個人情報ファイル検出ツール「すみずみ君」の紹介

Sumizumikun, personal information file detection tool

松下 英男*
Hideo Matsushita

すみずみ君は、個人情報管理、漏洩対策として、クライアントPC／共有サーバ内のすみずみまで個人情報・機密情報に該当するファイルを「簡単」「高速」「高精度」に検出する個人情報ファイル検出ツールである。ファイル名を検出するのみではなく、そのファイル内のどこに個人情報・機密情報があるかも簡単にチェックできる。

"Sumizumikun" is a personal information file detection tool which detects the file which corresponds to personal information and confidential information in client PC / shared server to "easy", a "high speed", and "high precision" as personal information management and a measure against disclosure. It not only detects a file name, but it can be confirmed easily where［in the file］personal information and confidential information are.

1. まえがき

近年、個人情報漏洩インシデントが多数発生しており、2013年も年間で約1300件、漏洩人数は約930万人、想定損害賠償総額2020億円にも上っている。(1) 2014年も大規模な個人情報漏洩事件が発生しており、個人情報の適切な管理がより一層求められている。

個人情報漏洩の原因は、誤操作、管理ミス、紛失・置忘れといったヒューマンエラーが大部分を占めており(1)、個人情報漏洩を防止するためにはヒューマンエラーを発生させない対策、つまり可能な限り人手に頼らない対策が必要である。

2. 「すみずみ君」とは

当社製品である、個人情報ファイル検出ツール「すみずみ君」は、クライアントPC／共有サーバ内に存在するファイルから、個人情報・機密情報に該当する情報を含むファイルを検出するツールで、「個人情報の保護に関する法律（個人情報保護法）」への対応として、2005年１月にリリースした。

個人情報・機密情報であると定義する文字列パターンは「辞書」として登録しておき、ファイルに含まれるテキストデータに該当パターンの文字列が含まれる場合に当該ファイルを個人情報・機密情報ファイルとして検出する。事前に個人情報･機密情報とする条件を設定するため、利用者に依存せず、同一条件で個人情報・機密情報ファイルを検出することができる。また、クライアントPC／共有サーバ内の全てのファイルを自動的に検査するため、検出漏れも防止できる。

辞書は、住所、電話番号、E-mailアドレス等の標準で搭載されているもの（標準辞書）に加え、キーワードや正規表現を使用して利用者が自由に追加できるようになっており、様々な業種で扱われる多様なデータ内容に柔軟に対応できる。

3. 「すみずみ君」の特長

本章では、すみずみ君が個人情報ファイル検出ツールとして、競合製品に対して優位性を有している特長について述べる。

3.1 検出文字列簡単チェック機能

クライアントPC／共有サーバを検査し、そこに含まれる個人情報・機密情報ファイルが検出された場合、そのファイルを破棄してよいのか、適切な保管場所に移動する必要があるのかを判断するために、ファイル内容を確認する必要がある場合がある。しかし、個人情報・機密情報に該当する文字列は、単純なキーワードではなく、特定のパターンを持った文字列（住所であれば、都道府県名に続いて市区町村名、番地が続く等）となっていることが多いため、ファイル内のどこに個人情報・機密情報に該当する文字列が含まれているか探し出すことは困難となる。

＊鎌倉事業部　ICTソリューション開発部

一般的な個人情報検出ツールでは、検出したファイルに各辞書に該当する文字列が何件含まれているかが示されるが（これは「すみずみ君」も同様である）、この情報ではファイルに含まれる個人情報・機密情報に該当する文字列を探しだすことはできない。

「すみずみ君」には、この課題を解決するための機能として「検出文字列簡単チェック機能」が備わっている。この機能は、辞書に該当した文字列の件数ではなく、文字列そのものをGUIで確認することができる。これにより、利用者は検出ファイル内に含まれる個人情報・機密情報に該当する文字列を容易に特定することができる。

