Vol.2010-CSEC-49 No.8
2010/5/11

機密情報の漏洩を防ぐための文書再利用検出技術

三品拓也[†1]　吉濱佐知子[†1]　浦本直彦[†1]

情報通信技術の発達と競争環境から，複数の組織がお互いの機密情報を共有しつつ協力して作業を行う機会が増えた．このような環境では，誤って機密情報を漏洩した場合の不利益が大きく，DLP (Data Leakage Prevention) 技術によって情報漏洩を防止する必要性が高まっている．本研究では特に不適切な再利用に起因する情報漏洩を検出するため，テキストの類似度を用いて再利用を検出する方法と，ベクタ画像の図形オブジェクトの地理的配置をグラフ表現してグラフマイニングによって類似度を比較する手法を提案する．提案方法を実装し，ビットマップ画像類似度を用いた場合と再利用検出の精度を比較したところ，テキスト類似度は大きなデータセットに対して高い性能を発揮し，グラフマイニングは部分的再利用をより正確に検出できることがわかった．

Document Reuse Detection Technologies for Data Leakage Prevention

Takuya Mishina ,[†1] Sachiko Yoshihama [†1]
and Naohiko Uramoto [†1]

The growth of the Internet and information technologies enables people in different organizations to share their confidential information in the form of digital documents among the alliances or the restructuring project teams. However, the sharing, especially inappropriate reuse of confidential documents, can cause unintentional and fatal data leakage. DLP (Data Leakage Prevention) is a generic term for IT products and services which detect such data leakage by content analysis and perform appropriate operations according to the result of the content analysis. In this paper we propose a document reuse detection technique utilizing both text similarity and vector image similarity. For the vector image similarity detection we adapt the graph-mining technique to the similarity detection on office documents.

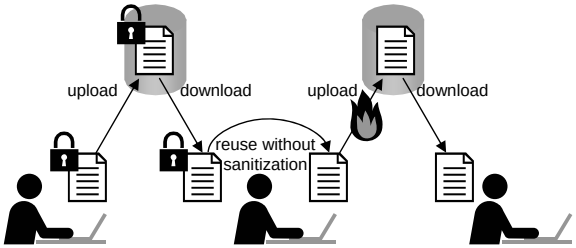

図 1　不適切な情報の再利用によって引き起こされる情報漏洩の例

1. はじめに

近年の情報通信技術発達に伴い，企業の命運を左右するような重要な情報が，異なる組織に跨って利用されるようになった．例えば，多国籍の企業連合を作ってひとつの財・サービスを共同で開発したり，企業内部の作業（機構変革・IT 導入など）をコンサルティングファームと共同で行う場合，共同作業を効率よく実行するために，電子メールの添付書類といったアドホックな手段のほか，Wiki や blog, CMS (content management system) などのアプリケーションによって文書が複数組織間で共有される．

複数の組織が協力することで得られる利益は大きいが，それは危険と隣り合わせである．誤って自社の社外秘情報を共有してしまえば直接的損害を被るし，他社の秘密を異なる会社に誤開示すれば，自社の信用失墜や訴訟を招く．例えば**図 1** のように，2 つの共同作業環境に属するユーザが，適切なデータ削除を行わずに文書を再利用した場合，ある共同作業体の秘密が別の共同作業体に漏洩する．本研究は，このような情報共有環境において発生する，所有者が意図していない情報漏洩を防ぐことを目的とし，対象データとしてはオフィス文書形式，中でもプレゼンテーション形式に着目する．また，プレゼンテーション文書を再利用するときはページ単位で再利用されることが多いと考えられるので，ページ単位での再利用関係検出を試みる．

DLP (data leakage prevention) は情報漏洩を防ぐための IT 製品・技術の総称であり，情報の内容を解析して不適切な情報の流れを発見し，情報漏洩とならないように適切な処

†1 日本アイ・ビー・エム（株）東京基礎研究所
IBM Research - Tokyo, IBM Japan, Ltd.

