

# テキスト検索技術と 個人情報・機密情報検出

加藤 守\* William S. Yerazunis\*\*\*  
柴田秀哉\*  
中川 誠\*\*

*Text Search Technology and Personal Information and Confidential Information Detection*

*Mamoru Kato, Hideya Shibata, Makoto Nakagawa, William S. Yerazunis*

## 要 旨

近年、企業などの組織で作成・管理される電子文書やメールの量の増加に伴って、高速・高精度なテキスト検索技術の重要性が増している。テキスト検索技術は、文書やメールの検索システムのほか、情報漏洩(ろうえい)防止のための個人情報・機密情報検出など、様々な応用が可能である。

従来のテキスト検索技術では、実時間処理のための文字列照合の高速化と条件設定の容易化、及び蓄積処理時の高速で漏れない全文検索が課題であった。

この課題に対して、次のテキスト検索技術を開発した。

- (1) 高速文字列照合技術sDFA (size-reduced Deterministic Finite Automaton)

大規模検索式でも約1億文字/秒の高速照合(他社比3万~20万倍)の性能を実現

- (2) 学習型フィルタSSC (String-based Statistical Classifier)

サンプルの自動学習によって条件設定を容易化

- (3) テキストフィルタCID (Combined Information Detector)

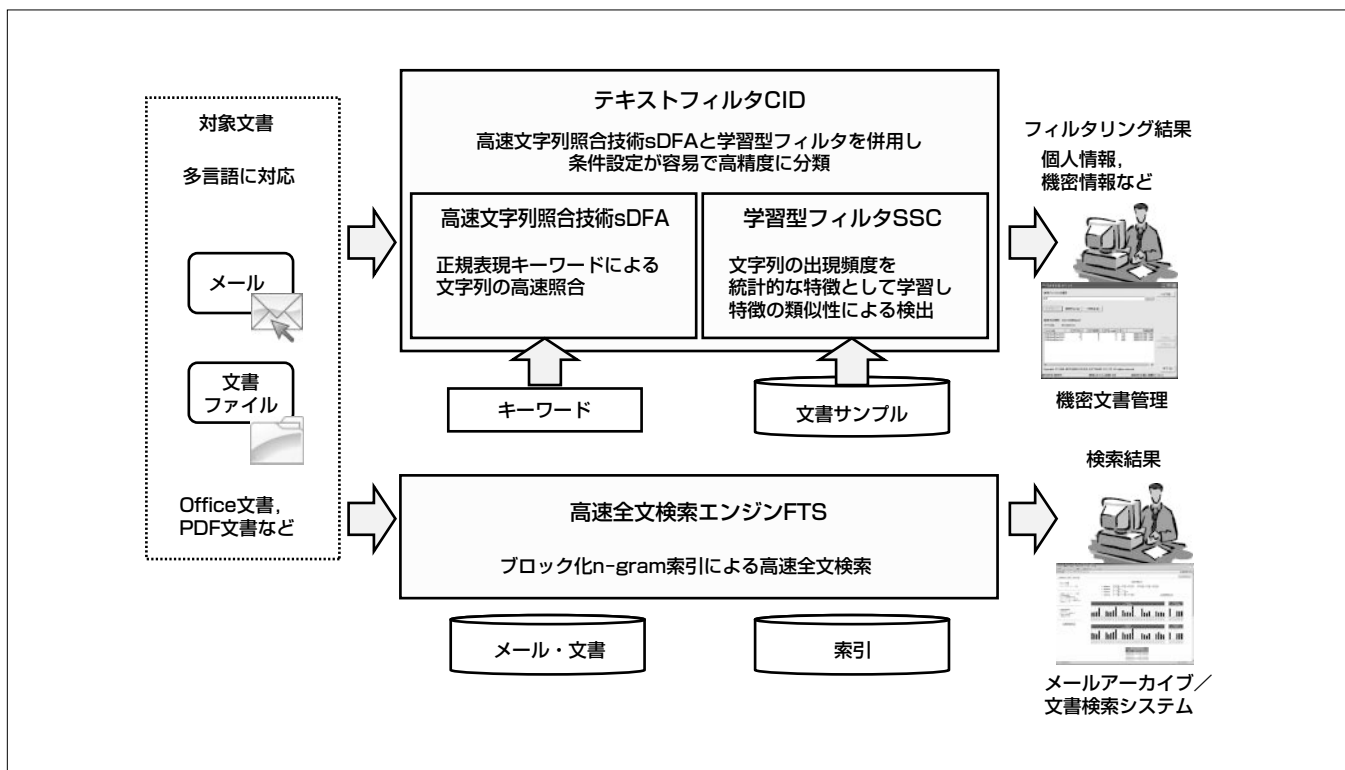
sDFAとSSCの併用によって条件設定が容易で高精度な分類を実現(再現率99.9%, 適合率98.7%)

