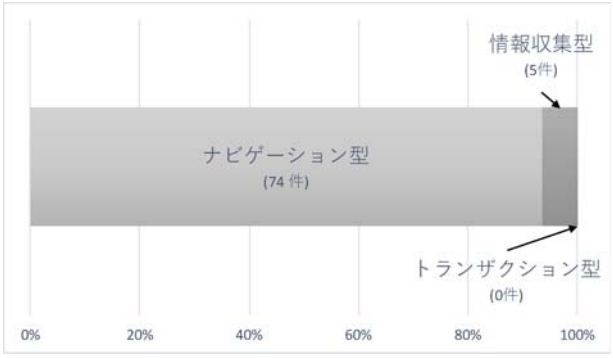

図- 4: アンケート結果の分析

名が文書 1 件につき閲覧または編集している時間を 10 分と仮定すると, 閲覧または編集される文書数は, 1 日平均 13.2 文書, スタッフ 19 名により閲覧または編集される文書数は, 1 日平均約 250 件となる.

また, ファイルサーバ上の文書をスタッフ 19 名が同時に参照しようとした場合, ファイルサーバおよび全文検索システムへの最大同時アクセス数は 19 である.

2.2 参照期間の分析

アンケートでは, スタッフが参照する文書への, 各自の記憶に基づく参照日時を回収した. 図-3 は, スタッフが記憶していた文書数とその参照日時の関係を示す.

スタッフが1ヶ月以内に業務で参照する文書がもっとも多く, 次に半年以内, さらに 1 年以内に参照する文書までがある. べき乗関数による回帰曲線によって次式のように近似できる.

$$y = 46.844x^{-0.533} \tag{1}$$

式 (1) は, スタッフが文書を参照する頻度が, その期間のべき乗に比例するべき乗則に従っていることを表す. これは, Web ページへの訪問 [3], UNIX コマンドの同じコマンドの入力 [4], 図書館での本の貸し出し [5], さらには人間の記憶に関する研究 [6] とも類似するパターンである.

2.3 検索意図の分析

表-2 は, アンケートの結果の一部である. このアンケート結果のそれぞれの文書について, 検索システムを利用して素早く見つけたい場合の検索意図を, ナビゲーション型, 情報収集型, トランザクション型の 3 つの検索意図 [7] のいずれかに分類した.

ナビゲーション型
ある特定の情報源を見つけ出すためにある 1 つの検索結果を得ることを目的とする.

情報収集型
あるトピックに関連する情報を収集するために複数の検索結果を得ることを目的とする.

トランザクション型
商品を探して購入したり, 旅行の予約をしたりといった取引を目的とする.

たとえば, 「鍵管理台帳」や「係・講座変換テーブル」は, 特定の情報源を見つけ出すことを目的としており, ナビゲーション型であるといえる. また, 「内部監査報

表- 3: ファイルサーバの状況

ディスク容量	1.4 TB
ディスク使用量	845 GB
ファイル数	139,694
ディレクトリ数	19,074
平均ファイルサイズ	6.39 MB
ファイル種別 (拡張子種別)	1,829
ディレクトリごとの平均ファイル数	7.32
ファイルごとのディレクトリの平均深さ	10.35

* 2016 年 7 月 20 日時点でのファイルサーバの状況.

表- 4: ファイルの種別の上位 20 件

順位	拡張子	件数	順位	拡張子	件数
1	pdf	21,163	11	tif	4,589
2	txt	13,876	12	html	3,930
3	xlsx	9,615	13	db	2,795
4	(なし)	7,196	14	log	2,793
5	xls	7,071	15	js	2,190
6	jpg	5,842	16	php	2,179
7	JPG	5,728	17	doc	2,179
8	docx	5,297	18	csv	1,776
9	gif	4,707	19	msf	1,542
10	png	4,592	20	pptx	1,500

* ファイル種別（拡張子）上位 20 件が, 全体の 8 割を占める.

告書」や「脆弱性診断関連ファイル」は, あるトピックに関連する報告書などの情報を収集することを目的としており, 情報収集型であるといえる. 一方, トランザクション型の意図を持ってファイルサーバから文書を探す場合は, 本アンケート結果には含まれなかった. 徳島大学においても, 人事や調達を担う部局については, トランザクション型の意図を持った文書の探索は存在しうるが, 本センターは情報インフラおよび情報セキュリティーの維持を担う部局であり, アンケート調査時期において, ファイルサーバを利用する業務に, 商品や旅行といった取引を目的とする業務が含まれなかったためであると考えられる.

