非構造化文書のライフサイクル管理の研究

－収集・分析・セキュリティ関連技術の開発

平成２３年３月

松　田　純　一

要旨

少子高齢化時代の日本の競争力強化には，量ではなく質で勝負することが必要であり，情報の戦略的活用が有効である。そのためには，従来よりも多くの情報を得ることが大前提である。有益な情報は，個々人のメモ書きや文書管理ソフトウエアなどの非構造化情報としてパーソナルコンピュータに蓄えられているだけで，有効に活用されていないことが多い。また，グローバル化が進む中で，必要な情報が日本語で存在するとは限らず，英語や中国語などで記述された外国語情報が，今後，ますます増えてくることが予想される。充分な情報が得られたら，活用できるように分析し，有益な知見を得ることが重要である。構造化情報は，従来からある程度の分析ができているが，非構造化情報は形式も様々であり，分析手法が確立されておらず，構造化情報の分析に比べると有益な知見を得ることが難しい。一方，より多くの情報が得られるようになると漏えいや不正利用のリスクが高まる。電子的な情報のセキュリティ対策は，従来から，認証技術や改ざん検知技術等が提案されてきているが，印刷して紙媒体になった後の情報のセキュリティ対策は，必ずしも充分とはいえない状況である。

組織で発生する膨大な情報を電子的に管理することは，電子文書管理と呼ばれる。文書管理とは，組織内で発生する様々な文書の蓄積，収集，検索，分析，印刷のライフサイクル管理であり，これによって，組織内で情報を共有し，戦略的に活用することが期待できる。本論文は，文書のライフサイクル管理における収集，分析，印刷の各フェーズにおける以下の３つの課題を解決することを目的とした。

(1) 外国語文書の収集

グローバル化の進展に伴い，外国語で書かれた情報の重要度が増している。より多くの情報を収集し，活用するためには，今後，外国語を含んだ文書を収集し，共有できることが必要である。外国語の中でも，特に，近年，経済や文化の面で交流が飛躍的に伸びているアジア圏の言語が重要である。

(2) 非構造化文書の分析

従来から，OLAP (Online Analytical Processing) を主体とする構造化文書向けの分析手法が使用されてきたが，多くの情報が非構造化文書として未活用のまま眠っていると言われている。

大量の非構造化文書中のテキストを対象とした新たな分析手法が必要である。特に，企業，団体，官公庁等の組織では，戦略立案に有効な地理的，時間的な観点での分析ニーズが高い。

(3) 印刷文書の管理

検索，分析結果を活用する際に，印刷することが多いが，セキュリティ面から情報の管理が必要である。最近では，情報漏えいの媒体として，紙文書が大多数(72.6%)を占めており課題となっている。情報漏えい抑止や真正性保証，トレーサビリティ管理のための対策が必要である。これらの３つの課題に対して，本論文では，以下の解決手法を提案した。

(ｉ)外国語文書の収集

外国語の文書の収集・共有に関しては，機械翻訳技術の活用が有効である。従来から，日本語－英語間の機械翻訳が研究されており，同じ方式でアジア圏の言語も機械翻訳することが試みられていた。本論文では，アジア圏の言語と日本語の部分的な類似性を反映できる新たな翻訳方式を提案した。具体的には，中国語を例にとり，トランスファ方式をベースに３段階のトランスファレベルを設ける翻訳方式を提案した。これにより，翻訳規則を簡素化して効率的に翻訳することが可能となる。

(ii)非構造化文書の分析

非構造化文書の分析に関しては，大量のテキストデータから固有表現を自動抽出して，データマイニングおよび統計処理を行う方式を提案した。本方式は，ニーズの高い時空間分析に有効な手法として，連続値である時間を離散化して分析する手法や平均距離を用いた空間分析手法に特徴がある。具体的な事例として，約５０００件の防犯メールデータを使用し，事件種別，場所，時間，等の固有表現を自動抽出し，事件と時間の関係ルールを導いたり，ランドマークとの距離と事件との関係性を見出すことができた。提案手法を用いることにより，分析者によるばらつきを防ぎ，大量のデータを短時間で分析することが可能になると期待される。

(iii)印刷文書の管理

印刷文書のセキュリティ対策としては，印刷文書の中に管理情報を埋め込むことが有効である。本論文では，モノクロ２値画像を対象に、地紋を使ってデータを埋め込む方式を採用し，印刷文書の内容に影響を与えることの少ないイラストを地紋に用いて，利用者がデータの埋め込まれた印刷文書から受ける違和感を軽減する手法を提案した。また，印刷文書からのデータ

抽出精度を向上させるための，誤り訂正符号を提案した。評価の結果，１０００文字程度の文字が記載されたＡ４用紙に10～20Byteの情報を違和感なく埋め込めることが分かった。

以上の提案により，組織内での情報共有が進み，戦略立案のための有益な情報が集約されることから，組織員のスムーズな業務遂行や組織としての競争力強化に貢献できるものと考えている。

Abstract

Under the aging society with fewer children, in order to strengthen international competitiveness, it is effective to improve individual skills by strategically utilizing information. To utilize information, it is necessary to get as much information as possible. Useful information tends to be stored away in personal computers as unstructured information such as "office suite" documents and text documents. Besides, in a global era, information written in foreign languages such as English, Chinese and so on will increase year by year. After getting enough information, the information should be analyzed and useful result should be obtained. Analysis of unstructured information is difficult as compared with structured information. On the other hand, much information increases the risks of information leakage or illegal usage. Although security technologies for electronic documents have been proposed, security technologies for paper documents are not sufficient.

Large amount of information generated in organizations is usually managed by using electronic document management system. Document management is document lifecycle management which consists of storing, gathering, retrieving, analyzing, printing etc. By using electronic document management system, it is expected to share and strategically use information. The purpose of this paper is to resolve the following three problems concerning gathering, analyzing and printing of the lifecycle phase.

(1) Gathering documents written in foreign languages

Information written in foreign languages becomes important in a global era. It is necessary to gather and share documents written in foreign languages. Recently Asian languages become comparatively important because communication with Asian countries is growing rapidly.

(2) Analyzing unstructured documents

While various analysis methods such as OLAP (Online Analytical Processing) for structured documents have been proposed, it is said that lots of unstructured documents

are not analyzed. New method for analyzing large amount of unstructured text data is necessary. For organizations, temporal and spatial analysis is effective for making strategies.

(3) Managing printed documents

Documents are often printed when using search and analysis result. Security management is necessary for printed documents. Recently, as more than seventy percent of information leakage incidents occur via paper documents, security management of paper documents is important.

This paper proposes the following solutions for the above mentioned problems.

(i) Gathering documents written in foreign languages

To gather and share documents written in foreign languages, it is effective to use machine translation technology. Japanese-English machine translation method had been proposed and applied to Asian languages. This paper proposed a new method which uses the similarity between Japanese and Asian languages. The method has three transfer levels according to the syntactic similarity. By two thousands sentence experiment for Japanese-Chinese translation, the method could simplify translation rules and get satisfied translation result.

(ii) Analyzing unstructured documents

This paper proposes the analysis method for large amount of unstructured text documents. The method is divided into two steps. The first step is to extract named entities such as temporal and spatial information which are effective for analysis from text and to transform them into structured documents. The second step is to analyze the structured documents. Two types of analysis method are used. One is using data mining technologies such as association rules and decision trees. A method for mining continuous values is proposed. Another one is using statistical analysis using average distance. The method is applied to about five thousands crime data in Tokyo Metropolitan area. Temporal

correlation between crime and time and spatial correlation between crime and landmarks are clarified. The method is expected to automatically analyze large amount of data.

(iii) Managing printed documents

This paper proposes a method for embedding information into monochrome printed documents for security management. The method is using dot pattern watermarking based on illustration so that impression of documents is natural. Since random dots are used for conventional dot pattern, documents with dot pattern often give strange impression to readers. The method is to use dot pattern image transformed from illustration image. Two key techniques are developed; one is a method for transforming illustration image to dot pattern image suitable for watermarking; another one is a method for embedding information that can reduce noise caused by errors when printing and scanning documents. By subjective experiments, it is possible to embed 10-20 byte information without strange impression in A4 size paper on which about one thousand characters are printed.

By the above mentioned proposal, it is possible that information sharing in organizations is promoted and valuable information is totally managed. The proposal is expected to contribute to running business smoothly and strengthening international competitiveness.

目次

第１章　序論

１．１　研究の目的

少子高齢化時代の日本の競争力強化には，量ではなく個人の質で勝負することが必要である。この対策として重要な要素の１つが，情報の戦略的活用である。ナレッジマネジメントや情報の共有・分析の仕掛けを組織内で確立させることにより個人の価値を向上させることが可能である[1] [2]。そのためには，情報を持つ企業・団体・官公庁等の組織において，様々な工夫をする必要がある。

まず，情報を活用するには，できるだけ多くの情報を得ることが大前提である。情報システムとしてデータベースに蓄積された構造化情報は，すでに適切に収集や共有がなされている場合が多いが，有益な情報は，個々人のメモ書きやワード，エクセルなどの非構造化情報としてパーソナルコンピュータ（ＰＣ）に蓄えられているだけで有効に活用されていないことが多い。また，グローバル化が進む中で，必要な情報が日本語で存在するとは限らず，英語や中国語などの外国語情報が，今後，ますます増えてくることが予想される。

充分な情報が得られたら，次にやるべきことは，活用できるように分析し，有益な知見を得ることである。構造化情報は，ＯＬＡＰ(OnLine Analytical Processing)や統計，データマイニングの手法により，従来からある程度の分析ができているが，非構造化情報は形式も様々であり，分析手法が確立されておらず，構造化情報の分析に比べると扱いが難しい。

一方，より多くの情報が得られるようになると漏えいや不正利用のリスクが高まる。情報が多くなればなるほど，また，重要であればあるほど，情報の取り扱いに慎重にならなければならない。組織内で，適切なセキュリティポリシーを策定・運用することはもちろんであるが，技術的にも対策を取るべきである。従来から，電子的な情報のセキュリティ対策は，認証技術や改ざん検知技術等が提案されてきているが，印刷して紙媒体になった後のセキュリティ対策は，必ずしも充分とはいえない状況である。

情報は，文書，画像，音声など様々な媒体で表現されるが，本研究では，文字によって記録された文書情報を対象とする。文書情報は，紙に記載される場合も電子的に記録される場合もある。また，文書中の客観的な事実を文字や数値で表したものをデータと定義する。

組織で生成される膨大な文書を電子的に管理することは，電子文書管理（Electronic

Document Management：ＥＤＭ）と呼ばれる。文書管理とは「組織の目的に適った文書の生成/格納/構成/転送/検索/操作/更新/廃棄」であり，ＥＤＭによって人々のコミュニケーションの改善が期待できる[3]。言い換えると，文書管理とは，図1-1に示すような文書のライフサイクルを管理することである。

個別に蓄積された文書は，収集され，統合的に検索・分析が行われる。検索・分析結果は表示，印刷した上で活用する。活用した結果は，新しい文書を生成し，再活用されるサイクルが確立される。

グローバル企業を例に，文書のライフサイクルの実現イメージを図1-2に示す。世界各地に拠点を持つグローバル企業では，各拠点で蓄積された文書を本社で一元的に収集し，管理することが多い。海外拠点からの文書は外国語で書かれていることもあり，その場合，必要に応じて日本語に翻訳している。収集した文書は統合データベース（統合ＤＢ）で一元的に管理され，検索・分析を通じて新たな知見を得た上で，その結果を表示，印刷して企業活動で使用する。その成果が，新たな経営戦略，営業戦略となる。戦略が生かされたかどうかは，新たに生成される文書を収集，検索，分析してみれば分かる。これが文書のライフサイクルである。ライフサイクルをいかにうまく管理し，企業価値向上に資することができるかが重要である[4] [5]。

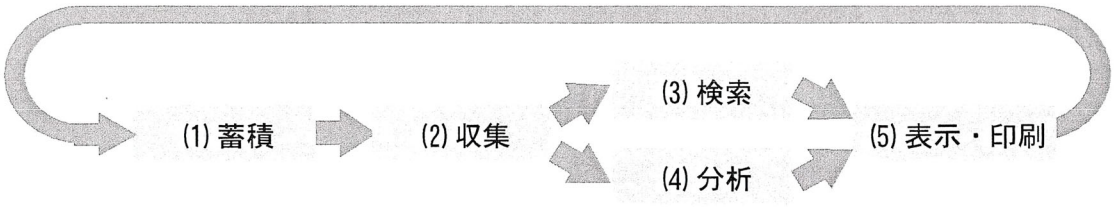


図1-1　文書のライフサイクル

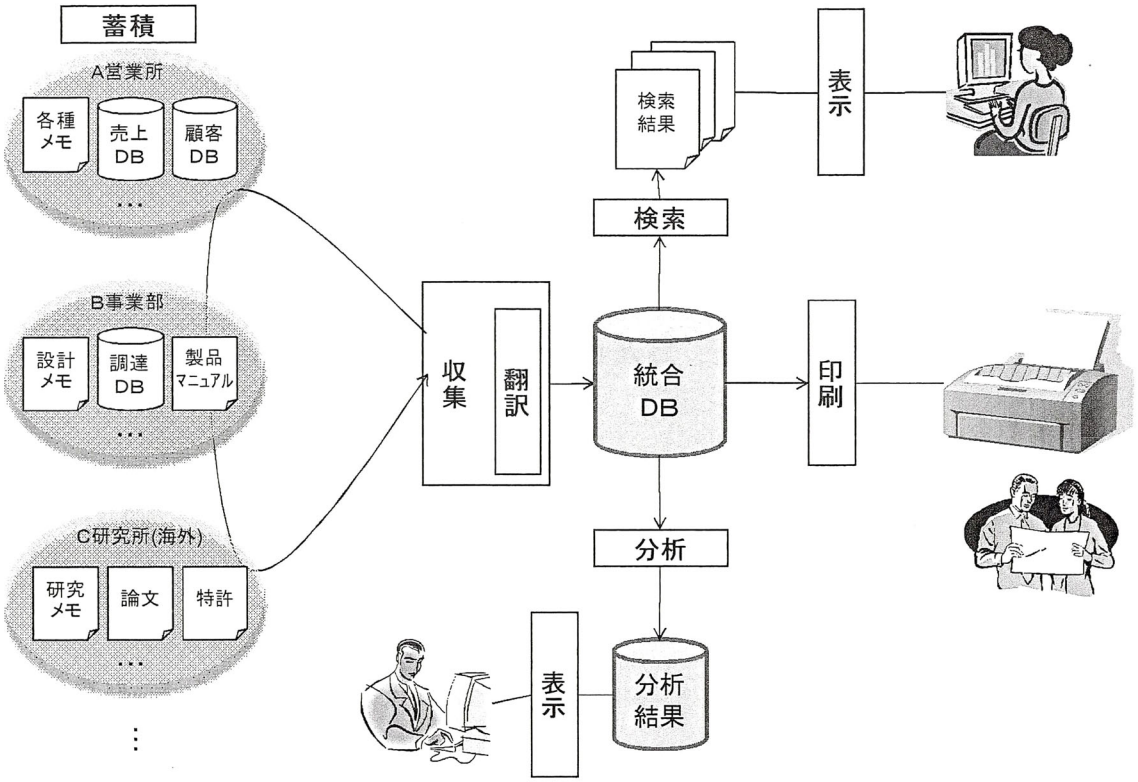


図 1- 2　文書のライフサイクルの実現イメージ

文書のライフサイクルの各フェーズの概要は以下のとおりである [6] [7] [8]