置（流れの完全遮断・危険部分の削除・ユーザへの警告）を行う．内容解析の方法としては，特定のパターンにマッチするかどうかを検査する方法や，機密扱いされているデータとの一致度を比較する方法が用いられる．

本研究では，類似性の中でも特にオフィス文書の再利用によって生じる類似性を検出することを目的とする．オフィス文書にはテキスト情報・画像情報が共に含まれているため，本研究ではまずテキストの類似度を用いて再利用を検出する方法を評価し，続いて図形オブジェクトの地理的配置をグラフ表現して，テキストを含むオブジェクトの特徴量に基づくグラフマイニングを行って類似度を比較する手法を提案する．提案方法を実装し，ビットマップ画像類似度を用いた場合と再利用検出の精度を比較したところ，テキスト類似度は大きなデータセットに対して高い性能を発揮し，グラフマイニングは部分的再利用をより正確に検出できることがわかった．以下，第 2 節では類似度に基づく文書再利用検出の概要と既知手法について述べ，評価実験を行って既知手法の改善点を考察する．第 3 節ではグラフマイニング技術を用いたベクタ画像類似検出手法を提案し，その評価を行う．第 4 節にて関連研究を紹介し，第 5 節でまとめと今後の課題を述べる．

2. 類似度に基づく文書再利用検出手法

文書の再利用を検出するためには大きく分けてふたつの方法がある．ひとつは，文書固有の ID と文書インスタンスとの紐づけをファイルシステムの拡張領域や専用の DB に登録しておき，情報の複製や再利用（ファイルシステム上でのコピーなど）を行ったときにこのメタ情報を参照して再利用関係を検出する方法[1] であり，もうひとつは，文書の内容を解析してその類似度から再利用関係を推測する方法である．前者の方法は再利用関係を厳密に追跡することができるものの，エンドポイントのコンピューティング環境が多様な場合は，情報が複数の計算機を移動する間に ID の紐付けが失われてしまう可能性が高い．ゆえに DLP で文書の再利用を検出する際には主に後者の方法が用いられる．

内容の類似度から再利用関係を推測する方法は，更に 2 種類に分類することができる．ひとつはある文書ストアに保管されている全ての既知文書に対して何らかのダイジェスト値を生成して保管しておき，分類対象文書のダイジェスト値と比較をする方法（シグネチャベース手法）であり，もうひとつは分類対象文書と全ての既知文書との間で類似度を計算する方法（類似度ベース手法）である．両者の特徴として，前者は動作が高速である代わりに微細な変更が行われただけでダイジェスト値が一致しなくなる欠点があり，後者は微細な変更に対して頑健である代わりに動作が遅い欠点がある．本研究では微細な変更に対する頑健性を重視して，後者の手法を用いることにする．その他の手法については第 4 節にて紹介する．

2.1 類似度検出手法

ここではまず基本的な類似度検出手法を再利用検出に応用した場合の性能について確認する．

一つ目の手法はテキスト類似度である．これは含まれる単語の出現傾向が似たページ対を類似であると判定する手法である．自然言語処理（情報検索）分野では多様な手法が提案されているが，今回はテキスト全文検索エンジンを使って類似するページを検索する手法を用いる．すなわち，予め既知文書集合に含まれる全ての文書を解析してテキストを抽出し，転置インデックスを生成しておく．分類実行時には分類対象文書からクエリを生成してこの転置インデックスを検索し，高いスコアが付与されたページを類似ページとみなす．この方法は事前に転置インデックスを作成する必要がある代わりに，検索時には全文書と直接比較する必要がないため，総当たり検索を行う必要のある類似度ベース手法に比べて計算量が少ない．また，ファイル共有サービスの場合は検索用にこのような転置インデックスを DLP とは関係なく持っている場合があり，そのようなサービスでは転置インデックスを流用できるメリットもある．

プレゼンテーション文書にはテキスト以外にも，図形・写真などの非言語情報も多く含まれており，それらの特徴を類似検出に利用することも有効であると考えられる．そこで二つ目の手法として，プレゼンテーション文書内の各ページを画像であるとみなしてビットマップ画像検索を行う手法が考えられる．ここで言う画像検索とは，一般にインターネットで見られるような画像の周辺に配置されたキーワードをキーとする検索ではなく，画像の内容をキーとして検索を行う画像内容検索（CBIR; Content-based Image Retrieval）である．

ビットマップ画像類似度検出では見た目の類似性が評価されるため，例えばページの一部分だけを再利用したり，別の情報と併合した場合にうまく類似関係を検出できない可能性がある．そこで非言語情報の利用方法としてもうひとつ考えられるのが，図形・写真の配置状況を何らかの構造で表現し，構造の類似性を検出するベクタ画像類似度検出である．ベクタ画像類似度検出では，構造の表現方法を工夫することで，再利用の際に情報が削除・追加・変更されても類似性を検出できる可能性がある．

