情報処理学会論文誌 Vol. 52 No. 2 645–655 (Feb. 2011)

文字コード処理方式による高速な印刷コントロール機能の開発

甲斐 賢[†1] 笈川光浩[†1] 伊川宏美[†1]
今一 修[†1] 森本康嗣[†1] 土田健一[†2]
手塚 悟[†3] 荒井正人[†1] 洲崎誠一[†1]

オフィスのセキュリティ対策では，文書の漏洩防止のため，印刷出力コントロールが重要である．本稿は，文書内容を解析したうえで各種コントロールを行う DLP（Data Loss Prevention）機構に着目し，印刷内容の解析を高速化した印刷コントロール機能の基本設計と開発結果を述べる．印刷文書を識別するために，文書 OCR と仮想プリンタドライバを組み合わせた画像処理方式（従来方式）では，フルテキストを取得するのに，仮想プリンタドライバで行ったレイアウト配置を再び文書 OCR で解析し直すという処理の重複がある．そこで文書 OCR を使わずに，仮想プリンタドライバ内でレイアウト配置は無視しつつ，ページ内で処理される文字出力を処理順番のとおりにつなげてフルテキストを取得する文字コード処理方式（提案方式）を考案した．提案方式を Windows XP 上で実装することで，仮想プリンタドライバでレイアウト配置を無視してもフルテキスト取得への影響が小さいことを確認し，印刷速度の低下が 0.4 秒程度以下に抑えられる見通しを得た．

Development of Rapid Print-out Control Using Character Code Processing Method

Satoshi Kai,[†1] Mitsuhiro Oikawa,[†1] Hiromi Igawa,[†1]
Osamu Imaichi,[†1] Yasutsugu Morimoto,[†1]
Kenichi Tsuchida,[†2] Satoru Tezuka,[†3]
Masato Arai[†1] and Seiichi Susaki[†1]

For office's security countermeasure, it is important to control print-out because of too many documents leakage. This paper addresses a DLP (Data Loss Prevention) mechanism which executes document controls after analyses contents of a document, and describes the basic design and the development result of rapid print-out control function. The conventional image processing method to identify the document used a document OCR and a virtual printer driver to acquire a full-text. But the conventional method had duplication processing between the layout arrangement in the printer driver and the layout analysis in the document OCR. So we propose the new character code processing method to identify the document without the document OCR, which method use the virtual printer driver to hook character output and to acquire a full-text ignoring the layout arrangement. We implemented the proposed method on Windows XP and confirmed that there was little influence ignoring the layout arrangement on acquiring the full-text. And we confirmed that print-out speed was added to only 0.4 seconds.

1. はじめに

企業や組織における情報セキュリティ対策として，オフィスの情報漏洩対策は重要な取り組むべき課題である．オフィスとは「組織が業務のために利用する建屋又は居室等」と定義されており[1)]，つまり，入退室管理で区切られた，特定の業務を行う役割が与えられた従業員の作業場所である．このようなオフィスで取り扱う文書は，紙文書も電子文書も両方存在することが通常であり，電子文書から紙文書に印刷するためのプリンタもオフィスに備えられることも多い．近年，企業で生成される文書の 93%が電子文書であるといわれる[2)] 一方で，企業からの情報漏洩経路として，紙文書が 55.9%と最多である[3)]．つまり，生成量として少ない紙文書の方が，漏洩の主な原因になっているのが現状である．これら紙文書の生成は，業務の IT 化の進展により，電子文書から印刷出力されることが多いと予想されるため，情報漏洩対策の一環として，印刷出力のセキュリティ強化は重要である．

本稿では，印刷出力のセキュリティ強化を目的とし，印刷内容を高速で判定可能とする印刷コントロール機能の開発について述べる．以下，2 章で既存セキュリティ対策と，本稿の貢献範囲を述べ，3 章で従来の印刷コントロールにおける課題を述べる．4 章で解決方針とフィジビリティ検証結果を説明し，5 章で高速性の評価を述べる．

†1 株式会社日立製作所
Hitachi Ltd.
†2 日立アイ・エヌ・エス・ソフトウェア株式会社
Hitachi INS Software, Ltd.
†3 東京工科大学
Tokyo University of Technology

2. 既存セキュリティ技術

2.1 印刷セキュリティ対策

印刷セキュリティ対策は，情報漏洩対策の一環として，およそ (1) 予防対策，(2) 検出対策，(3) フォレンジック対策に大別できる．

まず (1) 予防対策は，印刷しても情報漏洩にならないよう未然に防止する対策である．予防対策の例としては，印刷自体を防ぐ場合 (1a)〜(1c) と，印刷内容を加工する場合 (1d), (1e) がある．

(1a)　プリンタドライバのインストールを，OS セキュリティ機能で禁止する．

(1b)　電子文書を開くアプリケーションのセキュリティ機能で印刷を禁止する[4),5)]．

(1c)　印刷権限のある人のみが IC カードをかざすことで印刷できるようにする．

(1d)　あらかじめ電子文書の一部を墨塗りする[6)]．

(1e)　電子文書の必要な部分をモザイク状に暗号化し，権限のある人のみが復号して閲覧できるようにする[7)]．

次に (2) 検出対策は，漏洩されたとしても印刷出力結果から漏洩を見つける対策である．検出対策の例としては，あらゆる電子文書にあらかじめ検出のための特徴情報を付与しておき印刷時に特徴情報も印刷する場合 (2a) と，印刷時に印刷出力結果に特徴情報を付与する場合 (2b) がある．

(2a)　電子文書のヘッダやフッタ，背景文字などに，出所（Copyright など），機密レベル（極秘，社外秘など），取扱い方法（持ち出し厳禁，印刷不可など）を記述する[8),9)]．