(1) 蓄積

文書が各部署で生成，格納された段階である。文書は分散された状態にある。その形態は，電子，紙のいずれの場合もありえる。文書を活用するには，文書が収集，検索できる形になっていることが必要である。個人のＰＣに眠っているだけでは共有できないため，部署内に共有サーバを設置して，組織内で作成した文書は常に共有サーバに蓄えられるようにすべきである。また，紙文書は，スキャンイメージを共有することが有効であり，検索・分析ができるように文字認識によるテキスト化を行うことが望ましい。

(2) 収集

各部署に分散されている文書を中央 (本社，本部など) で一元管理できるようにする段階である。中央から収集ロボット (クローラー) が各部署の指定された共有サーバの文書を収集する。この際，文書の重要性や更新頻度等をもとに収集頻度をきめ細かく制御して効率的に収集する

必要がある。

(3) 検索

検索とは，所望の文書をすばやく見つけることである。インターネット検索サービスでも分かるように，検索結果には，不要なもの(ノイズ)が多く含まれていたり，逆に，所望の文書が入っていないことやランキングの後のほうに出てくることが少なくない。適切な文書を適切なランキングで出すことが重要である。

(4) 分析

検索が生の文書を１つずつ探し出すのに対し，大量の文書をもとに，①統計処理，②分かりやすく加工(たとえば，時間軸に沿って整理，地図上で整理，関係性のチャート作成)，③データマイニング(新たな知識の発見)，④予測，⑤異常値発見，等を行うのが分析である。分析結果が有効活用できるような様々な工夫が必要である。

(5) 表示，印刷

検索・分析結果を画面上で表示したり，紙に印刷したりして，人間に見える形で分かりやすく提示することである.

従来の組織内の文書管理は，ＤＢに蓄えられ構造化された文書の管理，ワークフローシステムによる文書の流れの制御が中心であり，いずれも，電子文書が対象であるが，もっと多くの文書をもっと多彩に活用する方向で検討がなされるべきである。具体的には，以下の点で，機能拡張のニーズが存在する。

(1) 対象となる文書の種類

より多くの文書を共有するには，構造化文書だけでなく，オフィス文書(Word®, Excel®，一太郎®，等)やテキスト，PDF 等の形式の非構造化文書も対象とするべきであるというニーズが強く，現在，ほとんどの文書管理システムでは，オフィス文書やテキスト，PDF 等の多様な文書形式を扱うことができる。また，紙文書のスキャンイメージを共有するために，文字認識，書式認識の機能がついた文書管理システムもある。さらに，グローバル化の進展に伴い，共有すべき対象として，外国語で書かれた文書の重要度が増している。外国語を含んだ文書も管理できることが望ましい。

(2) 分析の必要性

ＯＬＡＰを主体とする構造化文書向けの分析手法ではなく，非構造化文書を対象とした新たな分析手法が必要である。様々な形式の文書が集められることによる新たな分析手法の開発も期待される。特に，企業，団体，官公庁等の組織では，空間的，時間的な観点での分析ニーズが高い。

(3) 印刷文書の管理

検索・分析結果を活用する際に，セキュリティ面から文書の適切な管理，漏えい防止が必要である。電子的な文書のセキュリティ対策はこれまでいろいろな技術が実用化されているが，印刷された紙文書のセキュリティ対策に関する課題は多い。特に，漏えいの媒体としては，印刷文書がインシデント件数の大多数（７２．６％）を占めており[9]，印刷文書のセキュリティ対策は重要である。

本研究では，上記のニーズに鑑み，以下の課題を解決することを目的とする。

(1) 外国語で書かれた文書を共有すること

(2) 大量の非構造化文書の分析を行うこと

(3) 印刷された文書を適切に管理し，漏えい抑止や追跡に役立てること

１．２　研究の概況

本節では，まず，１．２．１において，文書のライフサイクル管理全体のこれまでの研究開発について概観する。このうち，本研究と関連の深い内容について，１．２．２，１．２．３，１．２．４で詳述し，本研究の範囲を明確にする。

１．２．１　電子文書管理

電子文書管理は，ＥＣＭ(Enterprise Contents Management)もしくは知識管理(Knowledge Management)と呼ばれることもあり，ＩＴ用語辞典の定義によれば，企業等の組織における文書の蓄積，管理，運用を統括的，包括的に行うための技術やシステムのことである[10]。ＥＣＭの世界的権威としてＥＣＭ技術の理解，導入，使用に関して主導的な役割を果たしているＡＩＩＭ(Association for Information and Image Management)では，「ＥＣＭとは，組織のプロセ

スに関連するコンテンツや文書を収集・管理・蓄積・保護・配布するための技術，ツール，手法」と定義している[11]。ＥＣＭソフトウェア製品は，文書の蓄積・管理・検索・配布等の基本機能を核として，紙媒体のデジタル化等の文書入力支援，デジタル複合機との連携等の機能を持っている。また，文書のバージョン/リビジョン管理，ワークフロー，プロセス管理等の統合的な機能を持っているものもある。近年では内部統制強化の流れにより社会的にも企業の文書管理能力を求める傾向があることからＥＣＭに対するニーズはますます高まっている[5]。

電子文書管理に関する研究内容は多岐に渡るが，以下，主な機能についての研究開発概況を述べる。

(1) 文書の蓄積，管理，配布

文書管理システムの基本機能であり，すでに研究段階を終え，数多くの製品が世の中に存在する。株式会社富士キメラ総研「２００８ ｅドキュメント市場マーケティング調査総覧」によれば，日本でシェアの高い製品として，富士ゼロックス，日立製作所，リコー，オープンテキスト，ＯＳＫの製品がある[12] [13] [14] [15] [16]。これらの製品では，文書のバージョン/リビジョン管理，ワークフロー，プロセス管理等の機能も備えているのが一般的である。外国語で書かれた文書を管理するための機械翻訳技術については，１．２．２で詳述する。

(2) 検索・分析

検索技術については，グーグルをはじめとするＷｅｂ検索技術が有名であるが，組織内の文書を対象とした検索はエンタープライズサーチと呼ばれ，分散された多様な形式の文書を検索する技術をベースとしている[17] [18]。すでに，世の中に多くの製品が出されており，継続的に機能拡張がなされている。株式会社ミック経済研究所「ミドルウェアパッケージソフトの市場展望（データ統合・コラボレーション編）２００９年度版」によれば，日本でシェアの高い製品として，ファストサーチ&トランスファ，日立製作所，ジャストシステム，オートノミー，アクセラテクノロジの製品がある[19] [20] [21] [22] [23]。これらの製品では，全文検索技術のほか，検索結果のランキング，検索結果のナビゲーション，アクセス管理等の技術が使用されている。

検索だけでなく，大量の文書から新たな知見を見つけ出す分析技術も重要になってきている。データベース等に蓄積された構造化文書の分析は，従来から，統計やデータマイニングの手法

が適用されてきているが，非構造化文書の分析については，まだまだ課題が多い。分析に関するの研究概況については，１．２．３で詳述する。

(3) 紙媒体のデジタル化

紙媒体を文書管理の対象として検索・分析できるようにするためには，文字認識を行うことが有効である。文字認識技術は，パターン認識の一分野として古くから研究が進められ，印刷文字に関しては認識率も高く実用化が進んでいる。一方，手書き文字認識や写真，映像等に含まれる文字の認識は認識率が不充分であり，精度向上のための研究開発が進行中である[24]。

(4) 印刷

組織内で管理している文書は，デジタル複合機等のプリンタと連携して，印刷した紙の形態で活用することが多い。電子文書のセキュリティや追跡技術は，暗号等の応用技術として従来から研究開発されているが，印刷文書の管理についてはまだまだ課題が多い。印刷文書の管理に関する研究概況は１．２．４で詳述する。

１．２．２　機械翻訳に関する研究概況

我が国において，外国語で書かれた文書を共有し，検索・分析を可能とするためには，日本語に変換することが望ましい。これに対応する技術として機械翻訳が従来より研究開発されている。機械翻訳は自然言語処理の代表的な応用技術である。機械翻訳の歴史は古く，１９３３年にロシア人が機械翻訳のアイデアを提案したのが始まりとされる。当初は，欧米を中心に，計算機と言語学の専門家により研究が進められ，一時期下火になったこともあるが，近年のグローバル化の進展に伴い，徐々に研究開発が盛んになってきた。日本でも，１９５５年に九州大学で研究が始まったのを皮切りに，電気試験所(現在の産業技術総合研究所)，京都大学，東京工業大学，東京大学，奈良先端科学技術大学院大学等で研究が盛んに行われている[25]。企業においても，１９８０年代から製品化が始まり，インターネットの普及とともに数多くの製品，サービスが提供されるようになってきている。

機械翻訳の方式は，大別すると，以下の３方式がある[26] [27]。

(1) トランスファ方式

原文を解析して構文木や意味表現を作成し，その構文木や意味表現を訳文のそれに変換して

訳文を生成する方式。言語対ごとに変換規則を設けることにより，きめ細かい翻訳を可能とする。

(2) 中間言語方式

原文を言語に依存しない中間言語に変換し，そこから訳文を一気に生成する方式。中間言語の仕様を決めることは難しいが，多言語の翻訳システムが実現しやすいという利点がある。

(3) 事例ベース方式

１９９０年代以降に出てきた新たな翻訳方式。膨大な対訳事例をもとに統計処理を応用して翻訳を行う。

機械翻訳の対象言語に関しては，欧米系の言語同士の翻訳が早くから進んでいたが，日本では，主に，日本語と英語の機械翻訳がトランスファ方式をベースに研究されてきた。しかしながら，日本とアジア諸国との経済，文化の結びつきが強まっていることから，日本語とアジア諸国の言語との機械翻訳の必要性が指摘されるようになり，通商産業省(現在の経済産業省)および(財)国際情報化協力センターが主体となって「近隣諸国間の機械翻訳システムに関する研究協力」が１９８７年から始められた[28]。対象言語は，中国語，インドネシア語，マレーシア語，タイ語である。翻訳方式は中間言語方式が採用されたが，日本語－英語間の翻訳で開発された中間言語方式を他の言語に適用したものであった。

アジア圏の言語は英語とは異なる文法構造を持つことから，日本語－英語間の翻訳方式をそのまま他の言語に当てはめただけでは最適な翻訳方式とはいえない。この考えに基づき，２章では，新たな翻訳方式を提案する。具体的には，近年，経済成長著しい中国で公用語となっている中国語を例に，トランスファ方式に基づいた日中翻訳方式について述べる。

なお，本論文で提案する手法の発表後，大量のディスクやメモリを安価に使えるようになったことから，精度の高い事例ベース翻訳が主流になってきた[29] [30]。また，インターネットの普及とあいまって多くの言語の翻訳サービス/製品が提供されている[31] [32] [33]。

１．２．３　非構造化文書の分析に関する研究概況

非構造化文書を対象とした分析の基盤技術として，テキストから定型項目を抽出する技術がある。定型項目としては，例えば，名前，住所，等の構造化文書によく現れる項目が挙げられ

る。名前，住所等は固有表現と呼ばれ，自然言語処理を用いてテキストから固有表現を抽出する技術が提案されている[34]。

テキストデータの分析に関しては，テキストマイニングと呼ばれる研究分野があり，自然言語処理の延長線上の技術として研究が進められている[35]。主に，テキストから抽出したキーワードに対して統計処理を行い，単語の頻度分析や相関ルール抽出，文書分類，時系列予測等を行うための技術である。

構造化文書の分析手法については，ＯＬＡＰだけではなく，大量のデータから有効な知識を発見するデータマイニング技術が１９９０年ごろから提唱されており，１９９４年の Agrawal の論文で，大規模データベースからの相関ルール発見アルゴリズムであるアプリオリが提唱された[36]。その後，実用化も盛んになってきたことから，ストリームデータやテキストデータ，Ｗｅｂデータ等，マイニングの対象範囲が拡張され，各種アルゴリズムも提唱されている[37]。これらのアルゴリズムは，離散データを対象に分析することが前提であるが，連続データである時間，空間データの分析も重要である。時間，空間データの分析も提案されているが，人手によるデータ編集作業が必要なため，適用事例としては数百件規模のデータにとどまっている[38]。３章では，大規模テキストデータにも対応できる時間，空間データの分析手法について提案する。