図-4 は, アンケートの集計結果を検索意図に分類した場合の割合である. アンケート結果に含まれる文書のほとんどが, ナビゲーション型であった. このことから, 本センターのスタッフの日常業務においては, 概ねナビゲーション型の検索意図を持って, 全文検索システムを利用するであろうことが予想される.

2.4 ファイルサーバの分析

本センターが利用するファイルサーバは, Windows Server® 2012 R2 で構築されており, ネットワーク通信速度は 1.0 Gpbs, ディスク容量は 1.4 TB, 2016 年 7 月 20 日時点でのディスク使用量は 845 GB (60.35 %), ファイル数 139,694, 平均ファイルサイズ 6.39 MB である. その他の統計データは, 表-3 に示す通りである.

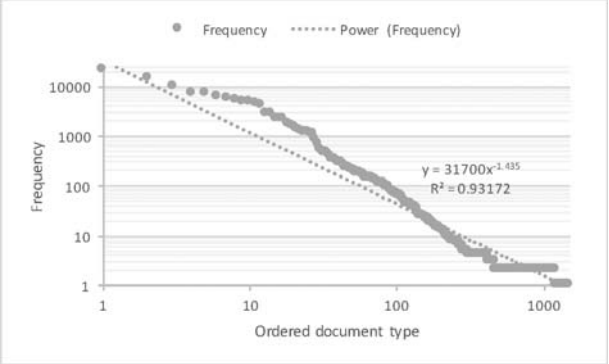

図- 5: ファイルの種別とファイル数

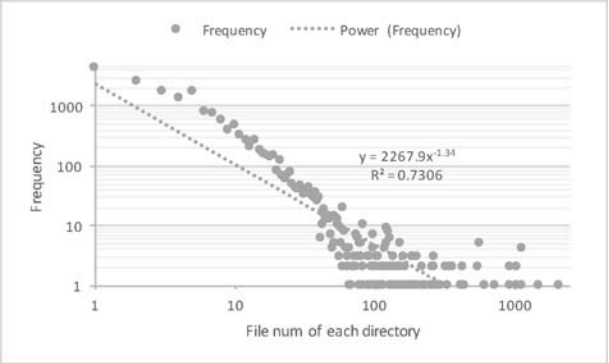

図- 6: ディレクトリごとのファイル数の分布

2.4.1 ファイル種別とファイル数の分布

ファイル種別ごとのファイル数は, 表-4 に示す通り, PDF, テキスト文書, オフィス文書, 画像ファイル, ログファイル, プログラムファイルがその多くを占めている. また, ファイル種別ごとのファイル数の分布は, 図-5 であり, ファイル数順に並べたファイル種別に対するファイル数は, べき乗関数による回帰曲線によって次式のように近似できる.

$$y = 2267.9x^{-1.34} \tag{2}$$

ディレクトリごとのファイル数の分布は, 図-6 であり, ディレクトリの深さに対するその頻度は, べき乗関数による回帰曲線によって次式のように近似できる.

$$y = 31700x^{-1.435} \tag{3}$$

このことにより, ファイル種別ごとのファイル数の分類およびディレクトリごとのファイル数の分布は, スケール不変性を有しているため, ファイル種別やディレクトリごとのファイル数を考慮した文書の探索コストの見積もりが, 比較的容易に可能となる.

2.4.2 ディレクトリ深さの分布

ファイルごとのディレクトリの深さの分布は, 図-7 に示すように, ディレクトリの深さ 9 と 11 に最頻値があ

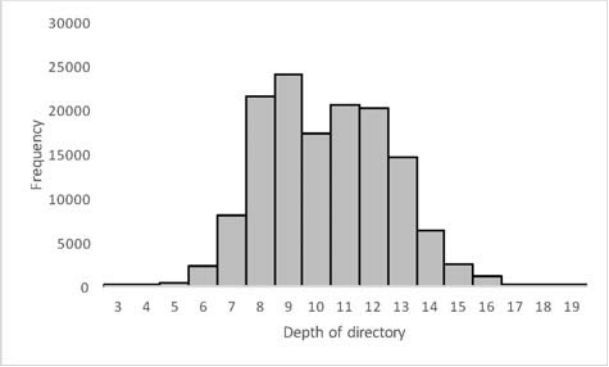

図- 7: ファイルごとのディレクトリ深さの分布

表- 5: ファイルごとのディレクトリ深さの統計

最大	75%	平均	中央	25%	最小
19	12	10	9	8	3

* ファイルごとのディレクトリ深さ (階層).