(2b)　印刷時に，ヘッダやフッタ，背景文字などに，機密レベル，取扱い方法などをプリンタドライバで追加して印刷する．

最後に（3）フォレンジック対策は，もし漏洩事故が起きたとしてもその被害拡大や復旧コストを最小に抑える対策である．フォレンジックの対策の例としては，事故発生後の調査 (3a), (3b) と，事故発生後の調査を確実化・効率化するための事前対策 (3c) とがある[10)]．

(3a)　流れるネットワークパケットを監視することで，漏洩の兆候を把握する．ネットワークフォレンジックと呼ばれる．

(3b)　漏洩が疑われる利用者の PC から決定的な証拠を探し出す．コンピュータフォレンジックと呼ばれる．

(3c)　印刷文書ごとに異なる ID 情報を電子透かしとして挿入し，事故後に，漏洩した紙文書からその ID 情報を抽出し，いつ，だれが，何を印刷したかを追跡する．

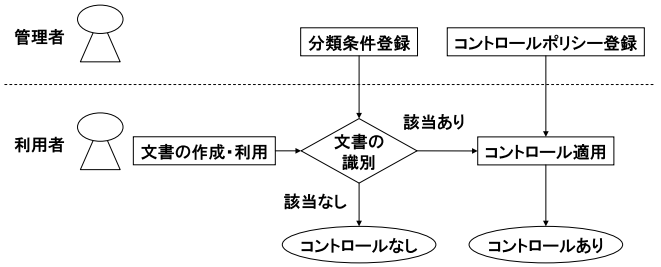

図 1　DLP 機構
Fig. 1　A mechanism of data loss prevention.

さらには，前述の（1）予防対策と（2）検出対策とを組み合わせたような，2.2 節に述べる新たな印刷コントロール方法も知られるようになってきた．

2.2 DLP（Data Loss Prevention）

情報漏洩対策として，近年，DLP と呼ばれるセキュリティ製品が注目を集めている[11),12)]．DLP は，印刷出力に限らず，メール転送，USB メモリ書き出し，Web アップロードなどの様々な漏洩経路に対してコントロールを行うことができる．

DLP を実現するアーキテクチャ（以下 DLP 機構と呼ぶ）の全体像を**図 1** に示す．DLP 機構では，まず管理者が「分類条件登録」と「コントロールポリシ登録」を行う．分類条件とは，たとえば「最重要書類」「個人情報（を含む）」「機密情報扱い」などに該当する文書の条件を登録し，それ以外の文書を「区分なし」と見なす，といったものである．コントロールポリシとは，たとえば「防止」「（管理者への）アラート」「（利用者への）警告」「ログ記録」といったコントロールの種類を，分類条件ごとにそれぞれ定義したものである．

次に利用者は業務を遂行するために「文書の作成・利用」を行う．同時に DLP 機構は，どの分類条件に合致するかを判定する「文書の識別」と，文書に対しポリシに従ったコントロールを行う「コントロール適用」を行う．本機構の特徴は，「文書の識別」という検出対策と，「コントロール適用」という予防対策の組合せにより，分類条件に合致する場合にのみ，コントロールポリシを適用する点にある．

このような文書の識別のためには，2.1 節の検出対策で述べた，電子文書に付与された特徴情報を見つけることで識別することもできる．しかし，こうした特徴情報は利用者の目に見えるため，悪意のある利用者が特徴情報を削除することや，別文書に一部コピーすることで特徴情報の検出を容易に回避されるおそれもある．そこで，文書の内容からその特徴を抽出する，

表 1　情報漏洩経路と文書の識別・コントロール方法
Table 1　Set of information leakage routes and how to identify documents and control.

情報漏洩経路	文書の識別方法	コントロール適用方法
紙媒体	画像ベース解析	印刷制御，複写制御，スキャン制御
Web・Net	HTTP 解析，ネットワークプロトコル解析	送信制御，受信制御
USB 等可搬記録媒体	ファイルベース解析	外部媒体（USB，CD-R 等）書き出し制御
Email	SMTP 解析，MIME 解析	送信制御，受信制御
PC 本体	ファイルベース解析	コピー制御，ファイル検出

フィンガプリント技術と呼ばれる手法が知られている．こうしたフィンガプリント技術は，文書中に含まれる文字列を取り出し，登録された文書との類似度を測るものである[13),14)]．

2.3　本稿の貢献

DLP 機構を実現するには，文書の識別とコントロール適用を，どう組み合わせるかが鍵となる．なぜなら，文書の識別を実現しやすい個所と，コントロール適用を実現しやすい個所とが必ずしも一致するとは限らないからである．文献 3) によると，情報漏洩経路として上位 5 経路だけで全体の 9 割以上を占め，その内訳は多い順に「紙媒体」「Web・Net」「USB など可搬記録媒体」「Email」「PC 本体」である．これらの主要な情報漏洩経路に対して，DLP 機構を実現するには，**表 1** に示す文書の識別方法およびコントロール適用方法の組合せが考えられる．

ここで文書の識別に関して，ネットワークプロトコル解析やファイルベース解析であればデジタル処理可能であるため実現しやすい．一方，文書の識別を画像ベースの解析で行うことはアナログ処理が残ってしまうため，DLP 機構を実現するにあたって，精度と処理速度の点で不利な立場にある．

本稿は，従来の画像ベース解析の代替手段として，精度と処理速度を向上させた文書の識別方法を提案するものである．これにより，利用者が電子文書を印刷するときに，その印刷内容に応じて，許可，禁止，アラート，警告，ログ記録，電子透かし挿入などを行うことを目的とする．

3.　印刷コントロールの従来技術

3.1　印刷環境の現状

オフィスにおける主要な業務の 1 つは印刷であり，印刷速度は年々向上していることは明らかである．

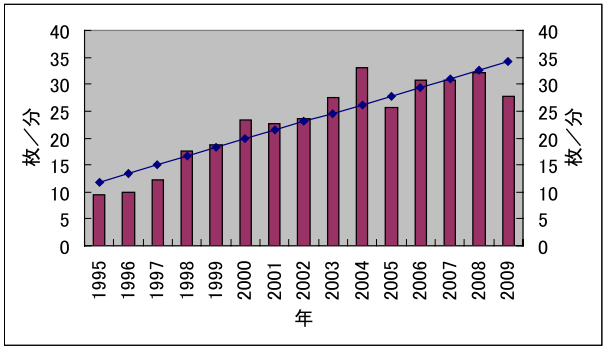

図 2　印刷速度の向上（A4 モノクロ印刷）*1
Fig. 2　Improvement of print-out speed (A4 monochrome print).