１．２．４　印刷文書の管理に関する研究概況

印刷文書の追跡や改ざん・偽造等に対する真贋判定を行うための手段として，文書画像にデータを埋め込む技術が提案されている。画像に各種のデータを埋め込む技術として，電子透かし技術があり，カラー画像へのデータ埋め込み手法や２値画像へのデータ埋め込み手法が提案されている[39] [40] [41]。しかし，これらは，データの埋め込みから取り出しに至るまでのサイクルにおいて，常に電子データのままであり続けることを想定しており，印刷原稿として一旦出力し，そこからデータを取り出すことは想定されていない。

一方，印刷物にした際にもデータ保持が可能な技術として，カラー画像にデータを埋め込むステガノグラフィーがあるが，モノクロ２値画像には適用できない[42]。モノクロ２値画像にデータを埋め込む技術としては，プリンタのトナーが赤外線へ反応する性質を利用したもの

[43]や，紙面の背景に微細な点を配置することでデータを表現する手法[44]が開発されているが，これらの方法により作成される点群や図形自体には意味がなく，利用者に違和感を与えたり，文章が読み難くなることがあると考えられる。

４章では，イラストを地紋に用い，その中にデータを埋め込む地紋透かしを用いることにより，データの埋め込まれた印刷文書から利用者が受ける違和感を軽減する手法を提案する。なお，組織内での文書はモノクロ２値画像が圧倒的に多いため，本論文では，データを埋め込む対象をモノクロ２値画像とする。

１．３　本論文の概要

本論文では，電子文書管理に有効な要素技術として，多言語処理，非構造化文書の分析，印刷文書へのデータ埋め込みの３つについて述べる。

２章では，外国語を含んだ文書を日本語文書と同様に管理するために必要となる機械翻訳技術について述べる。従来から，英語－日本語間の翻訳技術は知られていたが，英語とは言語構造の異なるアジア諸国の言語で効率的な翻訳方式が実現可能であることを述べる。

３章では，非構造化文書を分析する手法について述べる。従来からデータベースに格納された構造化文書の分析のための統計分析やデータマイニングの技術が研究されてきていた。一方，非構造化文書の分析については，自然言語処理技術に基づく固有表現抽出やテキストマイニングの技術が提案されてきたが，大規模な非構造化データを対象とした応用の事例が少なかったこともあり，どのような場面でどのように役立つのかが明確でなかった。本論文では，防犯メールデータを例にとり，時間および空間的な分析の手法と応用事例について述べる。

４章では，２値画像印刷文書にデータを埋め込む手法について述べる。従来から，紙面の背景に透かしを入れる技術が提案されていたが，透かしを入れることによって利用者に違和感を与えたり，文章が読み難くなることがないように埋め込む方式について提案する。

最後に，５章で，全体のまとめと今後の課題について整理する。

文書管理システムと本論文で提案する技術との対応関係を図 1- 3 に示す。

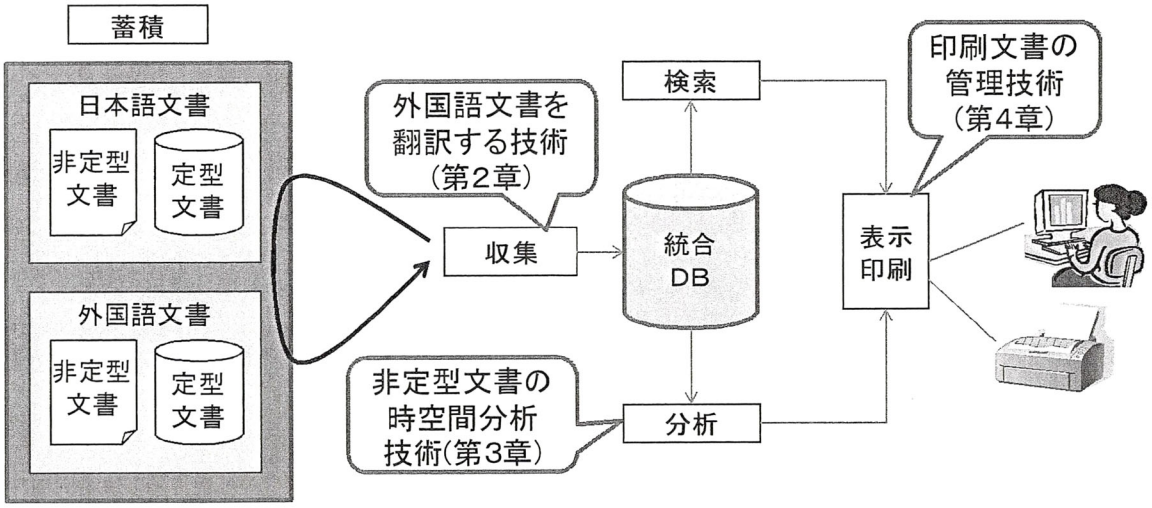


図 1- 3　文書管理システムと提案技術との関係

第２章　多言語情報の収集と翻訳

２．１　緒言

グローバル化が進む中で，外国語で書かれた文書を共有の対象とすることは，多くの組織で必要不可欠になりつつある。このための有効な手段として，機械翻訳技術がある。機械翻訳により外国語文書を日本語に翻訳しておけば，日本語と同じように検索・分析が可能となる。また，日本語で検索システムを使用することも可能となる。機械翻訳を用いた文書管理システムの利用イメージを図 2- 1 に示す。

機械翻訳技術は，従来から，欧米系の言語である英語を中心としてトランスファ方式や中間言語方式が研究開発されてきた。この方式は，日本語と言語構造がまったく異なる英語の特性を考慮した方式である[25]。

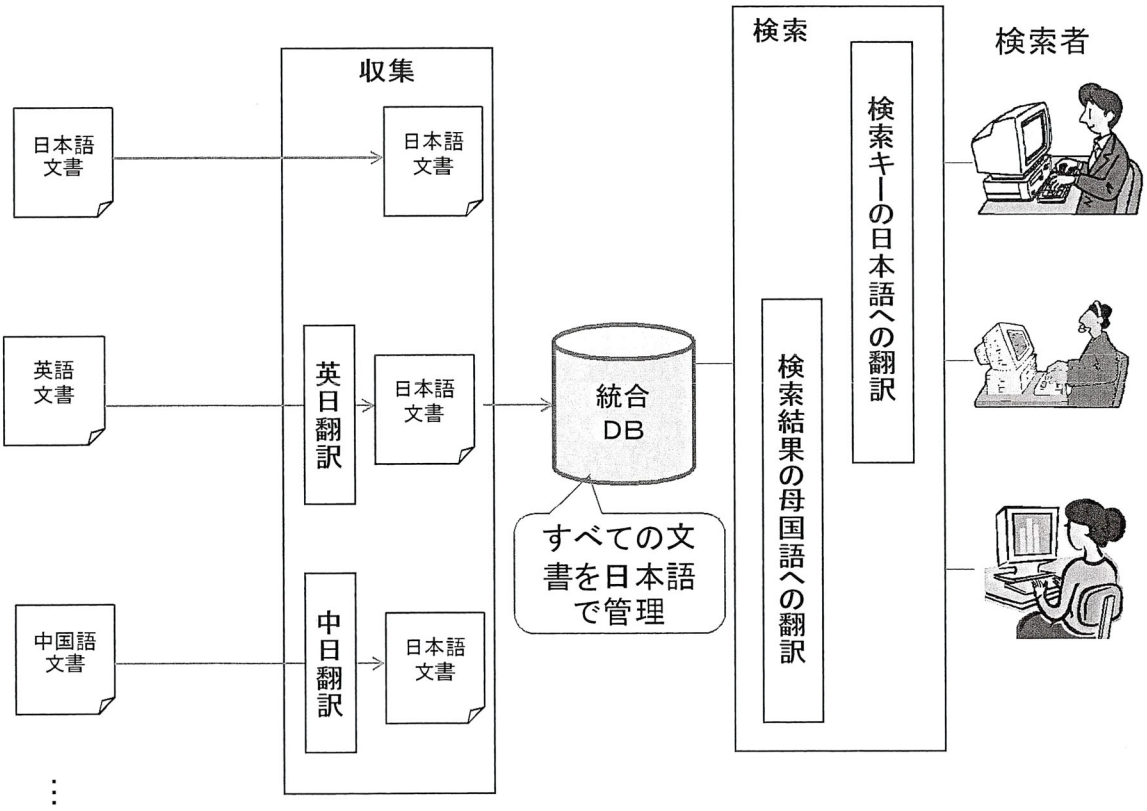


図 2- 1　機械翻訳による文書管理システムの利用イメージ

一方，近年，アジア圏との経済文化交流が盛んになるにつれ，アジア圏の言語の重要性が増している。また，インターネットの普及につれ，英語以外の言語で記述されたコンテンツも増えている。外国語といえば英語であった時代は，英語以外の言語も，英語を介在して翻訳すれば何とかなるとみなされてきた。実際，１９９５年頃まではインターネットで用いられている言語のうち英語が占める割合は８０％以上と言われていた[45]。しかしながら，英語以外の言語のコンテンツが増えてくると，日本語と該言語語との間でダイレクトに翻訳するほうが効率的である[46] [47]。

図 2- 2に２０００年と２０１０年での言語別のインターネットユーザ数の推移を示す[48]。中国語やスペイン語等の伸びが著しいことが分かる。

実際，図 2-3 に示すように，２０００年と２０１０年での言語別インターネット人口の比率を見ても，英語の比率が下がっていることが分かる[48]。

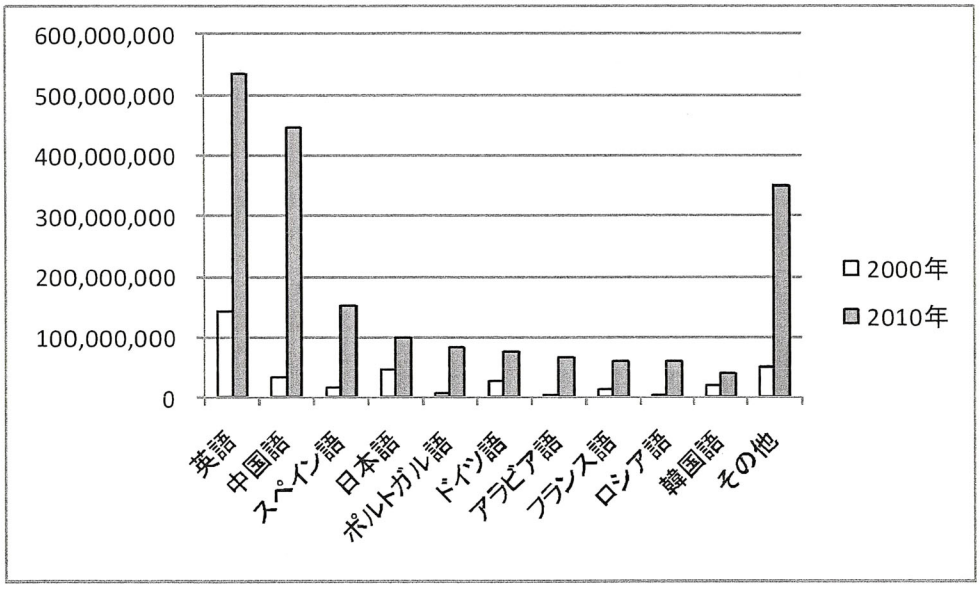


図 2- 2　言語別インターネットユーザ数[48]

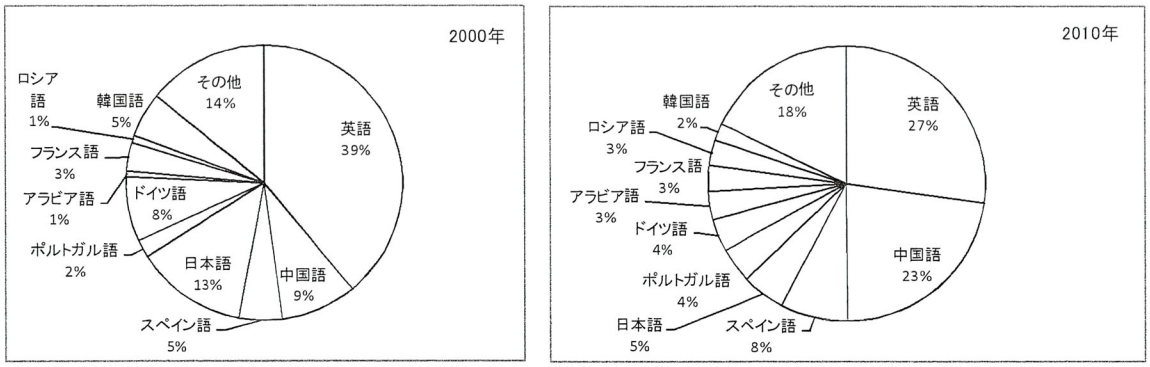


図 2- 3　言語別インターネット人口比率 [48]

財務省の貿易統計でも，アジア諸国の重要性が増していることが分かる。図 2- 4 は，日米間の貿易額の推移と全貿易に占める割合を示したものである。米国の占める割合が低下していることが分かる。

これに対して，日本と中国の関係は年々深くなっている。図 2- 5 は日中間の貿易額の推移と全貿易に占める割合を示したものである。貿易額も全貿易に占める中国の割合も飛躍的に伸びていることが分かる。２００７年からは，米国を抜き，中国が日本の最大貿易相手国になっている。