る二峰性分布である. 平均は 10, 25 パーセンタイルから 75 パーセンタイルは 8 〜 12 の範囲であった. 文書を探索する際に, 平均で 10 階層, 多くの場合 8 から 12 階層の範囲で, ディレクトリを移動する必要があることを意味する.

2.4.3 文書の探索時間

スタッフのファイルサーバの利用状況とファイルサーバの状況から, スタッフが必要とする文書を参照するための探索時間を見積もる. 実際には, 記憶違いやディレクトリ構成の変更などを原因とする再探索が何度か行われることになると予想されるが, 本論文では, スタッフがルートディレクトリから一度も迷わずに, 目的のファイルを参照できるまでを見積もることとする.

あるファイルをファイルサーバのルートディレクトリから辿って, 目的のファイルにアクセスするまでの平均探索時間 T_f [s] を, ディレクトリごとの平均ファイル数 N_d, ファイルごとのディレクトリの平均深さ D_f, スタッフのスキルやネットワーク環境などの要因に依存する 1 文書あたりの平均処理時間 α [s] を用いて次式で定義する.

$$T_f = \alpha N_d D_f \tag{4}$$

$N_d = 7.32$, $D_f = 10.35$, $\alpha = 0.5$ とすると, ファイルを探索する平均探索時間 T_f は 37.88 [s] となり, 1 文書を探索する平均時間は約 38 秒かかることが予想される.

年間を通して行われる日常の業務負荷のうち, 文書の探索にかかる時間を 38 秒, 閲覧または編集される文書数を, 1 日平均 13.2 文書とすると, スタッフ 1 人が 1 日に文書の探索にかける時間は 501.6 秒 (8.36 分) である. これは文書を閲覧または編集している時間の平均

表- 6: 全文検索システムの機能要件

機能要件	内容
ファイルサーバ接続	SMB
全文検索機能	クエリ入力, 結果表示
検索結果絞り込み	ファイル種別
検索結果並べ替え	検索スコア, 更新日時
ファイル本文解析	TXT, PDF, Office 文書
日本語の分かち書き	形態素解析
Web サーバ	HTTP

表- 7: 全文検索システムの非機能要件

非機能要件	性能
最大文書数	30 万件
平均レスポンス時間	2 秒
最大同時アクセス数	20
平均検索時間	0.1 秒
連続稼働時間	17 時間
インデックス更新時間	7 時間
平均文書追加時間	0.5 秒

2.2 時間 (約 132 分) のうち 6.3 % が文書を探している時間ということになる. また, スタッフ 19 名の年間の作業時間に換算すると 2,006 分 (33.44 時間), 年間で約 4.16 人日のコスト増となっている.

3 全文検索システムの要件と設計

3.1 機能要件

全文検索システムの機能要件は 表-6 とした. 本センターで利用しているファイルサーバへの接続には SMB(Server Message Block) を利用する. 全文検索機能を実現するためには, クエリ入力と結果表示のそれぞれの機能が必要となる.

アンケートの集計結果から, 全文検索システムの利用者は, そのほとんどがナビゲーション型の検索意図を持って文書を探索するであろうことから, Dumais らによる SIS(Stuff I ' ve Seen) の報告 [8] を参考に, ファイル種別による検索結果の絞り込み機能, 更新日時による並べ替え機能を備える.

ファイル種別の統計情報で上位に多く含まれる TXT, PDF, Office 文書をファイル本文の解析対象とする. また, そのほとんどが日本語であることから, 分かち書きのために形態素解析機能を備える. ユーザインターフェースとしては Web ブラウザの利用を想定しているため, Web サーバ機能も必要となる.