どの程度向上しているかを検証するため，国内主要プリンタメーカ 5 社が 1995 年から 2009 年にかけて販売したプリンタ 254 機種を対象に，A4 モノクロ印刷で，1 分あたり印刷可能な枚数を調べた．販売開始年度ごとの印刷速度の平均値をヒストグラムに示した結果を**図 2** に示す．販売開始年度と印刷速度とを回帰分析したところ，重相関係数 R = 0.91 が得られ，年々1.61 枚/分の割合で速くなっていることが判明した．印刷コントロールを行うには，こうした印刷速度を落とさないことが必要である．

3.2　画像処理方式による文書の識別

紙媒体の文書を識別するには，画像ベース解析として OCR（Optical Character Recognition）の利用が多い．実際，文献 15) では印刷時に OCR を利用して印刷内容を把握する複合機が公開されている．本稿では，OCR を利用した文書の識別方法を画像処理方式と呼ぶことにする．OCR はその用途により 2 種類に大別できる[16),17)]．

- 文書 OCR：書籍や新聞などあらかじめ形式の決まっていない印刷図書を読み取ることを目的としたもの．文書のレイアウトを自動的に解析，理解することで，表や文章と図や写真を自動的に分離して読み取ることが可能
- 帳票 OCR：定型業務に即されて作られた，あらかじめ決まった形式の帳票や伝票を読み取ることを目的としたもの

⋆1 2009 年の印刷速度の平均が 2008 年の平均よりも遅くなっている理由の 1 つに，印刷速度を落として低価格化を狙ったプリンタの数が増えたことがあげられる．

オフィスにおける印刷文書は多様であるため，その文書を識別するには，文書 OCR の利用が適当である．この文書 OCR の処理動作は，その特徴からすると「レイアウト解析」と「文字解析」に大別できる．

- レイアウト解析：画像全体から文章・表・図の範囲を区別する，文章についてその範囲や縦書き・横書きを区別するなど，文書のレイアウト情報を得る．
- 文字解析：使用言語を区別する，フォントや文字装飾（太字，斜体，下線，カラー，白黒反転など）の差異を吸収して文字だけを識別する，文章解析により誤検出文字を補正するなど，文字情報を得る．

特にレイアウト解析と文字解析は互いに独立した処理ではなく，文字解析の前提に必ずレイアウト解析が必要である．そのため画像処理方式は文書の識別に時間がかかる傾向にある．

3.3　仮想プリンタドライバによる印刷コントロール

オフィスにおけるプリンタは，様々な機種のプリンタが使われるのが通常であるため，プリンタ機種によらない印刷コントロールが望まれる．そうした印刷コントロールを実現するために従来技術として仮想プリンタドライバによる方式が提案されている[18]．本方式では，あらかじめ利用者の PC に仮想プリンタドライバをインストールしておき，業務で印刷するときには，利用者はアプリケーションから仮想プリンタドライバを呼び出して印刷操作を行う．すると，仮想プリンタドライバが印刷履歴取得や電子透かし挿入などのコントロールを行った後，その仮想プリンタドライバから実プリンタドライバを呼び出して，通常の印刷を行う．本方式のメリットは次に示すとおりである．

- 利用者がアプリケーションから印刷する操作性が既存の操作性と変わらない．
- 現在利用しているアプリケーションやプリンタを継続利用できる．
- 仮想プリンタドライバの出力は汎用性が高い Bitmap ファイルであり，どのような実プリンタドライバでも印刷連携できる．

さらに同文献[18] によると，利用者が印刷アプリケーションから直接実プリンタドライバを呼び出すことでコントロールを回避するという攻撃に対する対策も述べられている．

この仮想プリンタドライバ方式では Bitmap ファイルに出力できるために，その出力を文書 OCR の入力とし，文書 OCR の結果を受けて文書を識別する，という連携は可能である．しかし，このような連携では，依然として 3.2 節の画像処理方式で述べたように，文書の識別に時間がかかる傾向にある．

3.4　課　　題

印刷コントロールにおける DLP 機構を実現するには，以下の 2 つの方向性がある．

- 文書の識別方法として，画像処理方式（従来方式）を改良する．
- 文書の識別方法として，従来方式とは異なる別のアプローチを採用する．

いずれの方向性も重要であるが，前者のアプローチでは 3.3 節で述べたようにアナログ処理が残り，文書の識別に時間がかかるという問題が残る．そこで，本稿では後者のアプローチをとることにした．つまり，本稿における課題は，画像処理方式とは異なる方式で，印刷コントロールとして高速に文書を識別する方法を確立することである．

4.　文字コード処理方式による文書の識別

4.1　プリンタドライバでの文字コード処理

プリンタドライバとは，アプリケーションが印刷時に呼び出し，プリンタへの印刷出力のために使用されるドライバである．プリンタドライバの処理動作は，その特徴からすると「レイアウト配置」と「文字出力」に大別できる．

- レイアウト配置：どのページの，どの位置に配置するかなどを決める．
- 文字出力：印刷フォント，サイズ，カラー，装飾などを決めて文字を出力する．

特に文字出力では，電子文書に含まれる文字コードの集まりを，紙文書に印字する文字の形に変換する処理を行う（**図 3**）．たとえば電子文書中の「あ」は「U+3042」という文字コードで表されており，プリンタドライバによる処理で印刷時に，「あ」という文字の形に変換されることになる．3.3 節で述べた仮想プリンタドライバの場合には，プリンタ用言語の代わりに Bitmap ファイルに出力される．