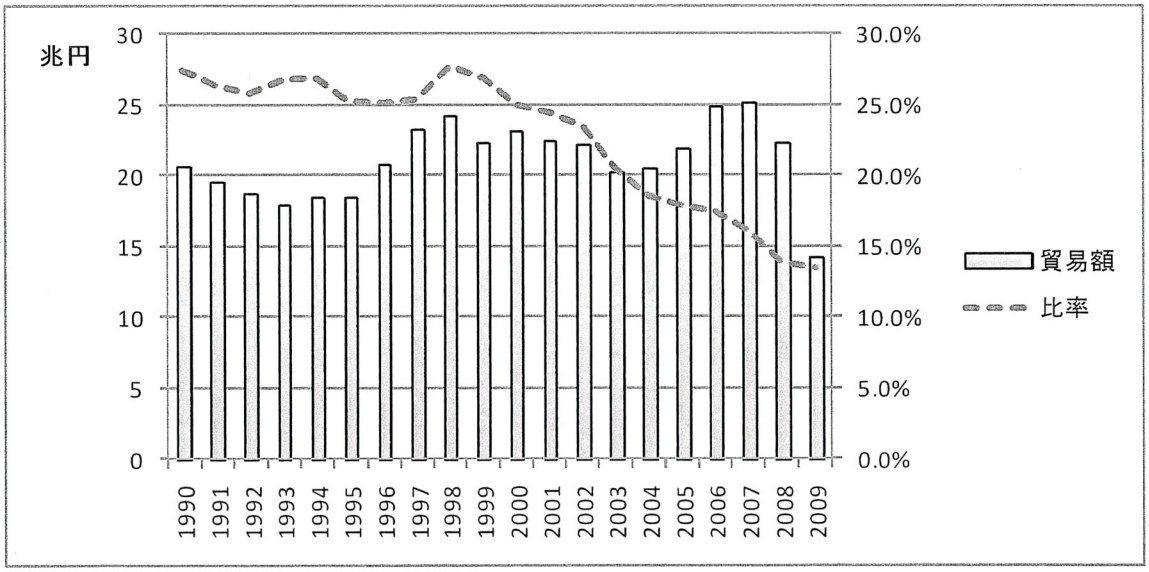


図 2- 4　日本と米国の貿易推移 (出典：財務省貿易統計)

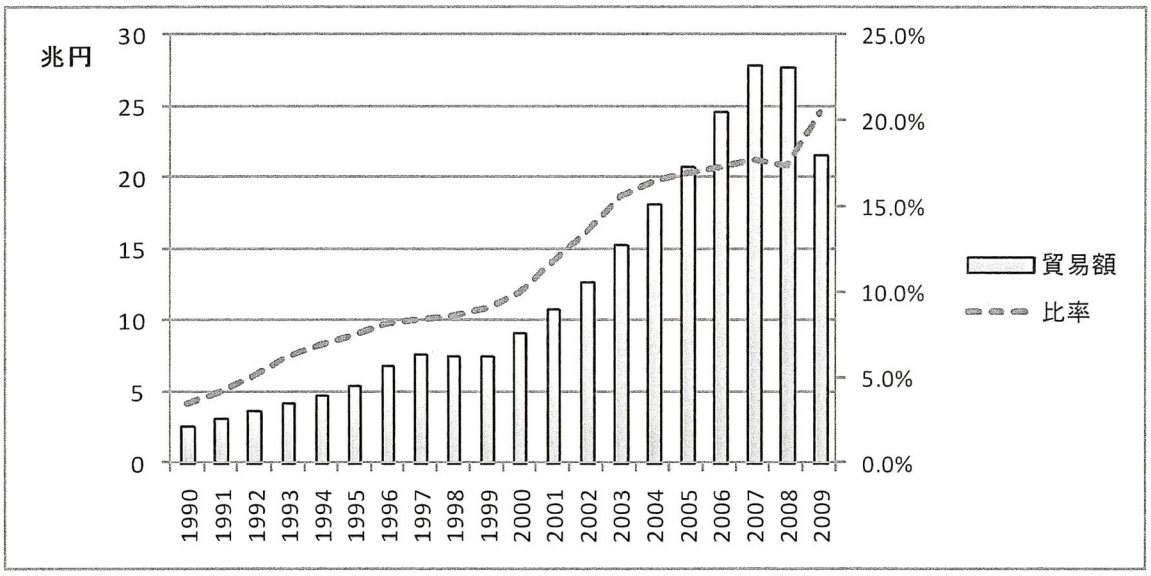


図 2- 5　日本と中国の貿易推移 (出典：財務省貿易統計)

以上のことから，インターネット，貿易のいずれの分野でも，英語以外の言語の比率が増えていると判断することができる。特に中国の伸びが著しい。

機械翻訳に関しては，従来から，日本語－英語間の翻訳技術が研究開発されていたが，１９９９年時点で，英語以外の言語との機械翻訳は，(財) 国際情報化協力センターが主体となって１９８７年から始められた「近隣諸国間の機械翻訳システムに関する研究協力」がほぼ唯一であった[28]。対象言語は，中国語，インドネシア語，マレーシア語，タイ語であるが，翻訳方式は日本語－英語間の翻訳で開発された中間言語方式をこれらの言語にそのまま適用したものであった。

アジア圏の言語は英語とは異なる文法構造を持つことから，日本語－英語間の翻訳方式をそのまま他の言語に当てはめただけでは最適な翻訳方式とはいえない。この考えに基づき，本章では，新たな翻訳方式を提案する。具体的には，近年，経済成長著しい中国で公用語となっている中国語を例に，トランスファ方式に基づいた日中翻訳方式について述べる。なお，ここでは，日本語から中国語への翻訳方式について述べるが，中国語から日本語への翻訳方式も考え方は同じである。また，本章で用いる中国語例文の構文と単語の意味は付録Ａで説明する。

２．２　入出力方式

１９９０年代当初，非英語圏で使用されているＰＣやワークステーションは，英語と母国語しか扱うことができなかったため，日本語と中国語を同時に表示させることが難しかった。しかしながら，１９９３年に，Unicode 体系に従った複数バイトの文字コードであるISO/IEC10646-1 が標準化され，複数バイトコード系の複数言語を扱える環境ができた。また，Unicode をサポートしたオペレーティングシステムが開発されつつあり，この問題は解決の方向に向かっていた[49]。実際，現在では，Windows をはじめほとんどのオペレーティングシステムはUnicode ベースとなっており，世界各国の言語を混在表示させることが可能になっている。

２．３　翻訳方式

２．３．１　日本語と中国語の言語構造

日本語はウラル＝アルタイ語族，中国語はシナ＝チベット語族という異なる言語族に属しており，表2- 1に示すように言語構造が異なる[50]。

一方，日本語も中国語も漢字を使っており，また，英語にはない以下の類似性がある。

(1) 名詞を修飾する句は名詞の前に置かれる。英語の場合は，名詞の後に置かれるため語順が変わる。

例１：　日本語：私の読んだ本

　　中国語：我读的书

表2- 1　日本語と中国語の言語構造

言語 / 項目	中国語	日本語
目的語の位置	動詞の後	動詞の前
前置詞/後置詞	前置詞	後置詞(助詞)
副詞の位置	動詞の前/動詞の後*1	動詞の前

*1：通常は動詞の前だが，頻度を表す副詞は動詞の後に置かれる

(2) 量詞を用いる。量詞自体は変わるが，日中間ではそのままの語順で翻訳できる。英語の場合は，「個」「人」のように一般に量詞がない。

例２：　日本語：５枚の紙

中国語：五张纸

(3) 方向名詞を用いる。英語の場合は，前置詞句等を用いて表現するので語順が変わることがあるが，中国語の場合は，そのままの語順で翻訳できる。

例３：　日本語：机の上の本

中国語：桌子上的书

(4) 動詞句の連接が様々な意味を持つ場合でも，日本語と中国語の構文に差異がなければ意味解析を行わずに翻訳できる。英語に翻訳する場合は，文脈やニュアンスを考慮する必要があり訳を１つに決めることが難しい。

例４：　日本語：彼は座って本を読んでいる

中国語：他坐读书

(5) 様々な意味を持つ助動詞も，日本語と中国語が一対一の対応があり，意味を考慮することなく翻訳できる。英語への翻訳は意味を考える必要があり難しい。

例５：　日本語：持ってくる

中国語：拿来

(6) 語順に自由度がある。英語は日本語，中国語に比べると自由度がない。語順を入れ替えた以下の２文は同じ意味である。

例６：　日本語：北京で私は彼に会った

中国語：在北京我见过他

例７：　日本語：私は北京で彼に会った

中国語：我在北京见过他

２．３．２　翻訳方式

日本語と中国語の構文構造は部分的に類似しているため，翻訳方式はトランスファ方式をベースとした。構文構造が類似していれば，解析のレベルは低くても構わないので，日本語の構

文構造をそのまま中国語の構文構造に変換することにより効率的な翻訳が可能である。中国語は豊富な熟語表現を持っているが，日本語の熟語表現も単語列をそのまま中国語の単語列に変換すれば翻訳できることも少なくない。そこで，トランファレベルを単語レベル，構文レベル，意味レベルの３段階に分ける方式を考案した。開発した翻訳システムは，３つのトランスファレベルを持ち，入力文に応じて，適切なトランスファレベルを選択する。これにより，高速かつ効率的な翻訳が可能である。図 2- 6 に翻訳の流れを示す。

翻訳処理の各トランスファレベルの概要は以下のとおりである。詳細は，２．３．３および２．３．４で述べる。

(1) 単語レベルのトランスファ

日本語の単語列をそのまま中国語の単語列に変換する。熟語表現はこのレベルで翻訳することが多い。

(2) 構文レベルのトランスファ

日本語と中国語の構文構造が類似している場合には，このレベルで翻訳する。なお，意味解析まで必要かどうかは，構文解析プロセスで判断する。

(3) 意味レベルのトランスファ

格関係を解析する必要のある場合は，このレベルで翻訳する。日本語－英語間のトランスファ方式による機械翻訳はこのレベルで行うことが一般的である。

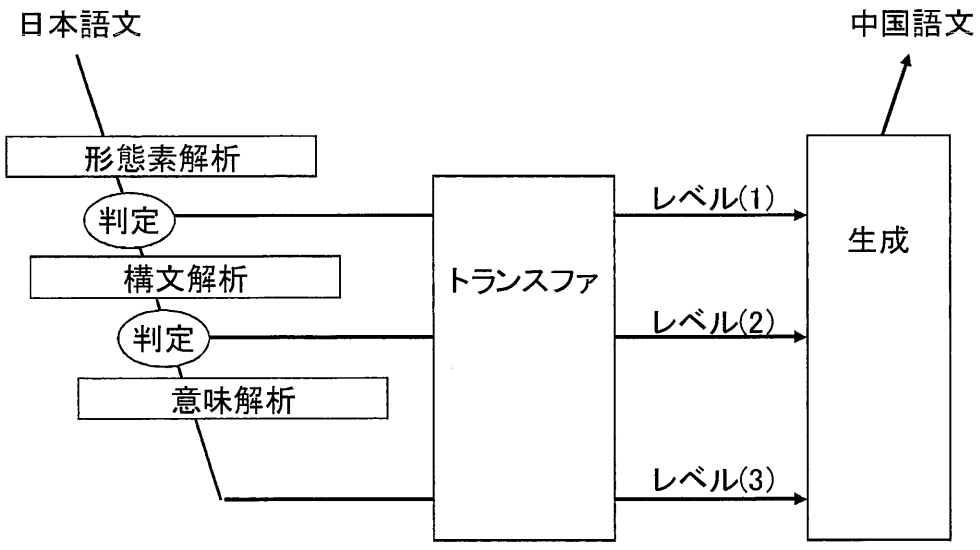


図 2- 6　日本語－中国語間の翻訳の流れ

２．３．３　中国語生成

構文解析や意味解析の結果は有向グラフで表現できる。ノードが単語を，アークがノード間の関係を表している。例として，以下の２文を取り上げる。

例８：　日本語：彼は座って本を読んでいる

中国語：他坐读书

（例４と同じ文）

例９：　日本語：来週北京に私は行きます

中国語：下个星期我去北京

例８の日本語文の構文解析結果を図 2-7 に示す。構文解析では，主に，係り受け解析を行い，上位ノードが下位ノードの係り先文節の単語となる。単語間の関係を表すアークは，この時点では，文節に付く付属語を記載する。図 2-7 の例で記載されている３つのアーク「ha」「null」「wo」はそれぞれ，助詞「は」，付属語なし，助詞「を」を表す。また，例９の日本語文の意味解析結果を図 2- 8 に示す。アークは文節間の格関係（意味的な関係）を表す。図 2- 8 の例で記載された３つのアーク「Time」「Goal」「Agent」は，それぞれ，時間，目的地，動作主の格を表す。

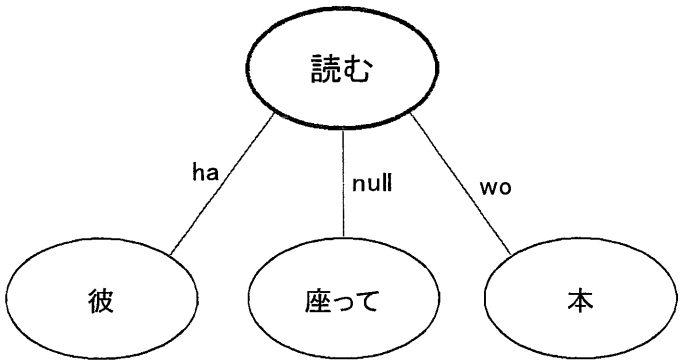


図 2- 7　構文解析結果の例

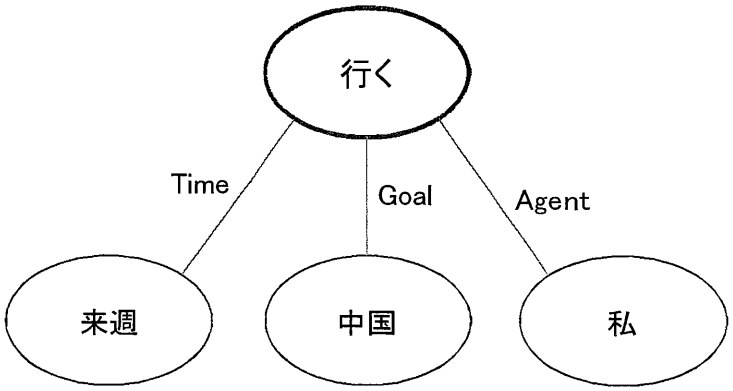


図 2- 8　意味解析結果の例

構文解析結果または意味解析結果から，中国語の依存構造を生成する。図 2- 9 に例 9 の中国語文に対応する中国語の依存構造を示す。この構造は，「下个星期」，「我」が「去」を前から修飾し，「北京」が後から修飾することを表している。