2.2 評 価 実 験

まずはテキスト類似度検出とビットマップ画像類似度検出の基本的な性能を確認するため，実際に行われた 3 つの業務プロジェクトで作成された Microsoft® PowerPoint ファイルから，各プロジェクト 50 個ずつランダムに選択して計 150 個のファイル（651 ページ）

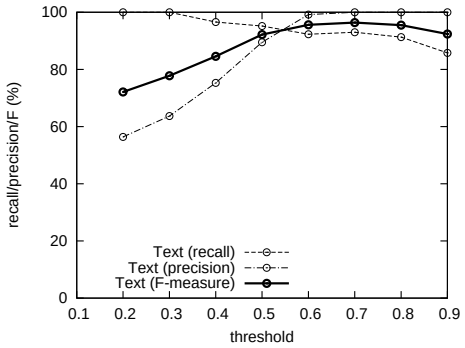

図 2 3 プロジェクト 150 ファイル（651 ページ）のデータセットに対する再利用検出性能

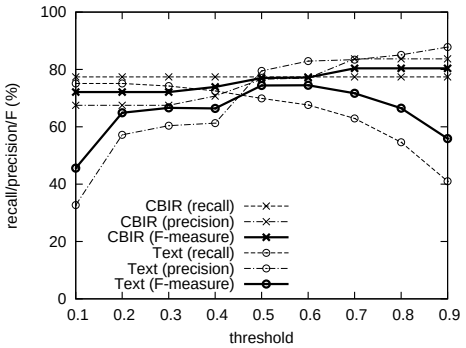

図 3 単一プロジェクト 16 ファイル（122 ページ）のデータセットに対する再利用検出性能

を対象として，再利用の検出性能を比較した．実験に用いたデータセットはページ数が非常に多く，また再利用か否かを示す明示的なメタ情報はないので，閾値を徐々に変化させて目視で再利用か否かの判定を行った．精度については再現率・適合率及び F 値で計測する．その他の条件については以下の通りである．

実行環境 Microsoft Windows® XP SP2 上の Java™Runtime Environment 6.0 (Sun Microsystems 社製)

データの読み込み Microsoft Office 形式ファイルに対してデータの読み書きを実現可能な Java ライブラリである Apache Poi[2] を用いて，プレゼンテーション文書の本文ページ及びマスターページから本文・ノート・ヘッダ・フッタに含まれるテキストを取得する．

テキスト検索エンジンとの統合 テキスト全文検索エンジンである Apache Lucene[3] を用いて事前にテストデータからインデックスを作成しておく．検索単語への分割は Lucene の `CJKAnalyzer` クラスを用いた[⋆1]．

実験の結果，ビットマップ画像類似度に関しては大量の誤検出を起こし，精度が極端に低くなってしまうことがわかった．これは文字主体のページがほぼ全て同一ページであるとみなされてしまうためである．一方テキスト類似度は**図 2** のような性能を得ることができた．

次に，ある特定のプロジェクトに属していて，再利用関係が既知である 16 ファイル（122 ページ）について同様の実験を行った結果を**図 3** に示す．このプロジェクトはテキストよりも図が多かったため，ビットマップ画像類似度も十分な性能が出ており，テキスト類似度を上回る性能となっている．

2.3 考　　察

データセットが十分に大きな場合，テキスト類似度が高い精度で再利用を検出できることがわかった．F 値最大となるのは閾値 70%の場合で，再現率 93.0%, 適合率 100.0%, F 値 96.4%であった．この値は約 20 個に 1 個の割合で再利用の検出漏れが発生することを意味する．ここから更に性能を改良するために，いくつかの実際のデータを観察した結果，以下のような再利用関係のあるページ対の場合，テキスト類似度ではうまく関係を検出できないことがわかった．

図形がほとんど同一なのにテキストが大幅に変更されている 翻訳元文書と翻訳先文書の対が代表例である．

テキストの一部だけを再利用している 情報の一部だけを再利用している場合，テキスト量が大幅に異なる上に特徴的な単語のいくつかが失われているため，検索スコアが低く出る傾向がある．