文書を識別するのに仮想プリンタドライバと文書 OCR を組み合わせた画像処理方式（従来方式）の場合には，仮想プリンタドライバでレイアウト配置した画像情報を，文書 OCR でもう 1 度レイアウト解析し直すという処理となる（**図 4** (1)）．文書の識別において重要なのは，レイアウト情報よりも文字情報の方であり，そのため画像処理方式で文書を識別することは，レイアウト配置とレイアウト解析が重複している．そこで，仮想プリンタドライバの中で，レイアウト配置する前に出力文字を文字コードとして取得することができれば，文書の識別にかかる時間を短くできると考えた（図 4 (2)）．こうした考えに基づき，仮想プリンタドライバで文字を取得し文書を識別する方式を，文字コード処理方式（提案方式）と呼ぶことにする．

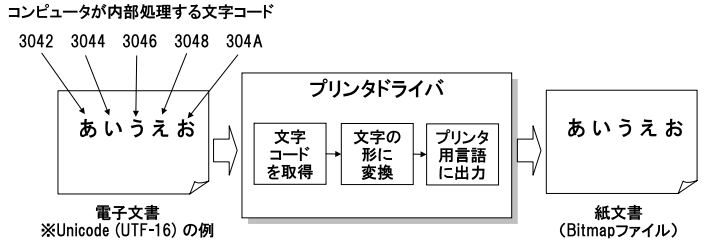

図 3　文字出力処理の概要
Fig. 3　Outline of character output processing in printer driver.

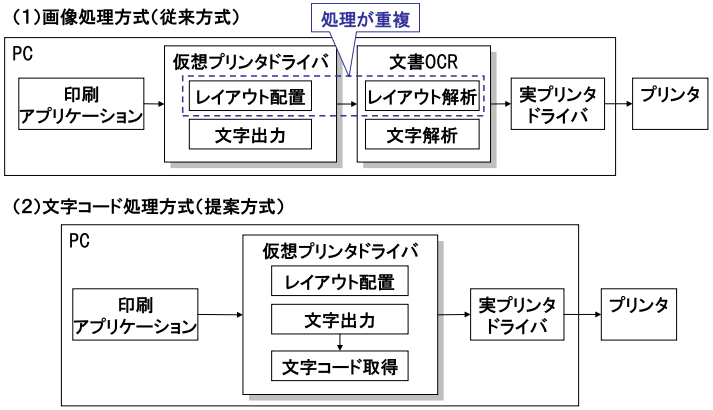

図 4　文字コード処理方式
Fig. 4　Character code processing method.

4.2　Windows XP [*1] 上での実装

Windows XP 上で文字コード処理方式を実装した．Windows XP ではプリンタドライバでの処理は，擬似コードで表すと**図 5** に示すとおりとなる[19]．文書全体の処理（DrvStartDoc〜DrvEndDoc），ページ単位の処理（DrvStartPage〜DrvSendPage），ページ内のレンダリング処理に大別できる．

これらの各処理で呼ばれる関数 DrvStartDoc，DrvStartPage などに対し，OS により DDI

```
DrvStartDoc()                  // 文書の印刷の開始
For each physical page {
    DrvStartPage()             // ページの印刷の開始
        Rendering operations   // レンダリング処理；レイアウト配置，文字出力
    DrvSendPage()              // ページの印刷の終了
}
DrvEndDoc()                    // 文書の印刷の終了
```

図 5　プリンタドライバ処理の擬似コード
Fig. 5　Pseudo code of processing in printer driver.

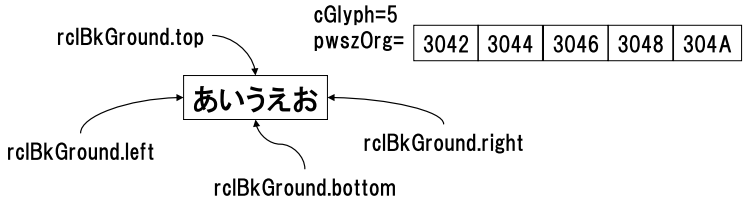

図 6　レイアウト配置と文字出力の例
Fig. 6　An example of layout arrangement and character output.

（Device Driver Interface）Hooking と呼ばれるフッキング処理が提供されている．フッキング処理を利用すると，印刷処理における各種制御情報の参照や変更が可能となる．DDI Hooking 処理可能な関数の 1 つが，レンダリング処理で呼ばれる DrvTextOut である．DrvTextOut 関数の内部では，**図 6** に示すようなレイアウト配置と文字出力の情報を取得・参照できる．

- レイアウト配置：rclBkGround 変数が文字の配置（境界矩形の上下左右のピクセル位置）を表す．
- 文字出力：cGlyph 変数が文字の数を表し，pwszOrg 変数は文字コード（Unicode 文字コードあるいは Glyph（印刷フォント）文字集合）を表す．なお，Unicode と Glyph は相互に変換可能である．

文字コード処理方式の考え方は，4.1 節で述べたようにレイアウト配置する前に出力文字を取得することである．そこでページ単位の処理（DrvStartPage〜DrvSendPage まで）において，レイアウト情報はすべて無視することとし，文字出力情報をページ内の処理順番のとおりにつなげ，ページのフルテキストとして取得することとした．さらに文字はすべて Unicode に統一した．

⋆1 Windows，Windows XP は，米国 Microsoft Corporation の米国およびその他の国における登録商標です．

表 2　フィジビリティ検証結果（1）．文字解析
Table 2　The feasibility result (1). Characters analysis.