図 2- 6 に示した生成プロセスでは，中国語依存構造に対応する生成ルールを用いる。例 9 の文を生成するための生成ルールは以下の通りである。

ルール 1 ：agent main goal

ルール 2 ：time main

ここで，「main」は依存構造の上位ノードに当たる主要語を表す。例 9 では，「行く」に対応する中国語「去」が主要語である。ルール 1 は，agent が主要語の前に，goal が主要語の後に置かれることを意味する。ルール 2 は time が主要語の前に置かれることを意味する。図 2- 10 に中国語依存構造を生成する処理の流れを示す。日本語依存構造に対して生成ルールを順に適用させ，マッチすれば生成ルールに記載された中国語依存構造を生成する。

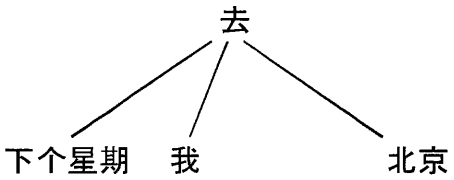


図 2- 9　中国語依存構造の例

上記の例では，日本語の意味解析結果から中国語文を生成したが，例８の場合は，一部を構文解析結果のままで中国語文を生成する。具体的には，「座って読む」の部分の「座って」と「読む」の関係を意味解析せず，単なる動詞の連用修飾という構文的関係だけの情報から中国語を生成する。中国語の場合も「座る」と「読む」の動詞を直接繋げるだけでよい。例８の日本語文から中国語文を生成するためのルールは以下の通りである。

ルール１：null main

ルール２：Agent main Object

図 2- 11 に例８の文の中国語生成処理の流れを示す。

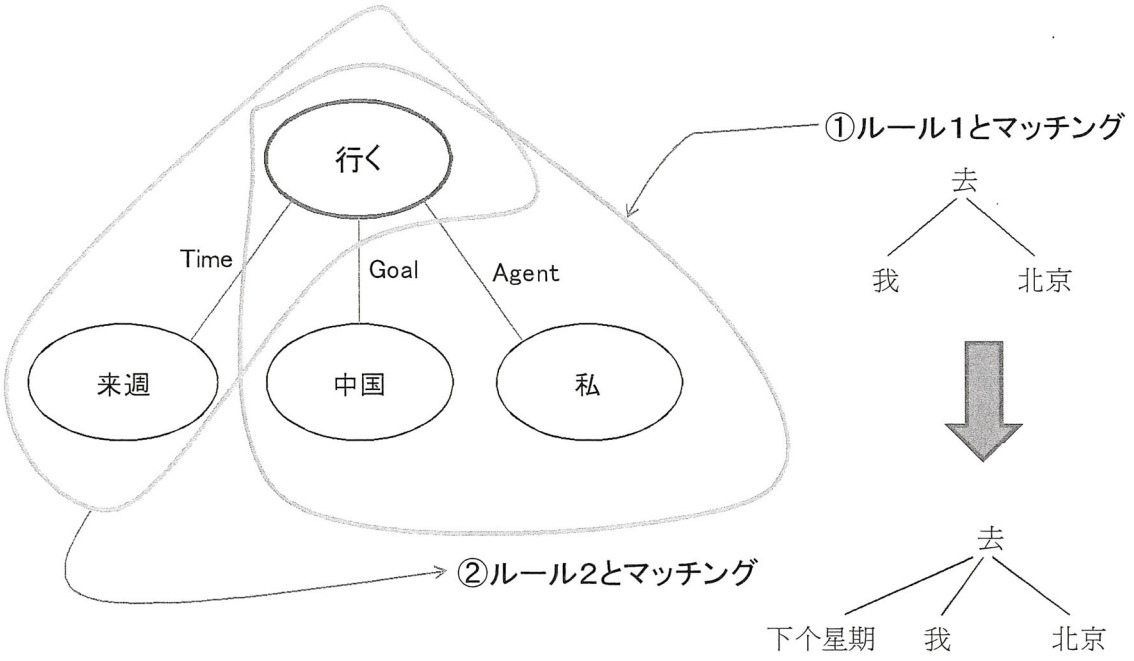


図 2- 10　中国語生成処理の流れ

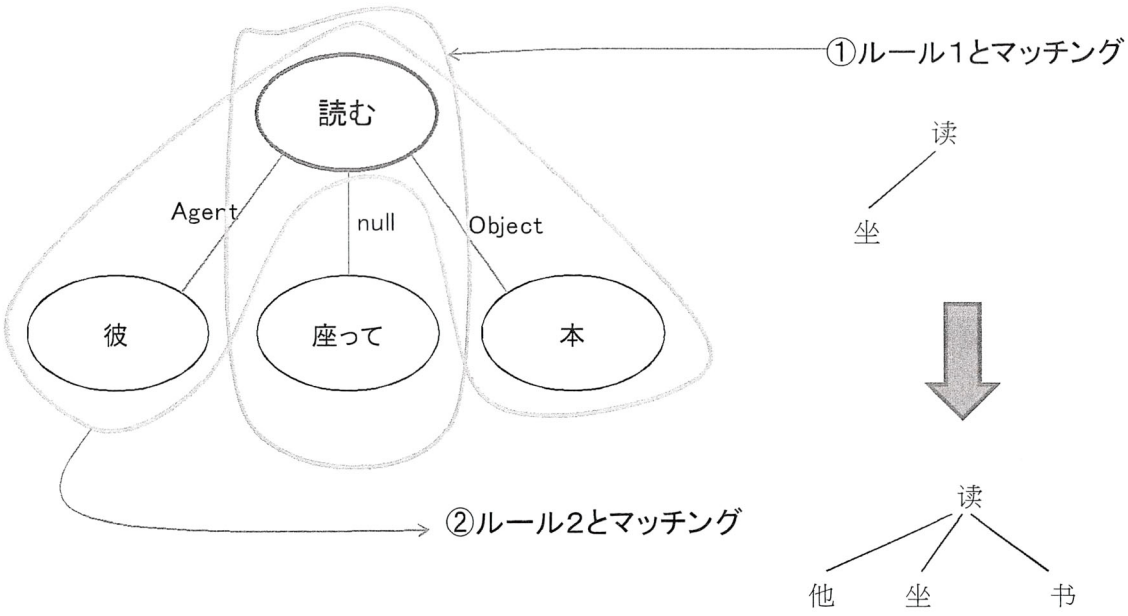


図 2- 11　中国語生成処理の流れ(その2)

2．3．4　熟語の翻訳

中国語は豊富な熟語表現を持つ。熟語の翻訳には，以下の理由から，単語レベルのトランスファが適していると考えられる。

(1) 日本語と中国語の熟語構造がしばしば似ている。構文レベルでの変換が必要なわけではない。

(2) 多くの熟語表現を簡素な変換ルールで記述することが可能である。

変換ルールの例を以下に示す：

x-ば+x-ほど+*y/越-x-越-*y/x:ps=adjective, y:ps=adjective

ここで，変数 x，y は句であり，ps は品詞を意味する。また，*はルールを適用する中心語を意味する。-は隣接する単語を繋ぐことを，+は係り受け関係にある単語を繋ぐことを意味する。ルールは「/」で区切られた３つの部分から成る。第１の部分は日本語の単語列を，第２の部分は中国語の対応する単語列を，第３の部分はルールの適用条件を記述する。

形態素解析結果が第１の部分とマッチして，かつ，第３の部分である条件部に適合した場合に，第２の部分の中国語単語列に直接変換する。x，y に当たる句のみが構文解析や意味解析の

パスを通ることになる。以下に例を２つ示す。

例１０：日本語文：　　データの送付は速ければ速いほどよい。

適用ルール：　x-ば+x-ほど+*y/越-x-越-*y/x:ps=adjective, y:ps=adjective

生成中国語文：数据　的　传送　越　快　越　好

例１１：日本語文：　　どうも風邪をひいたようだ。

適用ルール：　どうも+x-た-*ようだ/好象-x-了/x:ps*=verb

生成中国語文：好象　感冒　了

２．３．５　訳語の選択

一般に，日本語の単語に対応する中国語の訳語は複数あることが少なくない。適切な訳語を選択するために，３種類の共起データを用いて訳語選択処理を行った。

(1) 構文共起

構文的に共起する単語のことである。

例１２：　日本語：市場で野菜を求めた

中国語：在市场买蔬菜

例１３：　日本語：社長に面会を求めた

中国語：向总经理要求面会

この例では，「求める」の訳語を選択するために，動詞と目的語の共起を用いる。共起データは下記のように記載する。目的語の条件として，品詞，属性，スペルを記述できる。

买-object-noun(attribute:goods)

説明：目的語の品詞が名詞で物品を意味する場合に「求める」の訳語は「买」を用いる

要求-object-noun(spelling: 面会)

説明：目的語が「面会」の場合に「求める」の訳語は「要求」を用いる。

(2) 前接する単語との共起

該当語の直前に隣接する単語との共起である。

例１４：　日本語：講演プログラム

中国語：讲演节目

例１５：　日本語：計算機プログラム

中国語：计算机程序

この例では，「プログラム」の訳語を選択するために，直前にある単語との共起を用いる。共起データは下記のように記載する。

noun(spelling: 讲演)-节目

説明：直前の語が「讲演」の場合，訳語は「节目」を用いる

noun(spelling: 计算机)-程序

説明：直前の語が「计算机」の場合，訳語は「程序」を用いる

(3) 後接する単語との共起

該当語の直後に隣接する単語との共起である。

例１６：　日本語：プログラム言語

中国語：程序语言

この例では，「プログラム」の訳語を選択するために，直後にある単語との共起を用いる。共起データは下記のように記載する。

程序-noun(spelling: 语言)

説明：直後の語が「语言」の場合，訳語は「程序」を用いる

２．４　評価

本章で提案した手法を実装したシステムを開発し，技術文書，新聞，会話文等の２３００文に対して有効性を評価した。２３％の文は完全に正しく翻訳できたが，他の文は誤り箇所があった。ただし，訳文の文意がまったく分からないケースはなかった。誤り箇所の総計は４４７３であった。誤りの内容を表 2- 2 に示す。

いくつかのタイプの誤りは，共起データや熟語ルールを追加すれば解決できる内容であった。誤りの内容と解決方法を表 2- 3 に示す。

表 2- 2　翻訳誤りの内容

項番	誤りの内容	箇所数	割合 (%)
1	訳語の選択が適切でない	1528	34.1
2	語順の誤り	926	20.7
3	日本語単語が辞書にない	367	8.2
4	日本語の付属語が適切に訳されていない	521	11.6
5	日本語の熟語が適切に訳されていない	208	4.7
6	日本語の複合語が適切に訳されていない	190	4.2
7	時制やアスペクトを表す中国語が誤っている	345	7.7
8	その他	388	8.7

表 2- 3　対応できる翻訳誤り

項番	誤りの内容	解決方法
1	訳語の選択が適切でない	構文共起データを追加する
5	日本語の熟語が適切に訳されていない	熟語のルールを追加する
6	日本語の複合語が適切に訳されていない	隣接語の共起データを追加する

表 2- 3 に示した誤りは，誤り全体の約４３％であり，これらの誤りは翻訳アルゴリズムを変えることなくデータやルールの追加で解決可能であった。これらのデータやルールは外部パラメータであり，機械翻訳の知識を持っていないユーザでも簡単に追加することができる。

２．５　結言

トランスファ方式を用いた日中翻訳エンジンを開発した。本アルゴリズムでは，単語レベル，構文レベル，意味レベルの３段階のトランスファを適切に選択しながら翻訳する。また，日本語の構造を中国語の構造に変換し，３種類の共起データにより訳語を適切に選択する。これにより，従来の方式では難しかった日本語と中国語の構造が類似している文や語句を少ない翻訳

ルールで高品質に翻訳することが可能となった。

さらに，２３００文の評価の結果では，４３％のエラーは共起データや熟語ルールを追加することにより，解決できることが分かった。

言語表現は無限であり，日々新しい言葉が生まれ，専門分野が多岐にわたる中，機械翻訳システム提供側であらゆる分野のデータをあらかじめ作成し１００％完全な翻訳を行うことは不可能である。このため，機械翻訳システムのユーザが簡単にデータを追加し，翻訳の質を上げられる手段を持っていることが重要である。本章で述べた翻訳方式では，共起データや熟語ルールをユーザが自由に更新できるが，他にもユーザが自由に更新して翻訳の質の向上を可能とする手法を考案することが今後の課題である。

第３章　非構造化文書の分析

３．１　緒言

企業，団体，官公庁等の組織にとって，組織内外にある大量のテキストデータを分析してその結果を将来の施策立案に役立てることは重要である。従来，データの分析には，アソシエーションルール，ニューラルネットワーク，決定木等のデータマイニング手法および統計分析手法が様々な分野で用いられてきた。データマイニングの応用としては，スーパーマーケットの売上データや医療機関のレセプトデータの分析等が有名であるが，これらは，項目の決まっている構造化文書を対象としている[51]。

非構造化文書の分析には，従来からテキストマイニング手法が用いられ，名前，住所，電話番号等の固有表現の抽出，出現単語の頻度分析，重要語抽出，文書分類などが行われてきた。

テキストマイニングとデータマイニング，統計分析を組み合わせれば非構造化文書の分析は可能であるが，分析の方法は目的や対象文書の特性により様々であり，決まった手法を使えば効果的な分析結果が得られるわけではない。特に，ニーズの高い人や時間，場所に関わる分析においては，以下の課題が存在する。

(1) テキストからの固有表現抽出

従来から報告されている名前，場所，時間だけでなく，乗物，身長，衣服などの人の属性をテキストから抽出することが必要であるが手法が検証されていない。

(2) 連続値と離散値の混在する構造化データへのデータマイニング手法適用

時刻等の連続値と人の属性等の離散値が混在する構造化データをデータマイニングアルゴリズムに適用する手法が確立されていない。特に，アソシエーションルールは離散値のみに適用できるアルゴリズムであり，そのままでは連続値を扱えない。