上記のようなページ対については，テキストを補完する形で図形情報を参照すれば精度よく検出できると考えられる．しかし，ビットマップ画像類似検出は大きなテストデータに対してあまりよい性能が出ていないため，ベクトル画像類似検索が必要であると考えられる．以上から，次節ではテキスト類似とベクタ画像類似を利用して再利用を検出でき，かつ無害な再利用は可能な限り検出しない手法を提案する．

3. グラフマイニング技術を用いた再利用検出手法

本節ではカーネル法を使ったグラフマイニングを利用した文書再利用検出手法を提案する．グラフマイニングは分子構造などグラフ表現可能なデータの類似度を計算することができ，得られた類似度から特定の性質を持つ物質を探索する等の用途に用いられる．グラフマイニング手法の中でも鹿島ら[4] はランダムウォークとカーネル法を組み合わせた手法を提案している．この手法はパラメータであるカーネル関数及び遷移確率分布を目的に応じて適切に設計することで，タンパク質の分類以外にも応用可能である．そこで本論文では，文書再利用検出に適したカーネル関数・確率分布を設計し，得られたカーネル関数を再利用か否かの判定に利用する．

⋆1 空白による分かち書きが行われている文は空白単位で分割し，分かち書きの行われていない文（日本語文）は文字 bigram として分割する．

3.1 グラフマイニングの概要

ランダムウォークに基づくグラフマイニングにおいて，二つのラベル付き有向グラフ G, G' の間のカーネル関数 $K(G,G')$ は以下のように表される．

$$K(G,G') = \sum_{\ell=1}^{\infty} \sum_{\boldsymbol{h}} \sum_{\boldsymbol{h'}} p_s(h_1) \prod_{i=2}^{\ell} p_t(h_i|h_{i-1}) p_q(h_1) \times p'_s(h'_1) \prod_{j=2}^{\ell} p'_t(h'_j|h'_{j-1}) p'_q(h'_\ell) \times K(v_{h_1}, v'_{h'_1}) \prod_{k=2}^{\ell} K(e_{h_{k-1},h_k}, e'_{h'_{k-1},h'_k}) K(v_{h_k}, v'_{h'_k}) \quad (1)$$

ただし

$$p_s(i) : \text{ランダムウォークがノード } i \text{ から開始される確率} \quad (2)$$
$$p_t(j|i) : \text{ノード } i \text{ からノード } j \text{ への遷移確率} \quad (3)$$
$$p_q(i) : \text{ランダムウォークがノード } i \text{ で終了する確率} \quad (4)$$
$$K(v,v') : \text{ノード対 } (v,\, v') \text{ の類似度を示すカーネル関数} \quad (5)$$
$$K(e,e') : \text{エッジ対 } (e,\, e') \text{ の類似度を示すカーネル関数} \quad (6)$$

文献 4) では，p_s 及び p_t として一様分布を，p_s として定数を用いている．また，$K(v,v')$ 及び $K(e,e')$ については，ノードもしくはエッジに付与されたラベルが一致する場合に 1，一致しない場合に 0 を返す関数を用いている．

カーネル関数を端的に表現すると，ある特徴空間上のふたつの特徴ベクトル間の内積であると考えられるから，似通った特徴を持つベクトル対に対して高い値を，異なる特徴を持つベクトル対に対して低い値を返すような関数であると考えてよい．すなわち $K(G,G')$ は，二つのグラフ G, G' の構造がどの程度類似しているのかを表していると言える．よって，類似度を計測したいプレゼンテーション文書のページ対をそれぞれグラフに変換し，その間のカーネル関数の値を求めることで，そのページ対の類似度を得ることができる．

3.2 文書再利用検出へのグラフマイニング応用

オフィス文書に対してグラフマイニングの手法を適用するために，以下の節において，文書内に含まれる各ページをグラフ構造に変換する手続きと，グラフマイニングに必要なパラメータ（p_s, p_t, p_q, $K(v,v')$, $K(e,e')$）を決定する．