検証項目		MS-Word 2003 ファイル	MS-Excel 2003 ファイル	MS-PowerPoint 2003 ファイル
フォント	MS P ゴシック	○	○	○
	MS 明朝	○	○	○
	HGP ゴシック E	○	○	○
装飾	太字	○	○	○
	斜体	○	○	○
	下線	○	○	○
	囲み線	○	—	—
	網掛け	○	—	—
	上付き	○	○	○
	下付き	○	○	○
	取り消し線	○	○	○
	隠し文字	×	—	—
	影付き	—	—	○
サイズ	標準	○ (10.5pt)	○ (11pt)	○ (18pt)
	最小	○ (8pt)	○ (6pt)	○ (8pt)
	1pt	○	○	○
カラー	黒	○	○	○
	赤	○	○	○
	白	○	○	○
	白黒反転	○	○	○

○：成功，×：失敗，—：関係なし（機能なし）

4.3　フィジビリティ検証

文字コード処理方式では，仮想プリンタドライバでレイアウト配置を無視し，出力文字をその処理順番のとおりに文字列としてつなげるために，処理の高速化が可能となる．しかし反面，文字出力が連続して処理される間はその順番どおりに文字列としてつなげることができる一方で，連続して処理されない場合は正しく文字列として取得できない可能性が考えられる．もし正しく文字列として取得できないと，キーワード検索や正規表現などでヒットできなくなる．

そこで，通常の業務で頻繁に使われると考えられる MS-Word，MS-Excel，MS-PowerPoint [⋆1] の 3 つのアプリケーションを対象にテスト文書を作成し，その文書を文字コード処理方式で印刷するときに，正しく文字列を取得できるかどうかをフィジビリティ検証した．フォント，装飾，サイズ，カラーを変えて文字解析を行った結果を**表 2** に，段組みや縦書きと横書きなどのレイアウトを変えてレイアウト解析を行った結果を**表 3** に示す．

まず文字解析の結果，装飾の「隠し文字」を除きすべての検証項目で文字列の取得に成功した．特に，文字のサイズが 1 pt や文字のカラーが白といった，通常 OCR では文字列の取得が困難な場合にも，確実に文字列を取得できることを確認した．ところで「隠し文字」はそもそも文字列が印刷されていないため，取得できないことは当然である．

⋆1 Microsoft Word，Microsoft Excel は，米国 Microsoft Corporation の商品名称です．Microsoft PowerPoint は，米国 Microsoft Corporation の米国，および，その他の国における商標です．

表 3　フィジビリティ検証結果（2）．レイアウト解析
Table 3　The feasibility result (2). Layout analysis.

検証項目		MS-Word 2003 ファイル	MS-Excel 2003 ファイル	MS-PowerPoint 2003 ファイル
段組み	1 段組み	○	○	○
	2 段組み	○	—	—
	3 段組み	△（各段の右端で折り返された最下行を 2 回重複して取得）	—	—
縦書きと横書き	横書き＋上向き	○	○	○
	横書き＋左向き	○	—	—
	縦書き＋左向き	○	○	○
	縦書き＋上向き	○	—	—
	縦書き＋右向き	○	○	○
罫線	3×3 サイズ	○	○	△（最下行セルの右端で折り返された最下行を 2 回重複して取得）
図形描画	テキストを挿入	○	○	○
ヘッダとフッタ	ヘッダ	○	○	—
	フッタ	○	○	○（スライドマスタ）
コメント	テキスト記述	○	○	○
背景文字	透かし	×	—	—
変更履歴	テキスト記述	○	—	—
グラフ	グラフタイトル	—	○	—
	X 項目軸	—	○	—
	Y 数値軸	—	○	—
	系列名	—	○	—
	値	—	○	—
印刷オプション	スライド	—	—	○
	ノート	—	—	○
	配布資料	—	—	○

○：成功，△：一部失敗，×：失敗，—：関係なし（機能なし）

次にレイアウト解析の結果，背景文字の「透かし」を除いてすべての検証項目で文字列の取得に成功した．特に，「コメント」や「グラフ」といった複雑なレイアウトを含む文書であっても，そこに含まれる文字列を確実に取得できることを確認した．ところで，背景文字の取得が失敗した理由は，埋め込む文字列の設定はテキストで行う一方で，MS-Word ファイルへの貼りつけは画像で行われるためと考えられる．また，表 3 によると文字列の取得には成功するものの，右端で折り返しにより最下行に位置する文字列が 2 回重複して取得されることで，不自然な分断に見える項目が MS-Word ファイルの「3 段組み」と MS-PowerPoint ファイルの「罫線」で部分的に認められた．重複して取得されるために，キーワード検索や正規表現で重複してヒットする可能性があるが，少なくともヒットに失敗することはない．

以上のフィジビリティ検証結果からすると，実用上はほぼ問題ないと考えられる．

ところで，文字コード処理方式は，本節で述べたように，通常の業務で頻繁に使うと考えられる文書に対して，その印刷時に精度良くフルテキストを取得できる．しかし，その一方で，プリンタドライバで文字コード処理を行わないような文書を印刷する場合には，フルテキストを取得できない．このような文書には，手元にあるファイルで実験したところ，画像ファイル（Bitmap ファイル，JPEG ファイル，TIFF ファイルなど），PDF [*1] ファイルがあることが判明した．特に PDF ファイルは，業務で MS-Word などで作成した文書を PDF 化して印刷するケースも多く，対応が必要である．

そのため，実際のオフィスに適用するためには，文字コード処理方式とともに，従来方式である画像処理方式との併用を考慮する必要がある．たとえば，最初に文字コード処理方式でフルテキスト取得を試し，取得できたテキストが少なすぎる場合に，画像処理方式を試す，といった併用が考えられる．よって併用の場合にも，時間のかかる画像処理方式を使うかどうかを，高速な文字コード処理方式で判定することが有効となる．