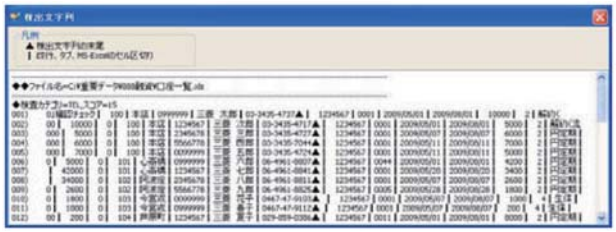

図１　個人情報・機密情報に該当した文字列の表示

3.2　高性能データストリーム検索エンジン「SP-Filter」

すみずみ君は、三菱電機株式会社　情報技術総合研究所が開発した高性能データストリーム検索エンジン「SP-Filter」を搭載している。

一般的な直接照合方式検索では、検索条件が大規模または複雑になると処理性能が劣化する。これに対して、SP-Filterは、検索条件の規模に依存せずに、高速文字列照合が可能なアルゴリズムを搭載した検索エンジンであり、広く使用されている正規表現文字列照合ライブラリと比較し、数万キーワード規模で3万～20万倍の高速な検索を実現している。

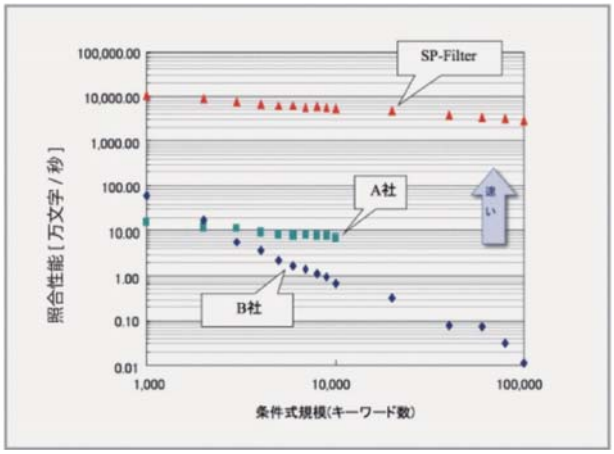

図２　SP-Filter照合性能

SP-Filterを搭載していることにより、すみずみ君では辞書に多数のキーワードや複雑な正規表現を使用しても検査速度が低下することなく検査が行える。前節でも述べた通り、個人情報・機密情報に該当する文字列は単純なキーワードであることは少ないため、辞書は複雑な正規表現で構成されることが多い。複雑な正規表現を使用しても検査速度が低下しないことは個人情報ファイル検出ツールとして、大きな優位点である。

3.3　高精度な辞書

既に述べている通り、すみずみ君の辞書は日本語正規表現を使用して作成することができる。さらに表記ゆれ（大文字／小文字、全角／半角、ひらがな／カタカナの拗音、促音、長音記号やハイフン等の同一視）を考慮して辞書が作成できる。

すみずみ君の標準辞書は、上記特長を利用した辞書となっており、ファイルに含まれる個人情報・機密情報を高精度で検出することができる。標準辞書の特長を表１に示す。

表１　標準辞書特長

No.	辞書	特長
1	電話番号	単純な数列ではなく、電話番号体系に則った固定電話、携帯電話を検出
2	住所	単純な市区町村名ではなく、都道府県＋市区町村＋番地というパターンを検出
3	E-mail アドレス	RFC に準拠し、ドメイン名で誤検出を低減
4	名字	代表的な日本人名字を基本に、後続文字で誤検出を低減
5	名字(カナ)	名字辞書に含まれる名字のカタカナ表記、全半角に対応
6	名字(英字)	名字辞書に含まれる名字の英字(ローマ字)、大文字、小文字に対応
7	生年月日	元号/西暦どちらの記載にも対応
8	クレジットカード番号	Luhn アルゴリズムでチェックデジットを照合し、検出精度向上
9	口座番号	都市銀行、地方銀行、第二地方銀行、旧長期信用銀行、ネット専業銀行等に対応

4. むすび

本稿では、個人情報ファイル検出ツール「すみずみ君」について述べ、特に特徴的な機能について説明した。

すみずみ君には、本稿で紹介した以外にも様々な機能があり、個人情報・機密情報ファイルの管理・運用を強力にサポートする製品となっている。すみずみ君の詳細については、下記窓口に問い合わせ頂きたい。

営業本部　ソリューション営業部　第一課
〒105-6132　東京都港区浜松町二丁目4番1号
世界貿易センタービル32階
TEL：03-3435-4737　FAX：03-3435-4745

参考文献

（1）2013年 情報セキュリティインシデントに関する調査報告 ～個人情報漏えい編～, NPO 日本ネットワークセキュリティ協会（2014/6/10）

執筆者紹介

松下　英男
2003年入社。
2012年からすみずみ君の開発に従事。