D-018

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

Proposal of a Confidential Information Detection Method with Regular Expression and Statistical Filtering

加藤 守† 柴田 秀哉† 郡 光則†
Mamoru Kato Hideya Shibata Mitsunori Kori

1．はじめに

近年、情報量の増大が著しい中、文書の機密性の判断を個人に委ねると故意や誤りによる情報漏洩を防止できないという背景があり、電子データに機密情報が含まれるか否かを自動的に判別する機密情報検出の技術が求められている。企業などにおいて機密文書の不正な管理や持ち出しを防止するシステムや、機密メールの不正な外部送信を検出するシステムなど、情報漏洩防止のための応用システムがすでに実用化されてきている。

しかし、現在広く用いられている文字列照合方式では、高い検出精度を実現するための検出条件作成が困難という課題がある。

本研究では、正規表現フィルタによる文字列照合と学習型フィルタを併用する機密情報検出方式を提案する。機密／非機密の文書サンプルを用いた学習により検出精度を向上させることができ、検出条件の作成が容易で高精度な機密情報検出を実現した。

なお、提案方式の評価結果は[1]で報告する。

2．機密情報検出方式

機密情報検出方式には大きくは次の 3 方式がある。

1. 文字列照合方式
2. フィンガープリント方式
3. 学習型フィルタ方式

以下、これらについて簡単に説明し、現状の課題を述べる。

2.1 文字列照合方式

文字列照合方式は、機密情報に含まれるキーワードや文字列パターンを予め検出条件として設定し、その検出条件に合致する文書を機密情報として検出する方式である。例えば、正規表現の高速照合方式[2]を用い、人名リストや住所・電話番号・メールアドレスのパターンなどを検出条件とすることで個人情報を効率的に検出する技術が実用化されており、個人情報のような固定的なパターンの検出には有効である。また、キーワードに設定した語が含まれるファイルを 100%検出することができ、検出された理由も明確であるという利点がある。

一方、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、適用先毎に細かな検出条件設定を人手で行わなければならない。何故なら、一般の機密情報は適用先の組織（企業や部署など）毎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。従って、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。
- 検出条件の設定方法論が確立されにくい。

という課題がある。

2.2 フィンガープリント方式

フィンガープリント方式とは、登録文書データをハッシュ値に変換して得られる証明データを記録しておき、検出対象文書データから得られる証明データと突き合わせて照合する方式である[3]。登録文書データから得られる証明データが検出条件に対応する。

フィンガープリント方式では、予め登録しておいた文書そのもの、あるいは登録文書を一部改変した部分一致文書のみが機密として検出される。従って、限られた文書のみを扱う場合に有効であるが、扱う文書の範囲が広いときや、多数の未知文書を扱う場合などには不向きである。

2.3 学習型フィルタ方式

学習型フィルタとは、事前にクラス別の訓練用データを学習させることにより、未知文書のクラスを判定するフィルタである。機密情報検出の場合、事前に機密／非機密の別に訓練用データを学習させ、これにより未知文書の機密／非機密を判定する。学習型フィルタはスパムフィルタ[4]として利用されるなど実用化が進んでいるが、日本語の機密情報検出における有効性については知られていない。

学習型フィルタを利用した場合、学習により検出条件が自動生成されるため、人手による検出条件の細かな設定・調整が不要となり、文字列照合方式の課題であった検出条件設定の困難さが解消される。また、学習型フィルタでは、入力文書が機密／非機密のどちらに近いかという判定を行うため、フィンガープリント方式より広い範囲の文書を検出することが可能となる。

一方、学習型フィルタには、

- 高精度な検出条件を生成するために大量の訓練用データを事前に用意する必要がある。そのため、環境の変化に弱く、学習結果が検出条件に反映されるまでに時間がかかる
- 100%の検出を保証することができない、検出された理由を明示することが難しいという不安定要素がある。

という課題がある。

† 三菱電機株式会社 情報技術総合研究所
Information Technology R&D Center,
Mitsubishi Electric Corporation

3. 正規表現・学習型フィルタ併用方式

3.1 併用方式

本提案の正規表現・学習型フィルタ併用方式による機密情報検出フィルタの構成を図 1 に示す。

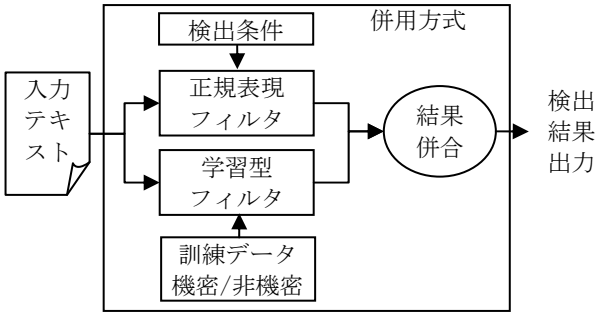

図 1 ：正規表現・学習型フィルタ併用方式の構成

この方式では、検出対象である 1 つの入力データ（テキスト）に対して、正規表現フィルタと学習型フィルタの両方により判定を行い、一方でも機密と判定されれば機密とする。理由は、以下の 2 点である。