3.2.1 グラフ構造への変換

変換の具体例を**図 4** に示す．まず，オブジェクトをノードに変換する．オブジェクトの

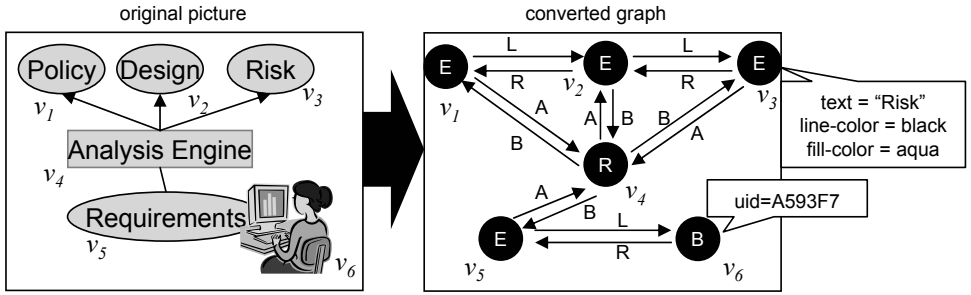

図 4 プレゼンテーション文書の一ページをラベル付き有向グラフへ変換する例

持つプロパティ（テキストを含む）をそのノードが持つ特徴量と考えて，後述する $K(v,v')$ の計算に利用する．続いてノード間をエッジで連結する．このときエッジに付与するラベルとして，連結されるノード間の地理的位置関係（上下左右）を用いる．意図的に荒い粒度のエッジラベルを用いることで，微修正に対して頑健なグラフ構造を目指す．

3.2.2 ランダムウォークパラメータ

次にランダムウォークに関するパラメータ $p_s(i)$, $p_t(j|i)$, $p_q(i)$ を決定する．ここで $p_s(i)$, $p_t(j|i)$ をノード毎に調整することで，ノードを考慮する度合いを変えることができる．そこで今回は主要なオブジェクトを重視して些末なオブジェクトを軽視するようにパラメータを調整する．具体的には，オブジェクトがページ上で占める面積比に比例して遷移確率を割り当てる．例えば図 4 において，v_6 の面積が 100 平方ピクセル，v_4 の面積が 50 平方ピクセル，全オブジェクトの面積の合計が 1000 平方ピクセルであった場合，$p_s(v_6) = 100/1000$ となり，$p_t(v_6|v_5) = 100/(100+50)$, $p_t(v_4|v_5) = 50/(100+50)$ となる．p_q は文献 4) と同様に定数とする．

3.2.3 ノードとエッジのカーネル関数

3.1 節で述べたように，カーネル関数は似通った特徴を持つベクトル対に対して高い値を，異なる特徴を持つベクトル対に対して低い値を返すような関数であり，いくつかの条件（$K(x,y) = K(y,x)$, $K(x,y) > 0$ など，詳しくは文献 5) 参照）を満たすものであれば任意の関数をカーネル関数として利用可能である．まず $K(v,v')$ については，以下のようなプロパティの一致度を線形補間して得る．

テキスト　ノード対に共通して出現する語の割合（Jaccard index）

Vol.2010-CSEC-49 No.8
2010/5/11

表 1 人工データに含まれるオブジェクト数

	1	2	3
データ系列 A	8	26	31
データ系列 B	16	50	59

ビットマップ画像 picture UID[⋆1]の一致

図形プロパティ 前景色・背景色・線種・横幅・縦幅の一致度

$K(e, e')$ については，ラベルが一致する場合 1，一致しない場合 0 を返す簡単な関数を用いる．

3.3 グラフマイニングを用いた再利用検出の評価

ランダムウォークに基づくグラフマイニングを前節で設計した関数を含めて実装し，まず再利用のパターンによって検出精度がどのように変わるかを実験した．実験は 2.2 節とほぼ同様の方法で行ったが，以下の点が異なる．

データセット 再利用のパターンによって検出精度にどのような違いが出るのかを確認するため，ある特定のページ A と別のページ B について，元のデータ（A3/B3）から 2 段階で図形数を減らしたページを作り（A1, A2; B1, B2），さらにそれらを組み合わせて作ったデータ（A1/B1, A1/B2, A1/B3, A2/B1, A2/B2, A2/B3, A3/B1, A3/B2, A3/B3）を作る（ページ A から生成した実験用ページを**図 5** に，A 及び B から組み合わせデータを作る例を**図 6** に，各ページが含むオブジェクト数を**表 1** に示す）．