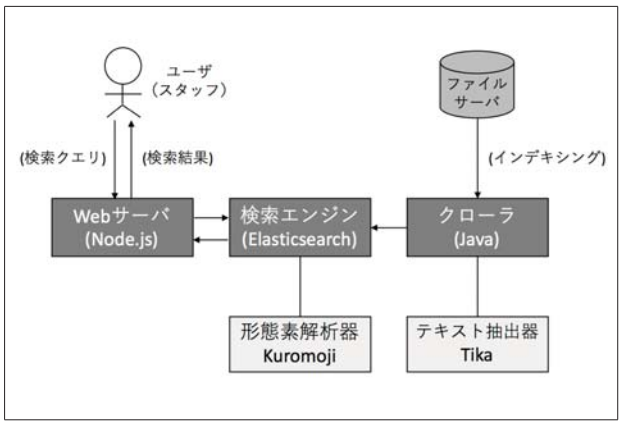

図- 8: 全文検索システムの全体構成

図- 9: 全文検索システムのインターフェース

3.2 非機能要件

直接的な機能として現れない非機能要件を, 表-7 に列挙する. ファイル総数は 139,694 件, ファイルサーバの使用率は 60.35 % のため, ファイルサーバの使用率が 100 % の場合を想定して, 最大文書数を 30 万件とする. また, 本センターの日常業務での使用で, 平均レスポンス時間を 2 秒以内とし, スタッフが特定の時間帯に集中して利用した場合を想定して, 最大同時アクセス数を 20 としたとき, 検索エンジンに許容できる平均検索時間は 0.1 秒である.

さらに, 午前 5 時から午後 10 時までの連続稼働時間を 17 時間としたとき, インデックスの更新時間にあてられるのは, 業務に支障のない時間帯である午後 10 時から午前 5 時までの 7 時間以内である. 平均文書追加時間は, ネットワークに大きな影響を与えない範囲として, 1 Gbps の 1 % の 1 Mbps を占有してもよいと仮定すると, 1 秒あたり 12.5 MB をファイルダウンロードできるので, 平均ファイルサイズ 6.39 MB に対して, 1 文書あたり平均 0.5 秒で追加できればよいと考えた.

3.3 全文検索システムの構成

全文検索システムの構成は, 図-8 とし, すべての構成要素は, 1 台の PC [*3] 上に構築する. まず, 非機能要件から, 検索対象文書数 30 万件に対して平均検索時間 0.1 秒で検索できなければならない. この性能を達成するため, 検索エンジンには転置インデックス型のものを採用する. また, 機能要件から, 検索エンジンは, 並べ替えや絞り込みができなければならない. このため, 転置インデックス型で, かつ REST API を利用して並べ替えや絞り込みの検索条件が変更可能な全文検索システム Elasticsearch [9] を採用した.

ファイルサーバに接続し, 検索対象文書のインデキシングに必要なクローラは, Java 言語を用いて開発した. 機能要件から, PDF や Office 文書といったバイナリデータの本文を解析対象とするため, Apache Tika [10] を用いた. また, 日本語の分かち書きと品詞分析を行うための形態素解析には, Kuromoji [11] を用いた.

非機能要件から, インデックスの更新を 7 時間で実行する必要があるが, 最大文書数 30 万件のインデックスを作成するためには, 1 文書を平均 0.5 秒でインデックスに追加できたとしても, インデックス更新時間は約 42 時間, 1 日あたり 7 時間インデキシングしたとして 6 日間を要するので, 初期インデックスについては, 全文検索エンジンの運用開始前に作成する必要がある.

運用開始後については, 仮に最大文書数 30 万件の 1 割に相当する 3 万件の文書が, 1 日で増加したと仮定すると, 1 文書を平均 0.5 秒でインデックスに追加できれば, 約 4 時間で登録可能であるため, 問題ないといえる.

Web サーバには, 同時アクセスに対する非同期処理と, Elasticsearch の REST API で用いる JSON(JavaScript Object Notation) データへの親和性の高さから node.js を採用した.

3.4 インターフェース

ユーザインターフェースは, 図-9 とした. 機能要件から, (1) クエリ入力欄と (3) 検索結果一覧を備える. また, ファイル種別による (5) 絞り込み機能, 更新日時による (2) 並べ替え機能を追加し, 一般的な Web 検索システムと同様に, (4) ページング機能を備え, 検索結果にはファイル名, ファイルの所在 (ファイルパス), 本文の一部を表示する. ファイル名またはアイコンをクリックすることで, ファイルを直接参照できる.

[*3] Mac OS® X: Version 10.11.6, CPU 2.5 GHz, Memory 16 GB, Disk 500 GB, Network 802.11n

```
[taniokah$ tree -d -L 1 public
public
|-- 00.admin
|-- 01.users
|-- 02.manuals
|-- 03.tmp
+-- 99.old

5 directories
```

図- 10: 対象とするディレクトリ

4 導入範囲と実験方法

対象となるファイルサーバの範囲, テストセットの作成方法と実験方法, 効果を測定するための精度評価の方法について述べる.