1. 機密情報検出においては検出漏れ件数を減少させることが最重要の要件となる。過剰検出されたデータは人手による確認が可能であるが、検出漏れは確認ができないためである。
2. 検出したい機密キーワードを 100%確実に検出し、検出理由が明確であるという正規表現フィルタの利点を生かすことができる。従って、学習型フィルタの学習が充分に進んだ状態では学習型フィルタのみを使用するように動的に選択することで検出精度が高くなる可能性はあるが、その方式はとらない。

本方式の下では、個々のフィルタでの過剰検出を抑える必要がある。併用により単体のフィルタを個別に使用した場合に比べて検出漏れを減少させることができるが、過剰検出は単純に増加するためである。学習型フィルタでは、学習が進むにつれ検出漏れ・過剰検出ともに減少するが、正規表現フィルタにて最初に設定した検出条件が全体の検出精度に影響するため、過剰検出の少ない検出条件設定が課題となる。

3.2 正規表現フィルタの検出条件設定方法

正規表現フィルタで過剰検出を抑えるための検出条件の設定方法論を以下に示す。

基本方針として、定型的な機密文書を限定的に検出するための条件を設定する。定型的とは、適用先の組織の文書管理規則あるいはセキュリティポリシー等に従ったフォーマットや表紙情報などを有することをいう。例えば、機密等級ラベル、定型文書名、組織名の略称など、機密文書に使用される語、組織内でのみ使用される語を検出条件として設定することで、定型的な機密文書を検出できる。以下に検出条件の具体例を挙げる。

1. 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キーワード。
2. 定型文書名
   システム開発計画書、新規事業提案書、経営計画書、等の文書形式や機密等級が規定された文書名。
3. 組織名略称
   品質管理部→品管部、システム開発第 1 課→システム開 1 、などのように組織内でのみ利用する略称が規定されている場合。

このような検出条件であれば、組織内で使用される語に限られ、一般の文書を過剰検出する可能性が低下する。一般的な語が含まれる場合には正規表現での限定を加えることにより過剰検出を減らすことが可能である。

例）「秘」の 1 文字を機密等級ラベルとして使用する場合、「秘」を含む熟語を検出対象から除外する。

また、検出対象文書の内容に精通していなくともキーワードのリストアップが容易であり、作成された検出条件は作成者に大きく依存せず、常に同程度の検出精度となることが期待できる。

定型的でない機密文書に関しては、学習型フィルタにより文書の内容に基づく検出を行なうことで、補完が可能である。

3.3 各フィルタの構成

本提案では、学習型フィルタとして、スパムフィルタにて用いられている方式[4]の中から、Support Vector Machine (SVM)方式、Orthogonal Sparse Bigram Bayes (OSB)方式、Bit Entropy方式の 3 方式を利用する。これらを個別に利用することも可能であるが、検出精度向上のために 3 種類の方式を併用する。学習段階では 3 種類の学習型フィルタ全てにて同じ訓練データを学習させる。検出段階では、3 種類の学習型フィルタ全てに同じ入力データを与え、検出結果のうち一つでも機密と判定されれば機密とする。これは、先に述べたように検出漏れを減少させることを優先したためである。

また、正規表現フィルタとして、正規表現の高速照合方式[2]で述べられている方式を利用する。

4．おわりに

本研究では新たな機密情報検出方式として、文字列照合方式と学習型フィルタを併用し特性を補完することで、検出条件の作成が容易で高精度な機密情報検出を実現する正規表現・学習型フィルタ併用方式を提案した。学習型フィルタにより、文字列照合方式の課題であった検出条件設定の困難さが解消される。本報告では検出条件設定方法論と設定の一例を示した。逆に、環境変化に弱い、検出結果に不安定要素があるという学習型フィルタの課題は文字列照合方式を併用することで補うことができる。

本研究で提案した機密情報検出方式の有効性評価は[1]で報告する。

参考文献

[1] 柴田 他, 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, 5D-2, 2009.
[2] 中村 他, 大規模正規表現の高速照合方式, 第 67 回情報処理学会全国大会 講演論文集(3)235-236, 2005.
[3] お手軽ツールから DLP まで 今どきの情報漏えい対策, 日経 NETWORK 2008.08, pp.57-69, 2008
[4] M. Kato, et al., Three Non-Bayesian Methods of Spam Filtration: CRM114 at TREC 2007, Proc. TREC 2007, 2007.