データの読み込み データ読み込みには同様に Apache Poi[2] を用いるが，テキストではなくオブジェクトを抽出する．Poi では図形を `Shape` インスタンスとして取得可能であり，ここから $K(v, v')$ の計算に必要なプロパティ（テキスト・Picture ID・前景色・背景色・線種・縦幅・横幅）を取り出す．オブジェクトの面積は，オブジェクトのタイプに関係なく（rectangle でも ellipse でも）横幅×縦幅で計算する．

閾値設定 テキスト・ビットマップ類似・グラフマイニングの各手法について、ページ A で F 値最大となる閾値を用いてページ B を評価し，同様にページ B で F 値最大となる閾値を用いてページ A を評価する．

実験結果を**表 2** に示す．図 2 や図 3 と比べてテキスト類似検出の精度が悪化しているが，これはインデックスを生成するためのデータが少なく，単語の特徴量を正しく捉えるのには十分でなかったからであると考えられる．また，検出に失敗した対を詳しく見ていくと，情報量の少ないページ（A1/B1）に類似するページや，テキスト量が大幅に異なる対（A1 と A3/B3 など）を検出できていないことがわかった．テキスト検出では「微修正してそのまま再利用」という再利用パターンに強く，「情報を再利用した上で新たな情報を追加する」という再利用パターンを検出しにくいことになる．これはクエリ文字列に含まれない文字列を多く含むページに対しては低い検索スコアが与えられることが原因と考えられる．ビットマップ画像類似検出はテキスト検索より高い精度で検出可能であったが，グラフマイニングには及ばない結果となった．図 6 の例のように，組み合わせ再利用の場合はページ全体で見ると違いが大きいため，このような結果に繋がったと考えられる．グラフマイニングはデータの多寡に関わらず一定の精度を得られており，ページ A よりもページ B の方がより高い精度で再利用検出が可能であった．表 1 に示したように，ページ B 系統のデータはページ A 系統のデータより多くのオブジェクトを含んでいることから，ページに含まれるオブジェクト数が多いことでよりページの特徴を捉えやすくなっている可能性が考えられる．

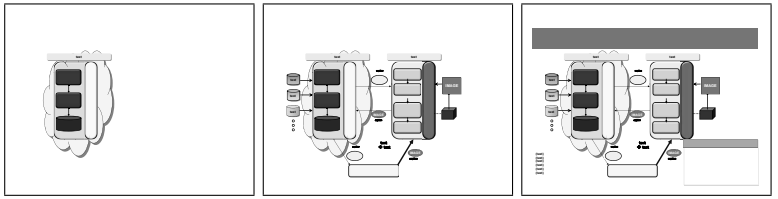

図 5 データ系列 A（左から A1, A2, A3）．実際のデータはテキストを含むが，機密情報であるためこの図からはテキストを削除したり “text” という文字列に置き換えてある．

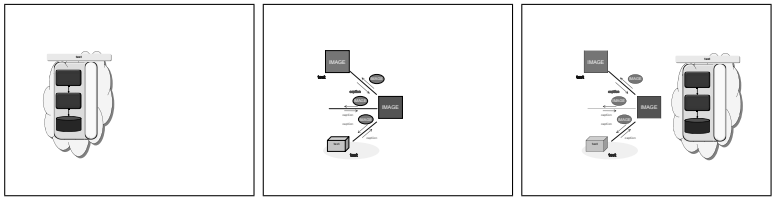

図 6 部分再利用データの組み合わせ例（左から A1, B1, A1/B1）

4. 関 連 研 究

第 2 節冒頭で述べたように，内容に基づいて再利用関係を検出する方法は，シグネチャ

⋆1 Microsoft OLE 形式のファイルに含まれる画像に対して割り当てられる一意の ID．クリップボード経由のコピー＆ペーストを行って文書を跨がるコピー操作を行っても ID は維持される．

表 2 部分再利用の検出精度

	再現率 (%)	適合率 (%)	F 値 (%)
テキスト (A)	46.2	90.9	61.2
テキスト (B)	27.7	100.0	43.4
ビットマップ画像類似 (A)	50.0	72.2	59.1
ビットマップ画像類似 (B)	50.0	72.2	59.1
グラフマイニング (A)	51.9	93.3	66.7
グラフマイニング (B)	100.0	79.4	88.5