4.4　文書の識別例

(1) 個人情報を含む文書の識別例

文書からフルテキストを取得すると，そのフルテキストを自然言語解析することで文書を識別可能である．ここでは識別の一例として，2.2 節で述べた DLP 機構として，文書が個人情報を含むかどうかを判定することとした．

個人情報の定義は個人を特定できる情報であり，文献 3) によると氏名が 93.7%と多い．そこでフルテキストから氏名（姓あるいは名）を抽出するために，形態素解析ツール「茶筌」[20] を利用した．茶筌で姓あるいは名を判定するための辞書を調べると，日本人の姓と名を合わせて 32,193 語を持つことが判明した，これは日本全人口の 96.27%以上にあたる [21] ため，十分な辞書のサイズであると考えた．

ところで，形態素解析ツールは氏名でないにもかかわらず氏名と誤検出する可能性がある．さらに，文書あたり個人情報を少量のみ含む場合であれば，大量に含む場合に比べて許容できるかもしれない．そこで，しきい値として 1 ページあたり 10 名以上の氏名が含まれていれば，その文書を個人情報と判定することにした．

システム構成の全体を**図 7** に示す．プロトタイプ開発したのは，印刷時にフルテキストを取得する仮想プリンタドライバと，形態素解析ツールを呼び出して氏名の数をカウントする自然言語処理サービスである．

筆者の手元にあった報告書ファイル（MS-Word の 1 ファイル）と名簿ファイル（MS-Excel の 1 ファイル）を対象に，文書の識別を試したところ，名簿ファイルのみを個人情報と判定できることを確認した．

さて上記プロトタイプでは，1 ページ単位での自然言語処理を行うこととしたが，この場合にはページをまたがる単語は認識できない．ただしこのような制限を設けたとしても，名簿などに含まれる固有名詞について，わざわざページをまたがって作成することは通常業務ではほとんどないと考えられる．

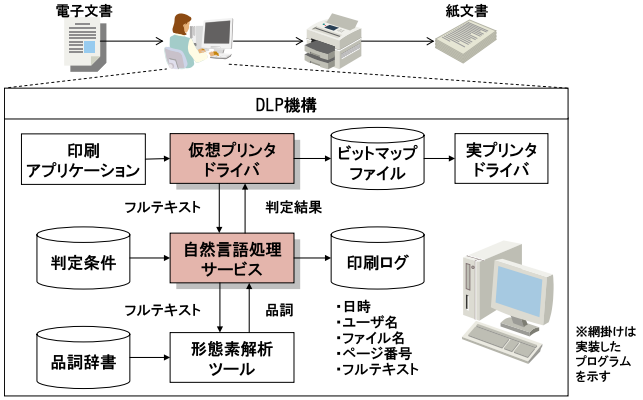

図 7　個人情報を検出する印刷コントロール機能
Fig. 7　Print-out control function for detecting personally identifiable information.

⋆1 PDF は，Adobe Systems Incorporated（アドビシステムズ社）の米国，および，その他の国における登録商標です．

(2)「最重要書類」「機密情報扱い」などの文書の識別例

オフィスにおける情報漏洩対策として，「最重要書類」「機密情報扱い」などの機密レベルや取扱い方法などを，電子文書のヘッダや背景文字に記載することは広く知られている．文字コード処理方式は，文書の本文のほかにも表 3 からするとヘッダに含まれる文字列もフルテキストの一部として取得できるが，背景文字の文字列は取得できない．そのため提案方式では，電子文書の機密レベルや取扱い方法などの文字列がヘッダに含まれるならば，それらを印刷時に取得し，文書の内容に応じて印刷禁止などのコントロールを行うことにも応用できる．ただし，背景文字の文字列は取得できないため，実運用上は少なくとも文書のヘッダ部に機密レベルなどを記載する必要がある．

5. 高速性の評価

5.1 実 験 方 法

文字コード処理方式は画像処理方式に比べて，文書からフルテキストを取得する時間が短いことが特徴である．そこで両方式を使って文書を印刷する場合を比較することにした．印刷時間の比較方法を**図 8** に示す．ここで印刷時間を分けて検討するために，下記に示す変数を定めた．

- $t1$：Bitmap 出力のみを行う仮想プリンタドライバ A による印刷時間
- $t2$：Bitmap 出力とフルテキスト取得を同時に行う仮想プリンタドライバ B による印刷時間
- $t3$：文書 OCR により Bitmap ファイルからフルテキストを取得する時間
- $t4$：Bitmap ファイルを実プリンタドライバで印刷する時間

従来方式である画像処理方式による印刷時間 Ta は，$Ta = t1 + t3 + t4$ で表される．提案方式である文字コード処理方式による印刷時間 Tb は，$Tb = t2 + t4$ で表される．さて，$t4$ は実際のプリンタによる印刷時間であるが，レーザプリンタやインクジェットプリンタといったプリンタの種類によって大きく異なり，さらに Ta と Tb で共通して出てくるため，本評価では $t4$ を除いて考える方針とした．印刷を行う PC の測定環境を**表 4** に示す．

仮想プリンタドライバ A と仮想プリンタドライバ B は，プロトタイプ開発した．文書 OCR は，文書 OCR 製品を 2 種類（A 社製 2005 年版，S 社製 2005 年版）比較したところ文字認識時間が倍近く違ったことから，高速な A 社製 2005 年版を採用した．印刷する文書には，実際の業務で作成した A4 サイズで 5 ページの MS-Word 文書を用意した．当文書に含まれる文字数は，1 ページ目 1,861 文字，2 ページ目 1,615 文字，3 ページ目 2,064 文字，4 ページ目 1,556 文字，5 ページ目 1,851 文字である．

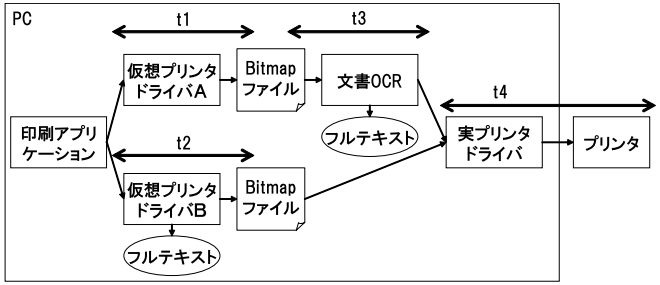

図 8　印刷時間の比較方法
Fig. 8　The comparison method of printing-out time.