- (4) 高速全文検索エンジンFTS (Full Text Search)

1テラバイトのメールを1秒で高速に検索

これらの技術の応用例として、パソコン内の文書ファイルの内容を検査し、個人情報を自動検出する三菱スペース・ソフトウェア㈱(MLS)の個人情報検出ツール“すみずみ君<sup>(注1)</sup>”と、機密情報を自動検出する“機密文書管理システム”を開発した。

(注1) すみずみ君は、三菱スペース・ソフトウェア㈱の登録商標である。



## テキスト検索技術の概念図

三菱電機のテキスト検索技術は、複雑な正規表現キーワードの高速照合を実現した高速文字列照合技術sDFA、サンプルを用いた学習によって条件設定不要とした学習型フィルタSSC、それらを併用し条件設定が容易で高精度の分類を実現したテキストフィルタCID、蓄積データの高速全文検索エンジンFTSの各技術からなる。多言語検索に対応し、個人情報・機密情報の検出、メールアーカイブや文書検索システムなどへ応用可能である。

## 1. ま え が き

近年、企業などの組織で作成・管理される電子文書やメールの量の増加に伴って、人手による文書管理が困難になっており、高速・高精度なテキスト検索技術の重要性が増している。テキスト検索技術は、文書やメールの検索システムのほか、情報漏洩防止のための個人情報・機密情報検出など、様々な応用が可能である。

本稿では、従来のテキスト検索の課題と、三菱電機で開発した高速・高精度テキスト検索技術、及び個人情報と機密情報の検出への応用例について述べる。

## 2. 従来のテキスト検索の課題

テキスト検索技術は、大きく実時間処理と蓄積処理の二通りに分けられる。実時間処理は入力文書に対して直接照合などの処理を行って結果を出力する方法で、正規表現<sup>(注2)</sup>による文字列照合を行う方法などがある。また、蓄積処理は入力文書をいったんストレージに蓄積し、蓄積された文書の検索処理を行う方法で、文字列照合のほか、蓄積時に検索用の索引を生成し、その索引を用いて全文検索を行う方法がある。次にそれらの課題について述べる。

### 2.1 実時間処理における課題

#### 2.1.1 高 速 化

従来、正規表現による文字列照合方式として、状態遷移機械のNFA(Non-Deterministic Finite Automaton)方式とDFA(Deterministic Finite Automaton)方式がある。NFA方式は、検出条件規模の増大に従って照合速度性能が大幅に低下する。DFA方式では、検出条件規模の増大に従って、状態数や状態遷移数が指数的に増加し、照合不可能になる。これらの方式は、欧米で文字種の少ない文字コードを対象として発達してきたため、文字種が非常に多いマルチバイト文字コードを含む場合に、高速な照合が困難である。例えば個人情報検出を行う場合、“頻出人名と住所パターンの組合せ”などの複雑な条件を高速に照合する必要がある、高速化が課題である。

#### 2.1.2 検出条件作成の容易化

文字列照合方式による検出を行う場合、検出対象によっては検出条件の作成が困難な場合がある。例えば、個人情報に限らない一般の機密情報を文字列照合方式を用いて検出する場合、適用先ごとに細かな検出条件設定を人手で行わなければならない。なぜなら、一般の機密情報は適用先の組織(企業や部署など)ごとに検出対象とする内容が異なるという性質を持つためである。したがって、適用先における検出対象文書の内容に精通していなければ、高精度な検出条件を作成するのは困難である。また、キーワードが少なければ検出漏れ件数が増大し、キーワードを多く設定

(注2) 文字列のパターンを表現する表記法

しすぎると過剰検出件数が増大する。結果として、文字列照合方式のみを用いる場合、精度の高い検出条件を設定するために試行錯誤的な調整が必要となり、検出条件設定が困難であるという課題がある。

### 2.2 蓄積処理における課題

企業などの組織内のテキスト検索システムでは、大量に蓄積された文書の中から漏れのない全文検索を高速に行うことが必要である。漏れのない全文検索方式としてn-gram索引方式があるが、文書中のすべてのn文字の組合せを索引化するために、文字種の多い文字コードでは索引情報が大きくなり、検索性能が索引のI/O(Input/Output)能力で制限される。