ベースと類似度ベースの 2 種類に大別することができる．シグネチャベースの手法では，何の加工もせずにデータをハッシュ関数に与えてダイジェスト値を計算すると完全に同一の文書対しか検出することができないので，類似文書を検出するためには，データの異なりがわずかな文書対には同一のダイジェスト値を与えるようにデータを前処理する必要がある．例えばテキスト文書に対するチャンキングを工夫したり[6),7)]，ダイジェストの計算対象から機能語（前置詞・冠詞等，内容と関係性が薄い語）を除去する I-Match signature[8),9)] といった手法が提案されている．多くの手法では，最終的なダイジェスト値の計算手法は SHA-1[10)] など既知のハッシュ関数がそのまま用いられているが，データの類似度によってダイジェスト値の衝突確率が決まる特別なハッシュ関数の存在が文献 11)，12) によって示されているので，これを用いている手法もある[13)]．

類似度計算の手法としてよく用いられるのは編集距離（X-Delta[14)] などのいわゆる diff プログラム）である．変種として文書を木構造のカテゴリに分類した上で木同士の編集距離を計算する手法[15)] もある．データが自然言語文である場合は,「微妙な変化」を言語間翻訳における表記変化と同一であると捉えた手法[16)] や，最尤推定で得た文書間の KL (Kullback-Leibler) 距離とタイムスタンプ・文書 ID 番号等のメタデータを組み合わせた手法[17)] を利用できる．また，データが画像である場合は，画像内容検索[18),19)] やその実装[20)] を適用できる．ベクトル画像の類似度検索手法としては 21) が挙げられる．本研究で対象とした画像とテキストが組み合わせされているデータに対しては，文献に含まれるイメージとそのキャプションから topic を検出する手法[22)] も利用可能である．近年のオフィス文書は XML 形式で表されることが多いので，XML 同士の類似度検出手法[23),24)] も利用できる可能性がある．なお，類似度ベースの手法は計算量が多くなるため，類似度計算を行う前に既知文書に対してフィルタリングを行い，計算回数を削減する技術（例えば文献 25)）も研究対象となっている．そのほか，再利用検出は情報漏洩対策以外にもプレゼンテーションのストーリーを見つけたり[26)]，テンプレートから作成された契約文書が法令に遵守した形になっているかどうかをチェックする[27)] ことにも用いられる．

本研究では機密情報の再利用に着目しているため，ある分類対象文書が機密なのかどうかを検出する手法については取り上げていない．実際の DLP システムでは，例えば文書内に個人情報保護法における個人情報を含んでいるか否か，といった文書そのものの機密性判定も必要である．通常この用途には正規表現等のパターンマッチング技術か，機械学習を利用した文書クラスタリングが用いられる．これらを組み合わせて効率よく機密情報漏洩を検出する手法[28)] が提案されている．

DLP は以上のような機密分類・再利用検出にポリシー管理・ポリシー強制の機能を持った製品のカテゴリ名である．DLP が必要とする技術については 29) が参考になる．

5. おわりに

本研究ではテキスト類似度・ビットマップ画像類似度を用いた再利用検出の精度を確認し，更にグラフマイニングを用いてベクタ画像類似度を考慮した方法について検討を行った．テキスト類似度は大きなデータセットに対して高い性能を発揮し，グラフマイニングは部分的再利用をも検出できることがわかった．今後の課題としてはパフォーマンスの改善が挙げられる．グラフマイニングは速度が問題になるため（二つのグラフ G, G' がそれぞれ n, m 個のノードを持っている場合の計算量オーダは $O((nm)^2)$），何らかのフィルタリングによって計算対象のデータ数を削減するか，シグネチャベース手法を取り入れることで計算量を削減する必要がある．また，より正確な評価を行うため引き続き様々なデータセットを用いて実験を行う必要があると考えている。

参 考 文 献

1) Mishina, T., Yoshihama, S. and Kudo, M.: Fine-grained Sticky Provenance Architecture for Office Documents, *IWSEC '07: the International Workshop on Security 2007*, Nara, Japan (2007).
2) Apache Poi. `http://poi.apache.org/`.
3) Apache Lucene. `http://lucene.apache.org/`.
4) Kashima, H., Tsuda, K. and Inokuchi, A.: Marginalized kernels between labeled graphs, *ICML '03: Proceedings of the Twentieth International Conference on Machine Learning*, AAAI Press, pp.321–328 (2003).
5) Bishop, C.M.: *Pattern Recognition and Machine Learning*, Springer Verlag, New York (2006).
6) Brin, S., Davis, J. and García-Molina, H.: Copy detection mechanisms for digital