(3) ランドマークとの関係性を明らかにする空間分析

テキストに書かれている事象と駅や学校などのランドマークとの空間的関係を分析する手法が確立されていない。データマイニングや統計処理の手法を適用するだけでは，ランドマークとの距離に関する分析結果を得ることはできない。

本章では，大量の非構造化文書を分析し，有益な知見を得るための方法を提案する。具体的な例として，警視庁が防犯電子メールサービスとして公開している「めーるけいしちょう」の

テキストデータを使い，実際に分析を行った結果についても報告する。対象データとしては，２００８年４月１８日から２００９年８月１７日に配信された４７８４件の都内全域を対象とした防犯メールを使用した。防犯メールは，東京都内で発生したひったくり，公然わいせつ，声かけ等の事件について記載されている。図 3- 1 にメール文書の例を示す。

今回対象とした防犯に関するデータの分析については，データマイニング手法等を用いた時空間分析 [52] [53] [54] [55]，回帰モデルによる分析 [56]，統計検定による分析 [57]，等が報告されている。これらの既往の研究において，目的や適用される手法は様々である。例えば，人口等の外部データベースとの相関関係分析や高層階，低層階別の犯罪特徴分析，駅や道路から犯罪発生場所までの同心円距離別分析などが実施されている。

しかしながら，上記の報告はほとんどが構造化データに対する分析である。一部，非構造化データを分析した例もあるが，対象となったデータ数は数百件にとどまっており，かつ，人手によるデータ編集作業を含んでいて分析の自動化ができていない。この理由は，テキストマイニング手法が非構造化データを構造化データへ充分な精度で変換できないため，人手による構造化データ作成作業が必要になり手間がかかるためと考えられる。

３．２　提案手法の概要

提案手法の処理の流れを図 3- 2 に示す。提案手法は２つのステップからなる。第１のステップは，メールテキストから固有表現を抽出し，構造化データに変換する処理である。固有表現としては，分析に有効と予想される空間情報や時間情報等を中心に抽出する。

Subject: 板橋警察署(公然わいせつ)
Date: Mon, 21 Apr 2008 14:38:01 +0900
From: info@keishicho.metro.tokyo.jp
To: xxxxxx@xxx.xxx.co.jp
4月21日（月）、午前11時10分ころ、板橋区氷川町の路上で、公然わいせつ事件が発生しました。（犯人の特徴については、20歳代、170cm 位、黒色っぽい上衣、色不明ジーパン、自転車利用）
【地図】http://www3.wagamachi-guide.com/Mail-Keishicho/index.asp?adr=13119041000
【問い合わせ先】板橋警察署 03-3964-xxxx（内線xxxx）

図 3- 1　防犯メールの例

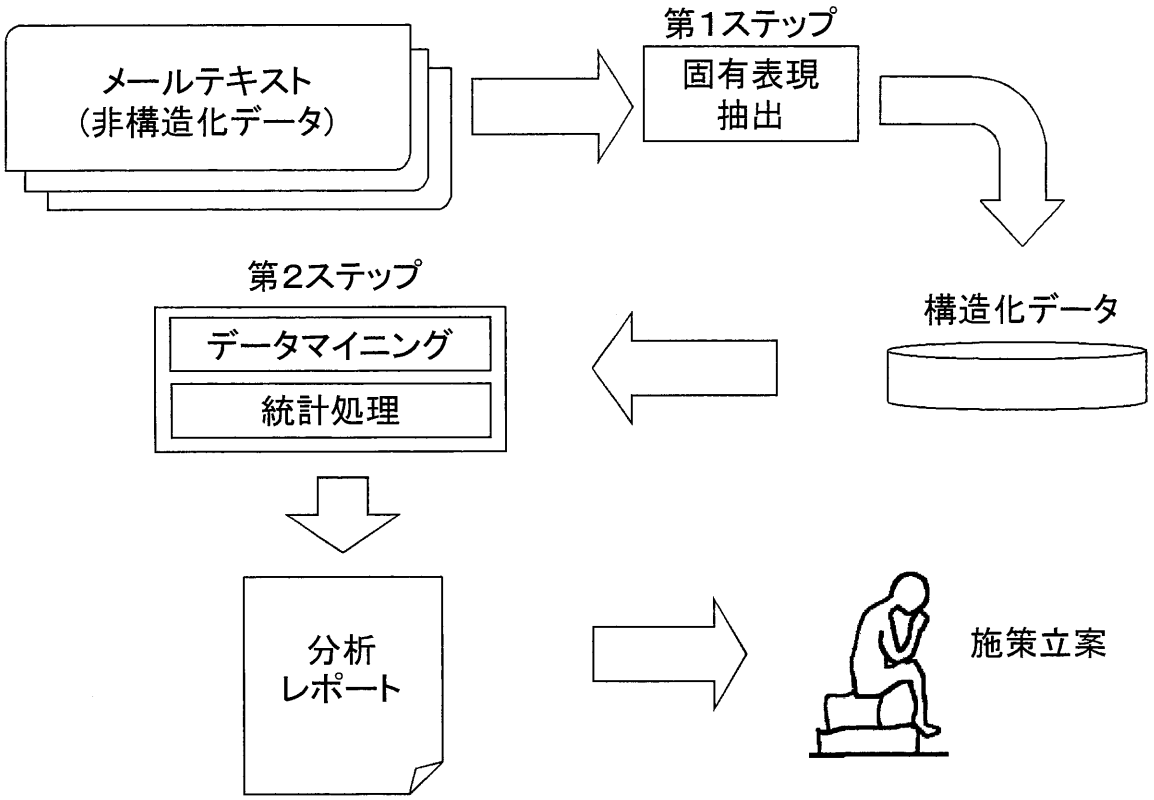


図 3- 2　分析処理の流れ

第２のステップは,構造化データを分析し,犯罪分類や犯罪パターン認識を行う処理である。ここでは，２種類の分析手法を用いた。１つは，データマイニングアルゴリズムのアソシエーションルールおよび決定木を使用する手法である。もう１つは統計処理を応用する手法である。分析の結果は，レポートとしてまとめ，関係者が施策立案の参考にすることが期待できる。

３．３　テキストからの固有表現自動抽出と構造化データの生成

３．３．１　テキストからの固有表現自動抽出

自然言語処理を用いて，表 3- 1 に示す固有表現をテキストから抽出した[58]。分析の基本情報となる事件種別と時間および空間属性のほかに，被害者や加害者の属性を抽出対象とした。

図 3- 3 に固有表現の抽出処理の流れを示す。まず，日本語のテキストを，形態素解析，構文

解析する。結果は依存構造で表される。形態素解析，構文解析は，２章で述べた機械翻訳と同様の処理である。なお，複数の文がある場合は，各文に対して，形態素解析，構文解析を行う。図 3- 3 に現れる文の例では，日付，時刻，場所，事件を意味する句が主動詞「発生しました」に依存する構造となる。図 3- 4 にこの文の依存構造を示す。

表 3- 1　抽出する固有表現

分類	固有表現	例
事件	事件種別	ひったくり，声かけ，強盗，進入窃盗，ひき逃げ
被害者	性別または年齢	女性，小学生，生徒，女性，女子高校生，児童
加害者	年齢	３０歳代，若い感じ
	身長	１６０cm 位
	髪型	長髪，短髪，スポーツ刈り，白髪まじり，坊主頭，はげ，パンチパーマ，パーマ
	衣服	ジャンパー，ジャージ，Ｙシャツ，ウィンドブレーカー，背広，作業衣，服，ジーパン，ズボン，トレーナー
	体型	中肉，がっちり型，小肥り，やせ型，太め
	乗物	徒歩，自転車，オートバイ，自動車，軽快車，スクーター，またがり式オートバイ，マウンテンバイク
	装身具	眼鏡，サングラス，帽子，マスク，ヘルメット，ノーヘル
時間	日付	２０１０年４月７日，平成２０年８月１日
	時刻	午後１１時１０分ごろ
	曜日	月，火，水，木，金，土，日
空間	警察署	板橋，高井戸，町田
	住所	新宿区中井一丁目，多摩市諏訪２丁目

犯罪種別，住所，年齢や身長を表す接尾語，髪型等は，解析用辞書に記載されている。事件種別，空間情報項目，髪型，衣服等は，解析時に辞書を参照することにより認識することができる。時間的項目，身長，年齢等は，品詞の並びのパターンにより認識する。

例えば，日付の抽出には，西暦と和暦の以下のパターンが対応する。

(数字)-年-(数字)-月-(数字)-日

(年号)-(数字)-年-(数字)-月-(数字)-日

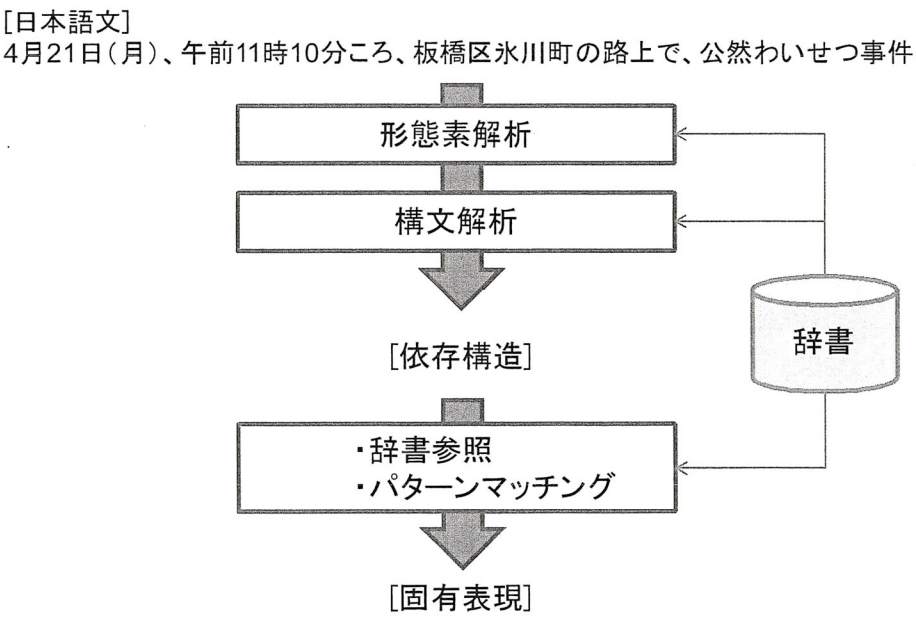


図 3- 3 固有表現抽出処理の流れ

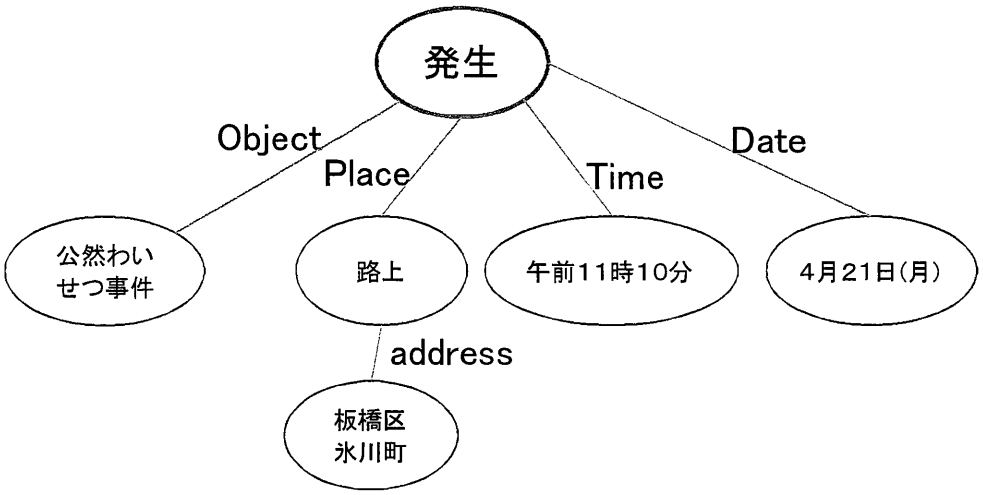


図 3- 4　依存構造の例

表 3- 2 に図 3- 1 に示した例文から得られた固有表現を示す。固有表現の値は，すべて正規化しており，例えば，テキスト中に西暦と和暦があっても，すべて西暦８桁に統一している。

表 3- 2　得られた固有表現

分類	固有表現	値
事件	事件種別	公然わいせつ
加害者	年齢	２０歳代
	身長	１７０cm 位
	衣服	上衣(色属性：黒)，ジーパン
	乗物	自転車
時間	日付	20080421
	時刻	1110
	曜日	月
空間	住所	板橋区氷川町

3．3．2　構造化データの生成

前節で述べた処理により，防犯メールから自動抽出した固有表現を用いて，分析のための構造化データを生成する。具体的には，以下に述べる２種類のテーブルを生成する。１つは，表3- 3 に示す事件ごとのテーブルである。各フィールドは，表 3- 1 に示す固有表現に対応している。防犯メールに書いてある情報しか抽出できないため，テーブルには欠損値が存在することになる。

もう１つのテーブルは，

表 3- 4 に示すような，丁目ごとの事件件数を表すテーブルである。空間分析を行うには，空間的なエリアごとに事件を分類してエリアごとに特徴を抽出することが適切である。東京都には，島嶼部等を除き５０９９の丁目が存在する。(一部，丁目がない住所表示があり，この場合は，上位住所の字で代用)。そこで，エリアとして丁目を利用した。空間分析としてランドマークを活用することが一般的であることから，丁目エリアの中心から最も近いランドマークまでの距離を計算して，テーブルの項目に加えた。ランドマークとしては，事件と空間的な関係がありそうな，駅，交番，神社，小学校を用いた。