4.1 導入範囲

導入範囲は本センターで利用しているファイルサーバ上で公開されている共有リソース全体とする. ルートディレクトリ public の直下に, 図-10 のように 5 つのディレクトリに分けて管理されている. 隠しファイル (.ファイル), サムネイルファイル (Thumbs.db) を除くすべてのファイルを検索対象とする.

また, 画像ファイルや, 本文の抽出が困難なバイナリファイルについても検索対象とするために, ファイル名やディレクトリ名を含むファイルパスを検索対象とする.

4.2 実験方法

全文検索システムを導入する前に, 導入効果のベースラインを把握する目的で, 全文検索システムのレスポンス時間と検索精度を計測した. 実験にはまず, 前述した導入範囲の文書を全文検索システムへ登録する. 次にアンケート結果を元に検索対象となりうる文書を正解セットとし, その文書を探す手がかりとなるキーワードを検索クエリとしてテストデータとする. 実験では, 正解セットのキーワードを全文検索システムのクエリとして検索する.

4.2.1 テストデータ

本実験においては, アンケートで得られた 63 文書のうち 17 文書を正解セットとし, それぞれの「タイトル」「ファイル名 (ファイルパス)」「ファイルの内容」を検索キーワードとして, テストデータ 51 件を生成した. 表-8 はその一部である.

表- 8: 正解セット

ファイル名	キーワード
【AIT2011-01-F32-0001】鍵管理台帳 20160701.xlsx	鍵管理台帳
【AIT2011-01-F32-0001】鍵管理台帳 20160701.xlsx	AIT2011-01-F32-0001 鍵管理台帳 20160701.xlsx
【AIT2011-01-F32-0001】鍵管理台帳 20160701.xlsx	ISMS のセキュリティ目的
a_職員証発行担当マスター_rev5.xlsx	係・講座変換テーブル
a_職員証発行担当マスター_rev5.xlsx	a_職員証発行担当マスター_rev5.xlsx
a_職員証発行担当マスター_rev5.xlsx	職員証発行関連のマスター参照
受付履歴管理台帳_pg.accdb	受付履歴管理台帳
受付履歴管理台帳_pg.accdb	受付履歴管理台帳_pg
受付履歴管理台帳_pg.accdb	対応履歴管理

4.2.2 実験手順

本実験では, 以下の処理を検索クエリの数だけ繰り返し, 精度の平均を求める.

1. 正解セットからキーワードを選択する.
2. キーワードをクエリとして検索する.
3. 検索結果の上位 50 件を記録する.
4. 正解文書の順位を集計する.

4.2.3 実験条件

全文検索システムの関連度スコアには, ターム (形態素解析などで文書から抽出された語) ごとに Elasticsearch のデフォルトの関連度スコア TF-IDF(Term Frequency - Inverted Document Frequency) を用い, ベクトル空間法 (Vector Space Model) によるスコアを採用する.

$$tf_{t,d} = \sqrt{freq_{t,d}} \tag{5}$$

$$idf_t = 1 + \log\{\frac{N_D}{N_{d,t}+1}\} \tag{6}$$

ここで $freq_{t,d}$ は 文書 d 中に含まれるターム t の頻度である. N_D は全文書数, Nd,t はターム t を含む文書 d を表す.

$$norm_q = \frac{1}{\sqrt{\sum_{t\,in\,q} idf_t^2}} \tag{7}$$

$$coord_{q,d} = 1.5 \cdot R_{t,q} \tag{8}$$

$$norm_{t,d} = \frac{1}{\sqrt{N_{d,t}}} \tag{9}$$

$norm_q$ は, 異なるクエリ q でスコアを比較するための正規化項である. $coord_{q,d}$ は, 文書 d に クエリ q のタームが含まれる割合 $R_{t,q}$ を用いた調整項. $norm_{t,d}$ は, ターム t を含む文書 d の数 $N_{d,t}$ の平方根の逆数による正規化項である.

$$tfidf_{t,d} = tf_{t,d} \cdot idf_t^2 \cdot norm_{t,d} \tag{10}$$

$$score_{q,d} = norm_q \cdot coord_{q,d} \cdot \sum_{t\,in\,q} tfidf_{t,d} \tag{11}$$

```
{
 "from":0,
 "size":10,
 "query":{
  "query_string":{
   "query":"ISMS 2016 OR \"ISMS 2016\"",
   "fields":[
    "body","filename.tokenized^10"
   ]
  }
 }
}
```