表 4　測定環境
Table 4　The measurement environment.

項目	スペック
ハードウェア	VMware Player 3.0 を利用して構築⋆1
CPU	Inter Core2 Duo 2.2 GHz⋆2
メモリ	512 MB
HDD	20 GB
OS	Windows XP Professional SP3

また時間測定方法は，ストップウォッチを使って次に示すレスポンス時間を 5 回ずつ測定し，その平均を計算した．

- $t1$，$t2$ は，MS-Word アプリケーションで印刷ボタンを押下してから，MS-Word アプリケーション表示領域の右下に印刷中に出るプリンタのアイコンが消えるまでを目視で確認した．
- $t3$ は，A 社製の文書 OCR で，文字認識を開始するボタンを押下してから，文字認識完了後に出てくる結果画面が表示されるまでを目視で確認した．

5.2 実 験 結 果

印刷ページ数を増やしながら時間を測定した結果を**表 5** に示す．フルテキスト取得時間だけに要する時間は，従来方式では $t3$，提案方式では $t2 - t1$ で表される．提案方式は従来

⋆1 VMware は VMware, Inc. の米国および各国での商標または登録商標です．
⋆2 Intel Core 2 Duo は，Intel Corporation の米国およびその他の国における登録商標です．

表 5　印刷時間の測定結果
Table 5　The result of print-out speed.

印刷ページ数	t1（秒）	t2（秒）	t3（秒）	t2 / (t1+t3)（%）
1	2.4	2.7	2.0	62
3	7.0	7.1	4.5	61
5	11.3	11.7	6.6	66

方式に比べて，フルテキスト取得に 0.1〜0.4 秒という短い時間で処理が完了している．さらに，印刷ページ数が増加していったとしても，フルテキスト取得時間も同様に増加することは認められなかった．よって，提案方式によるフルテキスト取得時間が印刷時間に比べて十分に小さいことが確認できた．

また印刷に要する時間は，従来方式では $t1 + t3$，提案方式では $t2$ で表される．この $t2$ は $t1 + t3$ に比べ，表 5 に示すように，61〜66%という高速性を達成している．

5.3 考　　察

文字コード処理方式におけるフルテキスト取得時間は，5.2 節で述べたように，印刷ページ数が増加していったとしても，フルテキスト取得時間も同様に増加することは認められなかった．これは，OS のマルチタスク処理により，最初の 1 ページの印刷に余計に時間がかかるだけで，次ページの印刷以降はフルテキスト取得時間が印刷全体時間に吸収されるためであると考えられる．

また文献 22) によると，レスポンスタイムとして利用者がシステムの反応が瞬時に行われていると感じる限界は 0.1 秒であり，利用者の考えの流れが妨げられない限界は 1.0 秒であるといわれている．従来方式では，表 5 によるとフルテキストの取得に数秒以上のレスポンスタイムの増加が見込まれることから，利用者が印刷が遅くなることに気づくと考えられる．一方，提案方式では，5.2 節で述べたようにフルテキストの取得に 0.1〜0.4 秒ほど増加するだけであり，そのため利用者にとって印刷が遅くなることにほとんど気づかないと考えられる．

6. おわりに

本稿では，印刷コントロールに関する DLP 機構の実現に向け，文書の識別を高速化するために，従来方式の画像処理方式とは異なるアプローチで，プリンタドライバにおける処理で文字情報を取得する文字コード処理方式を考案した．さらに，文字コード処理方式を実現するプリンタドライバを Windows XP 上で実装し，そのフィジビリティ検証および印刷速度への影響を評価し，実用上，問題のないことを確認した．

提案方式である文字コード処理方式は，既存オフィス環境に導入するうえで，プリンタやプリントサーバを変更・追加設置することなく，プリンタドライバだけで文書の識別を行うために，SOHO（Small Office/Home Office）といった小企業でもオフィスの情報漏洩対策として容易に導入できるセキュリティ対策である．

今後の課題は，下記に示すとおりである．

- 文字コード処理方式で取得したフルテキストを使いつつ，実際の業務で使われる各種文書に対応した文書の識別手法を確立すること
- 印刷コントロールとして，印刷禁止，アラート，ログ取得，出力プリンタ切替え，電子透かし挿入などとの連携を図ること

参　考　文　献

1) 財団法人ニューオフィス推進協議会：オフィスセキュリティマーク認証基準（Ver3.0）．http://www.nopa.or.jp/security/pdf/osm_v3.pdf (参照 2010-04-07)
2) Lange, M.C.S. and Nimsger, K.M.: *Electronic Evidence and Discovery: What Every Lawyer Should Know*, ABA Publishing (2004).
3) NPO 日本ネットワークセキュリティ協会：2008 年情報セキュリティインシデントに関する調査報告書．http://www.jnsa.org/result/2008/surv/incident/index.html (参照 2010-04-07)
4) Microsoft: Information Rights Management (IRM). http://office.microsoft.com/ja-jp/help/FX010937471041.aspx (参照 2010-04-07)
5) Adobe: Adobe LiveCycle Rights Management ES2. http://www.adobe.com/jp/products/livecycle/rightsmanagement/ (参照 2010-04-07)
6) CodePlex: Word 2007 Redaction Tool. http://redaction.codeplex.com/Wikipage (参照 2010-04-07)
7) 富士通：世界初！紙と電子データの暗号化技術の開発に成功—高いセキュリティを確保しながら情報共有が可能に．http://pr.fujitsu.com/jp/news/2008/06/10.html (参照 2010-04-08)
8) NRI セキュアテクノロジーズ：SecureCube/Labeling. http://www.nri-secure.co.jp/service/cube/labeling.html (参照 2010-04-07)
9) 富士ゼロックス：紙文書と電子文書を一つのセキュリティ環境下で管理できる初めての技術を開発．http://www.fujixerox.co.jp/company/news/release/2008/0512_security.html (参照 2010-04-08)
10) 特定非営利活動法人デジタル・フォレンジック研究会：デジタル・フォレンジック辞典 (2006).