## 3. 高速・高精度テキスト検索技術

ここでは、三菱電機の高速・高精度テキスト検索技術について述べる。これらはいずれも言語に依存しない処理方式を採用しており、多言語検索を可能としている。

### 3.1 高速文字列照合技術sDFA

三菱電機のsDFA方式<sup>(1)</sup>では、状態遷移表の冗長性を利用し複数の状態遷移を単一の遷移に置き換えるなどによって、DFA方式の課題であった状態遷移数を1/7,000に削減し、検索時のメモリ消費量を大幅に削減した。その結果、大規模検索式でも約1億文字/秒の高速照合(他社比3万~20万倍)の性能を実現した(図1)。

このsDFA方式によって、複雑な条件の高速照合を必要とする個人情報検出が実現可能となった。

### 3.2 学習型フィルタSSC

学習型フィルタは、あらかじめ該当データと非該当データに分けられたサンプルデータから、出現頻度の高い文字列を統計的特徴として学習し、新たな入力データと統計的特徴を比較することで、該当、非該当いずれかに分類を行うフィルタである。迷惑メール(スパム)を検出・排除するスパムフィルタで実用化され、広く用いられている。

三菱電機は、SVM(Support Vector Machine)方式<sup>(2)</sup>、OSB(Orthogonal Sparse Bigram Bayes)方式<sup>(2)</sup>、Bit Entropy方式<sup>(2)</sup>の3方式を採用した学習型フィルタSSCを開発した。特性の異なる3方式を併用することで検出精度

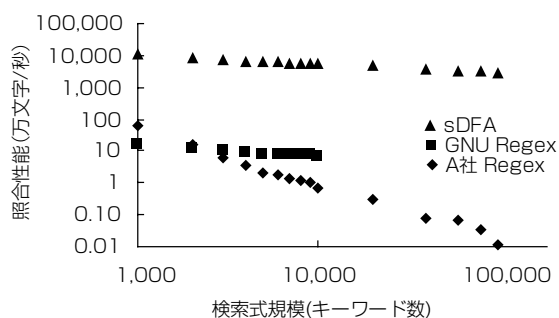


図1. 文字列照合速度比較

向上を図っている。学習段階では3方式の学習型フィルタすべてで同じサンプルデータを学習させる。検出段階では3方式の学習型フィルタすべてに同じ入力データを与え、検出結果のうち一つでも該当と判定されれば該当とするなどの方法によって結果を併合する。

学習型フィルタSSCは、人手による検出条件の細かな設定・調整が不要であるため、機密情報のように人手での検出条件作成が難しい場合に有効である。

### 3.3 テキストフィルタCID

高速文字列照合技術sDFAによる正規表現フィルタと学習型フィルタSSCを併用することで、互いの長所を生かして補完する正規表現・学習型フィルタ併用方式を用いたテキストフィルタCIDを開発した(図2)。

併用方式では分類を各フィルタで個別に行い、用途に応じて結果を併合し出力する。機密情報検出の場合、検出漏れを低減させるため、少なくとも一方のフィルタが機密と判定した文書を総合的に機密と判定する。

併用方式の特長は、簡易な条件設定のみで高精度な分類が実現できることである。正規表現フィルタは条件設定に細かな調整を要し、高精度な条件作成が困難な場合がある。併用方式では、学習型フィルタの併用によってこの課題を解決している。例えば機密情報検出では、適用先組織の規則などで定められた定型的な機密文書を正規表現フィルタで検出し、それ以外の機密文書は学習型フィルタで補完するという方針で条件を設定する。これによって設定キーワードのリストアップが容易となり、検出対象文書の内容に精通していない第三者による条件作成が可能となる(表1)。

また併用方式では、少数のサンプルデータのみで高精度の分類を実現できる。学習が不十分なときも、正規表現フィルタの併用によって大幅な精度低下は起こらず、安定した分類精度が実現される。

機密情報検出に正規表現・学習型フィルタ併用方式フィ

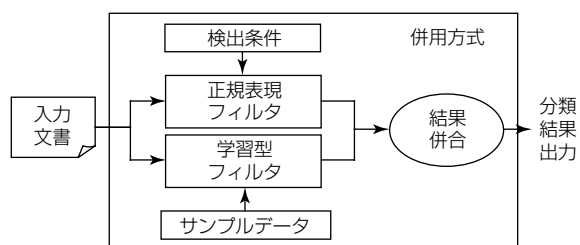


図2. 正規表現・学習型フィルタ併用方式の構成

表1. 機密情報検出における正規表現フィルタ条件設定例

|         |                                          |
|---------|------------------------------------------|
| 機密等級ラベル | 社外秘、極秘、人事秘など機密文書に附することが規定されている語          |
| 定型文書名   | システム開発計画書、新規事業提案書など、文書形式や機密等級が規定された文書名   |
| 組織名略称   | 品質管理部→品管部、システム開発第1課→シス開1のように組織内でのみ利用する略称 |