図- 11: JSON 形式の検索クエリ

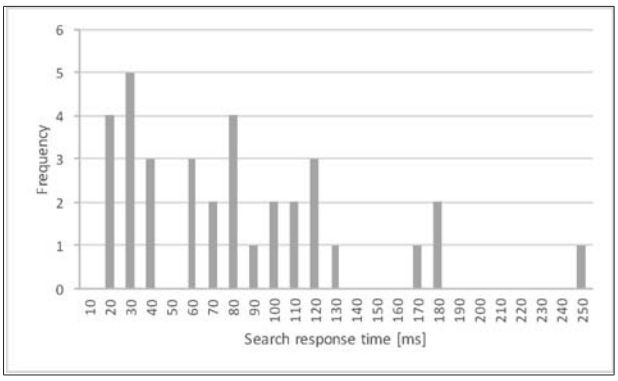

図- 12: レスポンス時間の分布

$itdif_{t,d}$ は, Elasticsearch のデフォルト設定による TF-IDF の重みである. $score_{q,d}$ は, TF-IDF を用いた内積に正規化項や調整項を組み合わせたスコアである.

検索クエリについては, ダブルクォートで囲んだクエリを追加することで, 完全一致するキーワードを含む文書がより高いスコアとなるようにした. 図-11 は, 実際に REST 形式で Elasticsearch へ送信される JSON の一部である.

検索クエリ「ISMS 2016」のとき, クエリタームは「ISMS」「2016」と「"ISMS 2016"」に分割し, OR 条件で連結する. 検索対象は, 文書の本文 (body) とファイルパス (filename.tokenized) である. ファイル名の検索を優先するために, filename.tokenized のスコアを 10 倍している.

4.3 評価方法

全文検索システム導入の妥当性について, 実験結果から速度と精度の観点で数値化し, 本センター業務の効率化がどの程度可能かについて評価する.

4.3.1 レスポンス評価

まず速度の観点で, 実運用に耐える性能であるかを判断するため, レスポンス時間 (Response time) を各検索クエリごとに集計し, その分布を分析する.

4.3.2 精度評価

次に精度の観点で, 実験結果から得られた各検索クエリに対する精度について, 次式に示すように適合率 (Precision), 再現率 (Recall), 精度 (Accuracy) を求め, 検索クエリ全 51 件についての平均値を算出する.

$$Precision = R/N \quad (12)$$

$$Recall = R/C \quad (13)$$

$$Accuracy = N_{found}/N_{query} \quad (14)$$

ここで R は検索された正解文書の数, N は検索結果の文書の数, C は全文書中の正解文書の数, N_{found} は正解文書が見つかった回数, N_{query} は検索クエリ数を表す.

4.3.3 妥当性評価

実験で得られたレスポンス時間の平均処理時間 β を 0.5 [s], 正解となる文書を探し当てるまでにまでに要する閲覧文書数 N_p を実験で得られた適合率の逆数で表し, 1 文書を閲覧する平均処理時間を α' [s] としたときの, ある 1 つのファイルの探索に要する平均探索時間 T_s [s] を次式で定義する.

$$N_p = Precision^{-1},$$
$$T_s = \alpha' N_p + \beta \frac{N_p}{10} \quad (15)$$

式 (15) で得られる平均探索時間と, 式 (4) で得られる平均探索時間を比較することで, 全文検索システムを導入することで得られるであろう効果とその妥当性を評価する.

5 実験結果と考察

本実験結果から, 本システムのレスポンス時間と検索精度を示し, ファイルサーバ上の文書の平均探索時間に関して, 全文検索システムを導入した場合と, ファイルサーバを直接的に探索する場合を比較し, 全文検索システムを導入した効果を見積もる.

5.1 レスポンス時間

レスポンス時間の分布を図-12 に示す. また, レスポンス時間の統計情報については, 表-9 に示した通りである. 統計情報を見ると, 75 パーセンタイルは 95.07 [ms] で, 非機能要件の平均検索時間 100 [ms] を下回っている.

表- 9: レスポンス時間の統計

最大	75%	平均	中央	25%	最小	標準偏差
248.51	95.07	69.41	51.74	28.61	15.78	54.36

* 検索リクエストから検索結果のレスポンス取得までの時間 [ms].