距離計算のためには，まず，丁目を緯度経度に変換する。変換には，東京大学空間情報センターの提供する「アドレスマッチングサービス」を利用した。ランドマークの緯度経度情報を国土地理院の発行する「デジタルマップ２５０００(公共)」で得たうえで，緯度経度情報から距離を計算した。距離に関しては，人が移動するための距離として扱うことが適切であることから，直線距離ではなく，道のりに近いマンハッタン距離で計算した。

表 3- 3　事件テーブルの例

事件	被害者	加害者			時間			…
		年齢	…	乗物	時刻	…	曜日	…
公然わいせつ	生徒	20 歳代		自転車	1110		月	
ひったくり	女性	50 歳代		バイク	1720		土	
声かけ	…							

表 3- 4　丁目ごとの事件件数テーブル

丁目	事件件数			ランドマークまでの距離 (km)		
	公然わいせつ	ひったくり	声かけ	神社	・・・	駅
新宿区中井1丁目	0	1	2	0.53		0.25
多摩市諏訪2丁目	1	2	3	1.02		0.96
・・・						

３．４　分析

３．４．１　統計処理による分析

前節の処理により得られた構造化データを用いて本格的な分析を行う前に，まず，従来から行われている単純な統計処理によって得たデータの全体傾向を表 3- 5 に示す。事件種別としては，全部で６種類あるが，ひったくり，声かけ，公然わいせつがほとんどを占めるため，以下の分析では，この３種類のみを対象とする。

表 3- 5　データの全体傾向

観点	統計処理の結果
事件種別	ひったくり（４５％），声かけ（３１％），公然わいせつ（２０％）の３種類で全体の９５％を占める
曜日別の傾向	火水木が比較的多く，土日は少ない。(水曜が日曜の１．６倍)
時刻別の傾向	夕方から深夜にかけてが比較的多い。ひったくりのピークは２０時ごろ。公然わいせつと声かけは１６時ごろがピーク (朝８時ごろにも小さなピークあり)。
加害者の年齢	２０代，３０代の若い年代が多い。年齢が上がるにつれ件数が減少。

図 3- 5 に曜日別の事件件数のグラフを示す。土日に少ないことが明らかである。図 3- 6 に事件の発生時刻に関する時系列グラフを示す。時刻に関わる傾向が容易に理解できる。図 3- 7 に加害者の年齢別の事件件数のグラフを示す。２０代，３０代の若い年代が多く，年齢が上がるにつれ件数が減少していることが分かる。

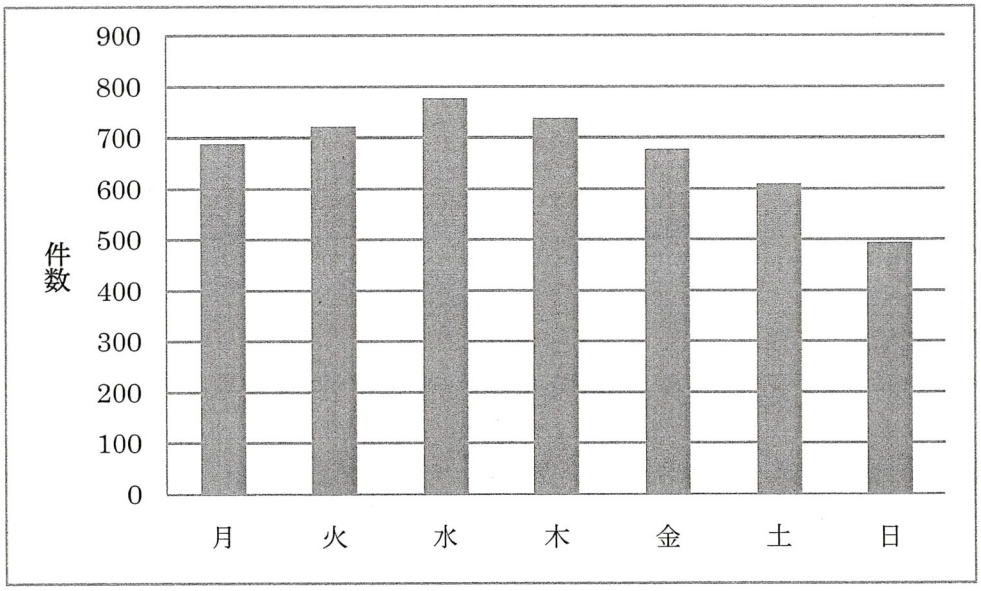


図 3- 5　曜日別の事件件数

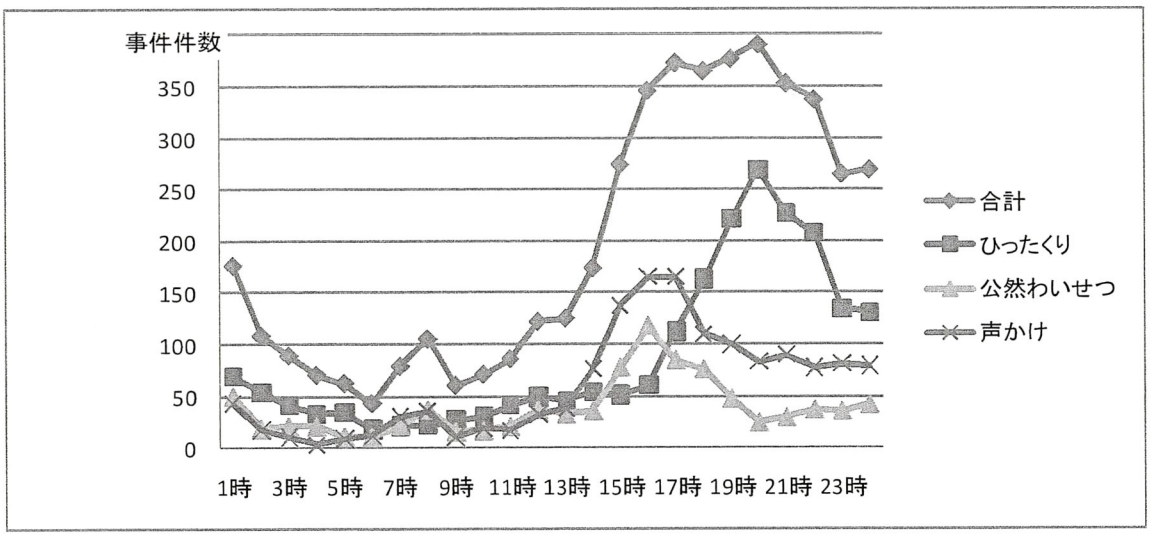


図 3- 6　事件の時系列グラフ

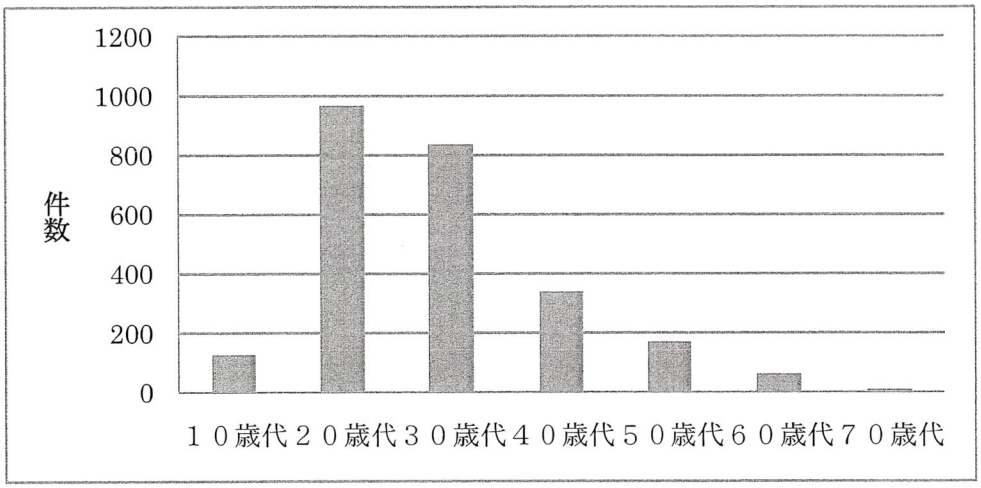


図 3- 7　加害者年齢別の事件件数

３．４．２　時間的な分析

固有表現として抽出した項目と事件との関係を明らかにするため，事件テーブルをデータマイニング技術を用いて分析した。結果として，主に時間と事件の関係を明らかにすることができた。データマイニングツールとしては「Ｗｅｋａ」を利用し[59] [60]，アソシエーションルールおよび決定木のアルゴリズムを適用した。

ここで,問題となるのは,データマイニングアルゴリズムが連続値を扱えないことである[61]。そこで，分析の前処理として，連続値である時刻を１時間ごとの離散値に変換して，データマイニングアルゴリズムが適用できるようにした。離散値変換の幅は任意に設定可能であるが,分析結果をパトロール等の対策に結び付けることを考慮して１時間とした。

以下，具体的な分析内容について述べる。まず，固有表現として抽出した各項目間の関係を分析するため，アソシエーションルールを事件テーブルに適用した。アルゴリズムとしては,欠落値があっても対応できるアプリオリを用いた[36]。その結果，時刻と乗物が事件種別と強く関係していることが分かった。得られたルールの大部分は自明であったが，いくつかのルールは活用できそうな知見であった。ルールは条件部と結論部から成り，条件部と結論部が共に成り立つ件数を条件部が成り立つ件数で割った値を確信度と定義する。表 3- 6 に確信度の高い上位１００ルールから得られた活用できそうな４つの知見を示す。なお，上位１００ルールすべてのリストは付録Ｂに記載する。

表3- 6　アソシエーションルールによる分析結果

ルール（得られた知見）	件数	確信度
徒歩の加害者が女性に対する場合は声かけがほとんどである	196	0. 995
２０/３０歳代および「若い感じ」の加害者が女性に対する場合は声かけがほとんどである	336	0. 995
午後３時/４時台に４０歳代の加害者が声かけをする相手は子どもがほとんどである	50	0. 994
オートバイの午前３時/午前１１時の事件はひったくりが多い	54	0. 982

ここで，各々の知見は必ずしも１つのルールに対応しているわけではなく，時間帯が継続している場合は，確信度が閾値を超えれば結合した。例えば，３番目の知見「午後３時/４時に４０歳代の加害者が声かけをする相手は子どもがほとんどである」は，表3-７に示す２つのルールを結合したものである。なお，前後の午後２時台，午後５時台では確信度の高いルールがなかったため，これ以上のルールの結合はできなかった。

次に，３種類の事件種別ごとの特徴を調べるため，決定木による分類を行った。アルゴリズムはＣ４．５を用いた。図3- 8に事件種別の決定木分類結果の一部を示す。

表3- 7　結合前のルール

分析結果（得られた知見）	件数	確信度
午後３時台に４０歳代の加害者が声かけをする相手は子どもがほとんどである	26	0. 99391
午後４時台に４０歳代の加害者が声かけをする相手は子どもがほとんどである	24	0. 99373

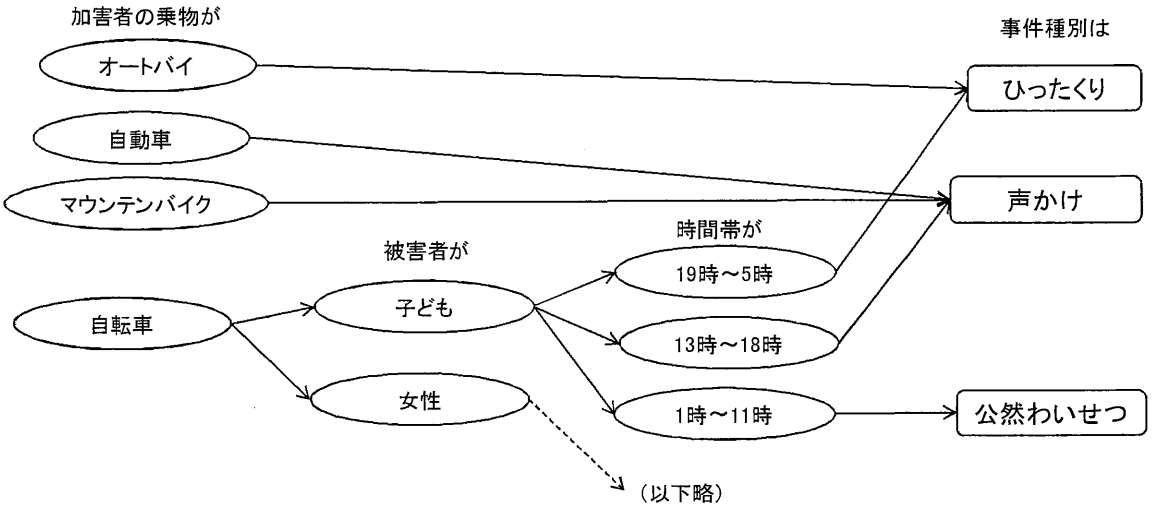


図 3- 8　決定木分類の結果

上記の２つのデータマイニングアルゴリズムを適用した分析結果によれば，事件種別は，主に，時刻，容疑者の乗物，被害者の性別や年齢に依存していることが分かる。

３．４．３　空間的な分析

事件は空間的にランダムに起こるわけではないことは，これまでにも言われている [56]。表 3- 8 に示すように，隣接する丁目でも事件発生件数は大きく異なる (これらの丁目での人口や面積に大きな差はない)。これは，事件発生場所には，何らかの空間的な要因があるのではないかと想定することができる。そこで，近くにあるランドマークとの距離に焦点を当てた。

表 3- 8　丁目ごとの事件件数

事件 / 丁目	ひったくり	公然わいせつ	声かけ	合計 (降順)
大田区西蒲田７丁目	6	2	5	13
練馬区桜台６丁目	7	2	3	13
練馬区桜台２丁目	3	4	4	11
・・・				
練馬区桜台４丁目	2	0	1	3
大田区西蒲田５丁目	1	1	0	2
・・・				

分析方法としては，まず，表 3- 3 の事件テーブルの各事件に対して，ランドマークとの距離を項目として追加して表 3- 9 に示す事件テーブルを作り，３．４．２で述べたデータマイニング手法を適用してみた。