ルタを適用した結果、機密メール12,575件、非機密メール1,701件からなるデータセットに対して、再現率99.9%、適合率98.7%という結果を得た。このことから、正規表現・学習型フィルタ併用方式が有効であることが分かる(図3)。

### 3.4 高速全文検索エンジンFTS

高速全文検索エンジンFTSは、独自のブロック化n-gram索引方式によって、大量に蓄積された文書から漏れなく高速に全文検索を行うことができる。ブロック化n-gram索引方式は、n-gram索引のデータ配置にブロック化トランスポーズドデータ配置<sup>(5)</sup>を適用することで、ディスクからの読み出しを最小限に抑え、データ駆動型並列処理<sup>(5)</sup>などの高速化技術を適用した。

FTSを応用したメールアーカイブシステムでは、1テラバイトのメールを約1秒で高速検索することが可能<sup>(6)</sup>である(図4)。

## 4. 応用事例

### 4.1 個人情報検出ツール“すみずみ君”

高速文字列照合技術sDFAの応用事例として、MSSの個人情報検出ツール“すみずみ君”について述べる。“すみずみ君”は、Windows<sup>(注3)</sup>パソコン内に保存されているファイルの内容を検査し、名前+住所・TEL・E-mailの組合せによる個人情報に該当すると思われるファイルを自動検出し、ファイル名を一覧表示するソフトウェアである。表示したファイルの内容の確認や削除なども行うことが可能である(図5)。