表- 10: 検索精度

Precision	Recall	Accuracy
0.1203	0.4872	0.6078

ただし, 最大は 248.51 [ms] であるため, 検索エンジンのパフォーマンスチューニングや, 検索クエリを工夫することで, 安定的に 100 [ms] 以下のレスポンス時間を保証する必要がある.

5.2 検索精度

検索精度について, 適合率 (Precision), 再現率 (Recall), 精度 (Accuracy) は, 表-10 に示した通りである.

適合率については, 式 (12) の N を, 実際にユーザが閲覧した文書数 (1 ページあたり 10 文書) を考慮して, 正解が得られるまでに閲覧したページ数の 10 倍 とした. また, 実際にユーザが閲覧するであろうページ数を 5 ページ (50 文書) とし, これを上限とした. 適合率 12.03 % は, スタッフが必要としている 1 つの文書を探し当てるまでに, 平均で 8.31 件の文書を閲覧しなければならないことを示す.

また, 再現率 48.72 % から, スタッフが必要としている全文書のうち, 約半数が見つけられない状況が伺える. 同様に, 正解の文書が得られた回数を全クエリの数で除した精度は 60.78 % であり, 全文検索エンジンを利用した場合に, 正解文書を見つけられる確率は約 6 割であった.

5.3 導入効果の考察

実験で得られたレスポンス時間から, 平均処理時間 β が 0.069 [s], 正解となる文書を探し当てるまでにまでに要する閲覧文書数 N_p が 8.31 件, 1 文書を閲覧する平均処理時間を α' が 0.5 [s] とすると, 式 (15) より, 全文検索システム利用したファイルの平均探索時間 T_s は 4.21 [s] となり, 全文検索システムを利用して正解文書が見つけられた場合は, 全文検索システムを利用しないファイルの平均探索時間 T_f の 37.88 [s] と比較して, 約 88.88 % のコスト削減が見込まれる.

正解文書が見つけられなかった場合は, ファイルサーバを直接的に探索する必要がある. このため, 全文検索システム利用した後に, ファイルサーバを直接的に探索する場合のファイルの平均探索時間を見積もる.

$$T_{max} = \alpha' N_{max} + \beta \frac{N_{max}}{10} \quad (16)$$

$$T_{one} = T_s \cdot Accuracy + (T_{max} + T_f) \cdot (1 - Accuracy) \quad (17)$$

$$T_{all} = T_s \cdot Recall + (T_{max} + T_f) \cdot (1 - Recall) \quad (18)$$

式 (16) から, 正解文書が見つからなかった場合に閲覧する最大文書数 N_{max} が 50 件のとき, 最大探索時間 T_{max} は 25.35 [s].

式 (17) から, 全文検索システムを利用して正解文書が見つかった場合の平均探索時間 T_s と, 正解文書が見つからなかった場合に, T_{max} に T_f を加算した時間とを, 1 つの正解文書が見つかる割合 $Accuracy$ で相加平均をとった平均探索時間 T_{one} は 27.36 [s] となり, 全文検索システムを利用しない平均探索時間 37.88 [s] より 10.52 [s] 短くなる.

式 (18) から, 全文検索システムを利用して正解文書が見つかった場合の平均探索時間 T_s と, 正解文書が見つからなかった場合に, T_{max} に T_f を加算した時間とを, 全正解文書が見つかる割合 $Recall$ で相加平均をとった平均探索時間 T_{all} は 34.47 [s] となり, 全文検索システムを利用しない平均探索時間 37.88 [s] との差は 3.41 [s] である.

この結果から, 全文検索システムを導入することにより, ファイルの平均探索時間は, 正解文書が見つけられた場合には, 約 88 % のコスト削減が見込まれる. 正解文書が見つからない場合も考慮すると, 1 つの正解文書を見つける場合, つまり, ナビゲーション型の検索に対しては, 27.78 % のコスト減が見込まれる. 一方, すべての正解文書を見つける場合, つまり, 情報収集型の検索に対しては, 8.99 % のコスト減にとどまるため, 検索意図によっては, 全文検索システムを導入する効果は限定的となる. ただし, アンケート結果から検索クエリの割合は, ナビゲーション型は 93.75 %, 情報集取型は 6.25 % なので, 全体としては 26.63 % のコスト減が予想される.

6 おわりに

本論文では, 徳島大学情報センターにおいて, ファイルサーバで管理している業務文書等の探索にかかる作業時間を短縮するために, 全文検索システムを導入することを検討した.