11) 日経コンピュータ：重要ファイルだけを「社外秘」に中身に潜むキーワードで判定，2010 年 04 月 28 日号.
12) 日経 NETWORK：なぜ今，日本に上陸するのか？ 情報漏えい対策の DLP 製品が続々発売，2010 年 06 月号.
13) Brin, S., Davis, J. and Garcia-Molina, H.: Copy Detection Mechanisms for Digital Documents, *Proc. ACM SIGMOD Annual Conference*, pp.398–409 (1995).
14) Shivakumar, N. and Garcia-Molina, H.: CAM: A Copy Detection Mechanism for Digital Documents, *Proc. 2nd Annual Conference on the Theory and Practice of Digital Libraries* (1995).
15) 特開 2006-261907：文字処理装置，文字処理方法及び記録媒体 (2006).
16) 富士キメラ総研：2008 e ドキュメント市場マーケティング調査総覧 (2008).
17) 社団法人電子情報技術産業協会：用語の解説. http://it.jeita.or.jp/document/OCR/vocab.html (参照 2010-04-07)
18) 藤井康広，海老澤竜，本多義則，洲崎誠一：マルチベンダ紙文書漏えい対策システムの一提案，電子情報通信学会技術研究報告 ISEC，情報セキュリティ，Vol.108, No.161, pp.51–58 (2008).
19) Microsoft: Windows Driver Kit: Print Devices, Rendering a Print Job. http://msdn.microsoft.com/en-us/library/ff561943.aspx (参照 2010-04-07)
20) 茶筌．http://chasen-legacy.sourceforge.jp/ (参照 2010-04-07)
21) 日本の苗字 7000 傑．http://www.myj7000.jp-biz.net/ (参照 2010-04-07)
22) ヤコブ・ニールセン：ユーザビリティエンジニアリング原論，東京電機大学出版局 (1999).

(平成 22 年 5 月 15 日受付)
(平成 22 年 10 月 4 日採録)

甲斐　賢（正会員）

1998 年京都大学大学院理学研究科修士課程修了．同年（株）日立製作所入社．システム開発研究所にてネットワークセキュリティ，コンピュータセキュリティ，情報セキュリティ，デジタルフォレンジックの研究開発に従事．情報処理学会第 60 回全国大会大会奨励賞受賞．日本セキュリティ・マネジメント学会 09 年度論文賞受賞．

笈川　光浩（正会員）

1998 年上智大学大学院理工学研究科電気・電子工学専攻博士前期課程修了．同年株式会社日立製作所入社．以来，同社システム開発研究所において，セキュリティ技術，特に電子認証の研究開発に従事．

伊川　宏美（正会員）

2004 年筑波大学大学院システム情報工学研究科博士前期課程修了．同年（株）日立製作所入社．システム開発研究所にて情報セキュリティに関する研究開発に従事．日本セキュリティマネジメント学会会員．

今一　修

1969 年生．1993 年京都大学工学部電気工学第二学科卒業．1995 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．1998 年同博士後期課程修了．博士（工学）．同年（株）日立製作所入社．中央研究所にて，自然言語処理，情報検索の研究開発に従事．言語処理学会，人工知能学会各会員．

森本　康嗣（正会員）

1988 年名古屋大学大学院工学研究科修士課程修了．同年（株）日立製作所入社．システム開発研究所，中央研究所にて機械翻訳，情報検索の研究に従事．言語処理学会会員．

土田　健一

1998 年株式会社日立東サービスエンジニアリング入社．1996 年日立アイ・エヌ・エスソフトウェア株式会社転属．ネットワークシステム，コンピュータセキュリティ，情報セキュリティ，WEB 基盤の製品開発に従事．

手塚　　悟（正会員）

2009 年度より東京工科大学コンピュータサイエンス学部の教授，現在に至る．1984 年度より（株）日立製作所入社．マイクロエレクトロニクス機器開発研究所に勤務し，パーソナルコンピュータのオペレーティングシステム，デバイスドライバ，LAN システム等の研究開発に従事．その後，システム開発研究所に勤務し，パーソナルコンピュータを中心とした LAN システムの構築・運用管理の研究開発，さらに電子政府，電子自治体等を主に情報セキュリティシステムの研究開発に従事．特に，PKI 技術を用いた電子署名，電子認証等の研究．慶応義塾大学理工学部特別講師，大阪大学非常勤講師等歴任．2004 年度情報処理学会論文賞，2008 年度情報処理学会論文賞，IEEE-IIHMSP2006 Best Paper Award．工学博士．著書に『Inside CORBA』（共訳，アスキー出版，1998 年），『インターネットコマース新動向と技術』（共著，共立出版，2000 年），『インターネット時代の情報セキュリティー暗号と電子透かし』（共著，共立出版，2000 年）．

荒井　正人（正会員）

1992 年日本大学大学院理工学研究科博士前期課程修了．同年（株）日立製作所入社．システム開発研究所にてネットワークシステム，セキュリティ技術等の研究開発に従事し，製品化に貢献．現在，研究開発本部研究戦略統括センタ所属．博士（情報学）．

洲崎　誠一（正会員）

1991 年 3 月横浜国立大学電子情報工学科卒業．同年 4 月（株）日立製作所システム開発研究所に入所．以来，情報セキュリティ技術の研究開発に従事．工学博士．