また、“すみずみ君”を統合制御する“すみずみ君SV”を使用することによって、検査結果をサーバに収集することが可能である。

(注3) Windows, Excelは、Microsoft Corp.の登録商標である。

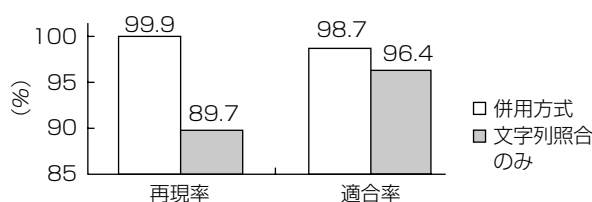


図3. 正規表現・学習型フィルタ併用方式の精度

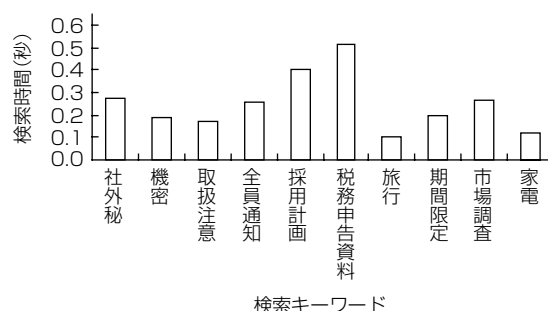


図4. 1テラバイトのメールの検索時間

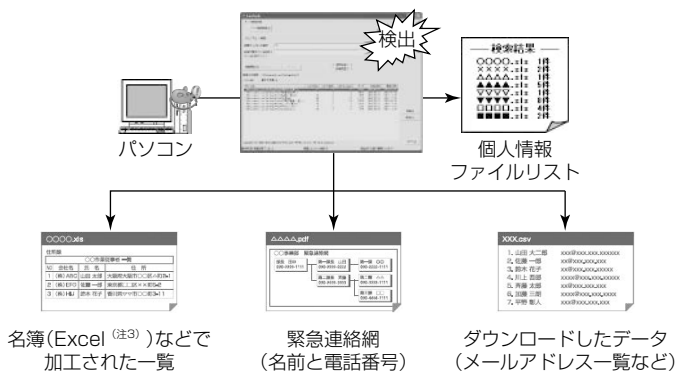


図5.「すみずみ君」機能概要

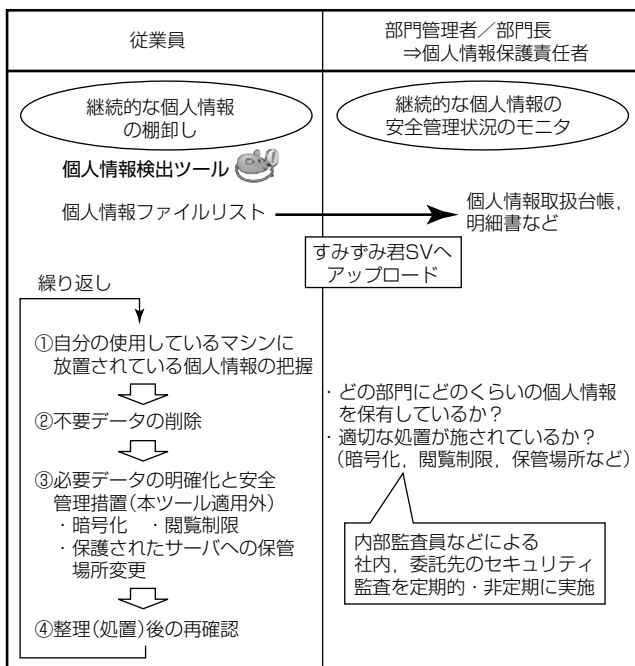


図6. 個人情報の安全管理運用例

“すみずみ君”及び“すみずみ君SV”を用いることによって、図6に示すように個人情報管理を継続的に実施できるようになる。

#### 4.2 機密文書管理システム

テキストフィルタCIDの応用事例として、機密文書管理システムについて述べる。機密文書管理システムは、パソコン内のファイルの内容を検査して機密情報と思われるファイルを自動検出し、適正に管理されていない機密文書を検出・是正するためのシステムである(図7)。企業などの組織で、共用ファイルサーバや各個人が使用するパソコン内の機密文書管理の徹底による情報漏洩防止を実現する。

管理サーバでは、セキュリティ管理者が検出条件作成及び検査結果の確認を行う。検出条件作成では、人手作成キーワードを作成し、機密文書/非機密文書のサンプルを準備して自動学習させる。適用先の部門ごとに異なる機密情報に対応し、部門ごとに検出条件を作成することが可能である。管理サーバでの検査結果確認によって、各検査対象

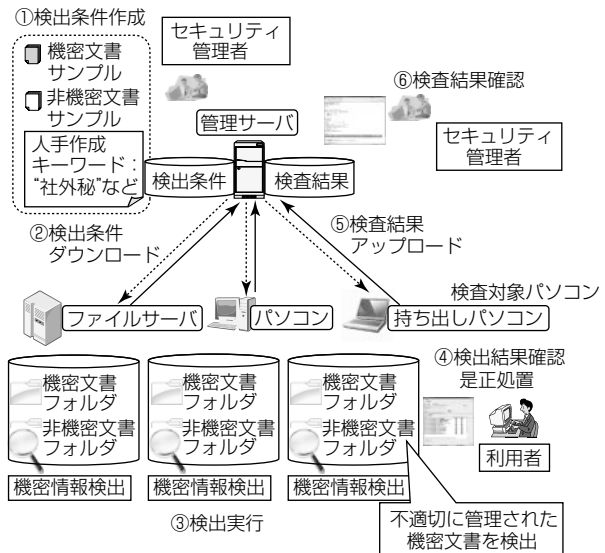


図7. 機密文書管理システム

パソコンの機密文書管理状況を一元管理可能となる。

検査対象パソコンでは、利用者が検出された機密文書の一覧を確認し、本来機密文書を置いてはいけないフォルダに不適切に置かれた機密文書の把握、削除などの処置を行う。

#### 5. む す び

本稿では、三菱電機の高速・高精度を特長とするテキスト検索技術について述べ、それらの応用事例としてMSSの個人情報検出ツール“すみずみ君”と、機密文書管理システムについて述べた。今後は個人情報検出、機密情報検出といったセキュリティ関連の応用に加え、他のテキスト検索・分類への適用拡大を図っていく。

#### 参 考 文 献

- (1) 中村隆顕, ほか: 大規模正規表現の高速照合方式, 情報処理学会全国大会講演論文集(3), 235~236 (2005)
- (2) Kato, M., et al.: Three Non-Bayesian Methods of Spam Filtration: CRM114 at TREC 2007, NIST Text REtrieval Conference (2007)
- (3) 加藤 守, ほか: 正規表現・学習型フィルタ併用方式による機密情報検出の提案, 第8回情報科学技術フォーラム講演論文集(2), 157~158 (2009)
- (4) 柴田秀哉, ほか: 正規表現・学習型フィルタ併用方式による機密情報検出の評価, 第8回情報科学技術フォーラム講演論文集(2), 159~160 (2009)
- (5) 郡 光則, ほか: 高性能並列情報検索技術, 三菱電機技報, 83, No.12, 705~708 (2009)
- (6) 中村隆顕, ほか: 大規模並列全文検索エンジンにおける多国語検索対応索引方式, 第8回情報科学技術フォーラム講演論文集(2), 115~116 (2009)