まず, 本センター業務の業務分析を行い, 全文検索システムを導入する際に求められる要件分析を行った. この機能要件及び非機能要件に従ってシステム設計を行い, アンケートを基に作成したテストデータを用いて,

実験を行った. 実験結果から, サーバ側からみた 1 クエリあたりのレスポンス時間は 69.41 [ms] であり, 同時アクセス数が 20 の場合でも, ユーザからみた平均レスポンス時間は 1.39 [s] となり, 実用上は問題ない.

ただし, 今後, 対象とするファイルサーバの規模がさらに大きくなった場合, 転置インデックス (Inverted Index) 方式を採用している Elasticsearch とはいえ, 必ずしも充分であるとはいえない. また, インデックスの更新時間に対しても, 大きな影響があるため, 検索エンジンやクローラを並列化するなどの工夫が必要となる.

検索精度に関しては, 精度 60.78%, 再現率 48.72%, 適合率 12.03% であり, 「最終的に探し当てる可能性は 6 割.」「平均 8 件の文書を閲覧する必要がある.」「探している文書のうち見つかるのは半数程度.」を意味する. 全文検索システムを利用して探している文書が見つからない場合は, 結局, ファイルサーバを直接的に探索する必要がある.

具体的には, 文書管理の関連業務の『定常的な業務負荷』について, 全体として 26.63 % のコスト減が期待されるため, 一定の有用性が認められる. ただし, すべての正解文書を見つける場合の平均探索時間 (T_{all}) は 34.47 [s] で, 全文検索システムを利用しない平均探索時間 (T_f) の 37.88 [s] と同程度のため, 改善の余地がある.

一方, 『突発的なインシデント対応』について, 全文検索システムを利用して求める文書が見つからない場合には, 最悪の場合, T_{max} に T_f を加えた 63.23 [s] を要し, 約 2 倍のコスト増となる. この問題を回避するためには, 再現率及び適合率を上げて, 全文検索システムを利用して求める文書が見つからない場合を減らす必要がある.

以上の課題の改善手段としては, 今回の実験では使用しなかったファイル種別などによる絞り込み機能や, 更新日時などによる並べ替え機能を組みわせることが考えられる. 実験方法も含めて今後の課題であるが, 実運用の開始後, さらに検索ログからテストセットの追加や機能の改善を行い, 継続的な評価実験を行うことで, PDCA(Plan-Do-Check-Action) サイクルによる業務改善につなげたい.

謝辞

本研究にあたり, 業務の合間を縫ってアンケートにご協力いただいた徳島大学情報センターのスタッフの皆さんに御礼申し上げます. 論文執筆にあたってご助言いただいた上田哲史先生, 佐野雅彦先生, 大平健司先生に深く感謝いたします.

参考文献

[1] 上田哲史, 佐野雅彦. 組織評価と ISMS. 情報処理学会研究報告インターネットと運用技術（IOT）, Vol. 2012, No. 41, pp. 1–6, Mar 2012.

[2] 佐野雅彦, 八木香奈枝, 上田哲史. 徳島大学情報センターにおける ISMS の効果. 学術情報処理研究 Journal for academic computing and networking, No. 18, pp. 90–98, Sep 2014.

[3] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the world-wide web. *Comput. Netw. ISDN Syst.*, Vol. 27, No. 6, pp. 1065–1073, Apr 1995.

[4] Saul Greenberg. *The Computer User As Toolsmith: The Use, Reuse, and Organization of Computer-based Tools*. Cambridge University Press, New York, NY, USA, 1993.

[5] John Mingers and Quentin L. Burrell. Modeling citation behavior in management science journals. *Inf. Process. Manage.*, Vol. 42, No. 6, pp. 1451–1464, Dec 2006.

[6] John R. Anderson and Lael J. Schooler. Reflections of the environment in memory. *Psychological Science*, Vol. 2, No. 6, pp. 396–408, 1991.

[7] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[8] Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. Stuff i've seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pp. 72–79, New York, NY, USA, 2003. ACM.

[9] elasticsearch. http://www.elastic.co, 2015. Online; accessed 30 May 2016.

[10] Chris Mattmann and Jukka Zitting. *Tika in Action*. Manning Publications Co., Greenwich, CT, USA, 2011.

[11] atilika. "kuromoji: japanese morphological analyzer.". http://www.atilika.org, 2012. Online; accessed 30 May 2016.