documents, *SIGMOD Rec.*, Vol.24, No.2, pp.398–409 (1995).
7) Forman, G., Eshghi, K. and Chiocchetti, S.: Finding similar files in large document repositories, *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, ACM, pp.394–400 (2005).
8) Chowdhury, A., Frieder, O., Grossman, D. and McCabe, M.C.: Collection statistics for fast duplicate document detection, *ACM Transactions on Information Systems*, Vol.20, No.2, pp.171–191 (2002).
9) Kołcz, A., Chowdhury, A. and Alspector, J.: Improved robustness of signature-based near-replica detection via lexicon randomization, *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, ACM, pp.605–610 (2004).
10) of Standards, N. I. and (NIST), T.: FIPS 180-3: Secure Hash Standard (SHS) (2008). `http://csrc.nist.gov/publications/fips/fips180-3/fips180-3_final.pdf`.
11) Indyk, P. and Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality, *STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing*, New York, NY, USA, ACM, pp.604–613 (1998).
12) Charikar, M.S.: Similarity estimation techniques from rounding algorithms, *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, New York, NY, USA, ACM, pp.380–388 (2002).
13) Manku, G.S., Jain, A. and DasSarma, A.: Detecting near-duplicates for web crawling, *WWW '07: Proceedings of the 16th international conference on World Wide Web*, New York, NY, USA, ACM, pp.141–150 (2007).
14) MacDonald, J.P.: File System Support for Delta Compression, Technical report, Masters thesis, Department of Electrical Engineering and Computer Science, University of California at Berkeley (2000).
15) Lakkaraju, P., Gauch, S. and Speretta, M.: Document similarity based on concept tree distance, *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, New York, NY, USA, ACM, pp.127–132 (2008).
16) Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A. and Zobel, J.: Similarity measures for tracking information flow, *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, New York, NY, USA, ACM, pp.517–524 (2005).
17) Yang, H. and Callan, J.: Near-duplicate detection by instance-level constrained clustering, *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, ACM, pp.421–428 (2006).
18) Deselaers, T., Keysers, D. and Ney, H.: Features for image retrieval: an experimental comparison, *Information Retrieval*, Vol.11, No.2, pp.77–107 (2008).
19) Owen, T.: A Comparison of Systems for Content-based Image Retrieval, *Multimedia Systems Conference* (2007).
20) FIRE (the Flexible Image Retrieval Engine). `http://www-i6.informatik.rwth-aachen.de/~deselaers/fire/`.
21) Fonseca, M.J., Barroso, B., Ribeiro, P. and Jorge, J.A.: Retrieving Vector Graphics Using Sketches, *Proceedings of the Smart Graphics Symposium* (2004).
22) Chen, X., Lu, C., An, Y. and Achananuparp, P.: Probabilistic models for topic learning from images and captions in online biomedical literatures, *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, New York, NY, USA, ACM, pp.495–504 (2009).
23) Wang, Y., DeWitt, D.J. and Cai, J.-Y.: X-Diff: An Effective Change Detection Algorithm for XML documents, *ICDE '03: Proceedings of the 19th International Conference on Data Engineering*, pp.519–530 (2003).
24) Viyanon, W. and Madria, S.K.: A system for detecting xml similarity in content and structure using relational database, *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, New York, NY, USA, ACM, pp.1197–1206 (2009).
25) 立石健二，久寿居大：Multi-level prefix-filter を用いた高速重複文書照合，日本データベース学会 Letters， Vol.5, p.4 (2007).
26) 中沢拓磨，久保田秀和，角 康之，西田豊明：再利用部分の抽出によるプレゼンテーションストーリーの変遷の可視化，人工知能学会全国大会 (2007).
27) Sayeed, A., Sarkar, S., Deng, Y., Hosn, R., Mahindru, R. and Rajamani, N.: Characteristics of document similarity measures for compliance analysis, *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, New York, NY, USA, ACM, pp.1207–1216 (2009).
28) 加藤 守，柴田秀哉，郡 光則：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，FIT2009: 第 8 回情報科学技術フォーラム講演論文集 (2009).
29) Understanding and Selecting a Data Loss Prevention Solution (2007). `http://securosis.com/publications/DLP-Whitepaper.pdf`.