しかしながら，ランドマークまでの距離をどんな幅で離散化しても，アソシエーション分析や決定木を用いて距離に関する確信度の高いルールを導き出すことはできなかった。これは，ランドマークとの距離が，時間的要因ほど顕著に事件に影響しているわけではないためと考えられる。実際，ランドマークとして駅を例にとり，事件ごとのランドマークと距離との関係をグラフ化してみると，図 3- 9 に示すように，場所に関してランダムに事件が発生した場合と大きな差がないことが分かる。

表 3- 9　ランドマークとの距離を追加した事件テーブル

事件	被害者	加害者		・・・	ランドマークまでの距離(km)		
		年齢	…		神社	…	駅
公然わいせつ	生徒	20 歳代			0. 38		0. 54
ひったくり	女性	50 歳代			0. 96		1. 25
声かけ	…						

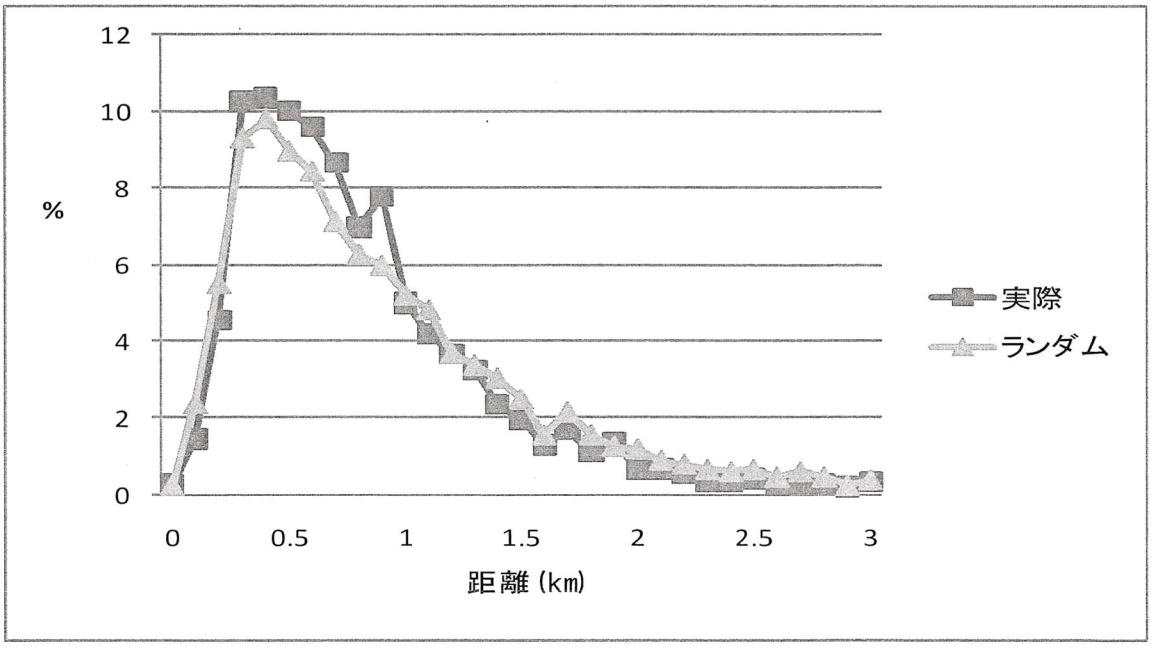


図 3- 9　事件発生場所と最も近い駅までの距離の関係

次に，事件が起きている場所と起きていない場所の比較を実施した。

表 3-4 に示す丁目ごとの事件件数テーブルを対象に，３．４．１で述べたような方法でデータマイニング技術を適用したが，「～ならば事件件数は０件である」というタイプの事件が起きていない場所に関するルールばかりが生成された。これは，ほとんどの丁目では事件が起きていないためと考えられる。

以上のことから，事件とランドマークとは，もっと弱い(確信度の低い)関係しかないのではないかと仮定し，その弱い傾向を把握するために平均距離を用いた分析を実施した。具体的には，事件の起こりやすさとランドマークまでの距離との関係を調べた。丁目ごとの事件の起こりやすさを，そこで発生した事件件数で定義する。表3- 10に，ひったくりを例に，各々の丁目での事件の起こりやすさとランドマークまでの距離との関係を示す。

例えば，神社との距離の項目を見ると，明らかに事件発生頻度が高いほうが神社との距離が近くなる傾向が読み取れる。図3- 10は，縦軸を事件発生件数，横軸をランドマークとの距離としてグラフに表した結果である。神社と異なり，駅，交番，小学校については，距離と事件発生頻度とは明確な傾向が見られない。同様に作成した公然わいせつ，声かけについてのグラフをそれぞれ，図3- 11，図3- 12に示す。

表3- 10　ランドマークまでの平均距離（ひったくり）

ランドマーク / 事件発生件数	駅	交番	神社	小学校
事件が起きていない丁目	1.012	0.548	0.644	0.524
事件が１件以上起きている丁目	0.861	0.511	0.586	0.445
事件が２件以上起きている丁目	0.839	0.519	0.540	0.438
事件が３件以上起きている丁目	0.829	0.519	0.506	0.437
事件が４件以上起きている丁目	0.885	0.507	0.438	0.464
事件が５件以上起きている丁目	0.975	0.555	0.408	0.492

単位：km

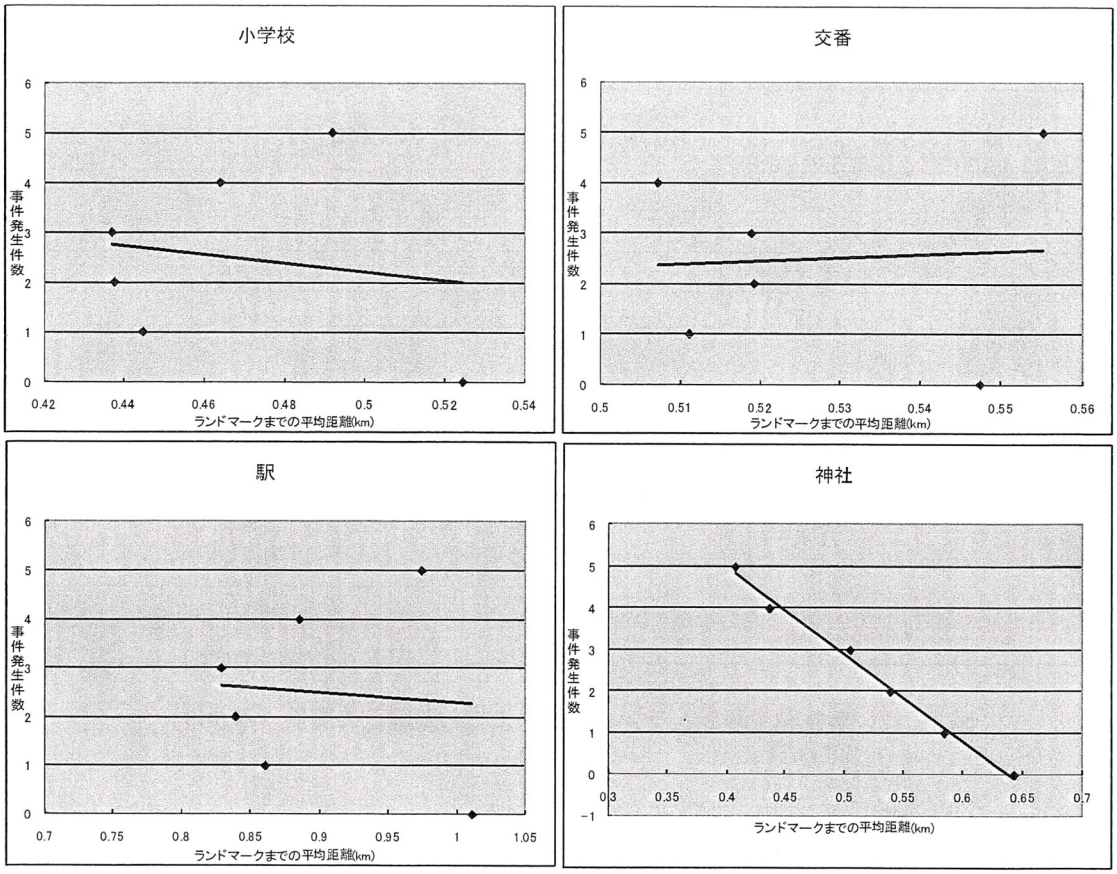


図 3- 10　事件発生頻度とランドマークまでの距離との関係（ひったくり）

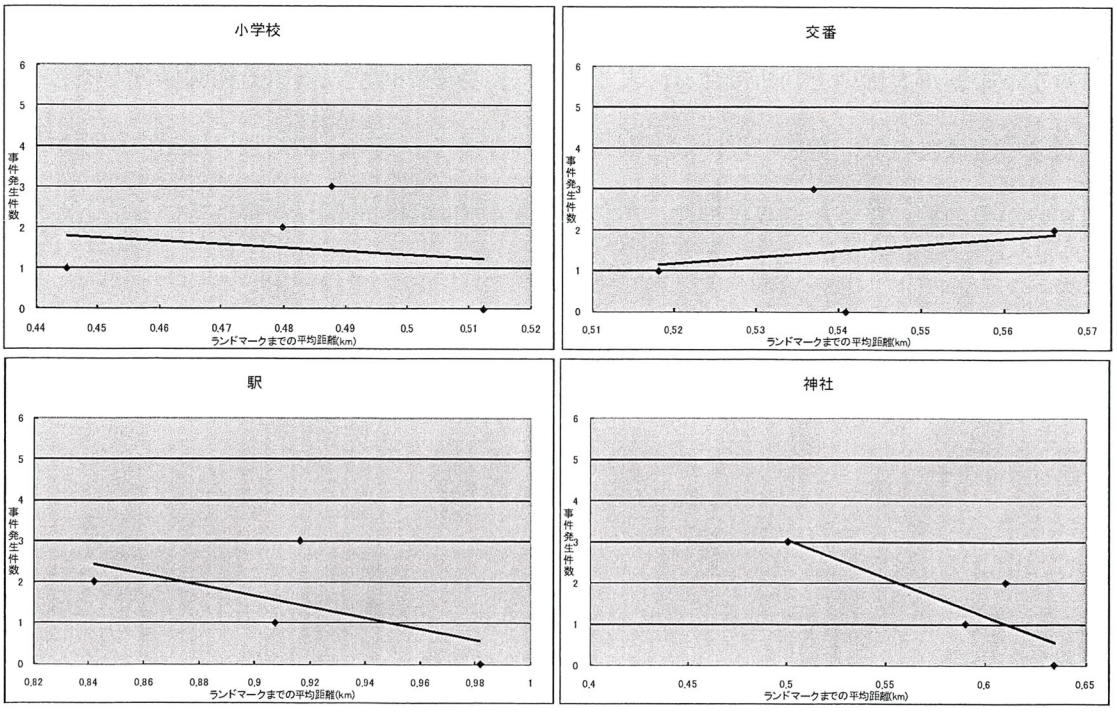


図 3- 11　事件発生頻度とランドマークまでの距離との関係（公然わいせつ）

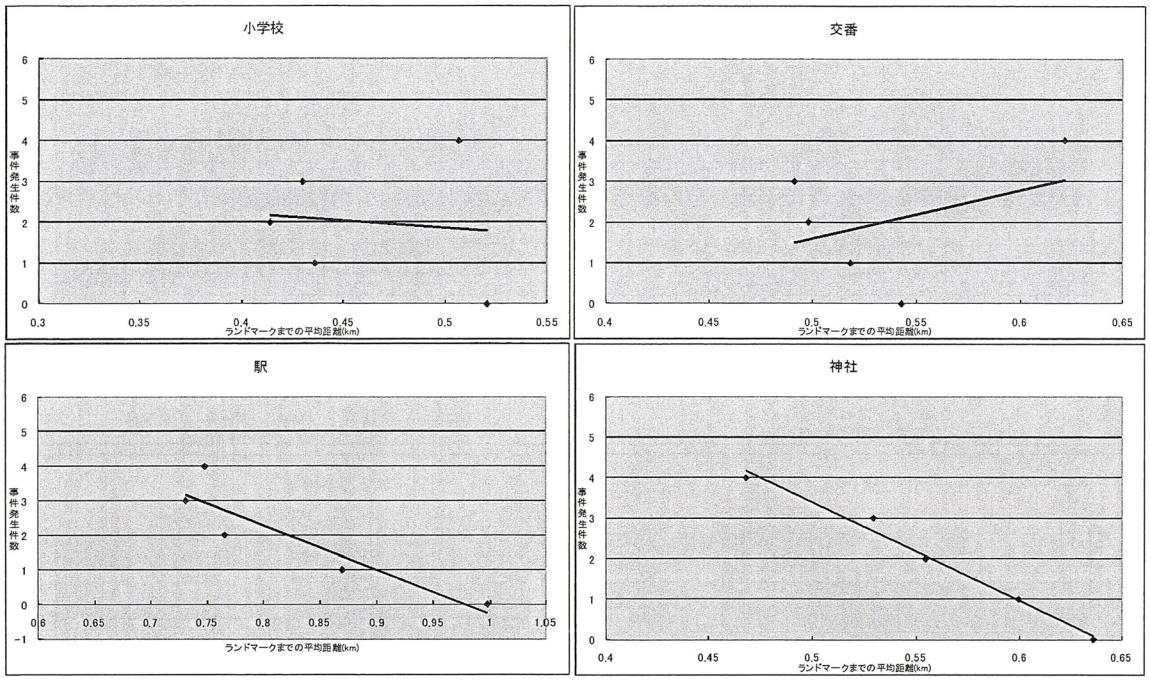


図 3- 12　事件発生頻度とランドマークまでの距離との関係（声かけ）

ランドマークまでの距離と事件発生頻度に関係があるのかどうかを判断するために，これらのグラフに対して，相関係数と回帰直線の傾きを計算する。相関係数は，ランドマークと事件との間にどの程度関係があるのかを判断できる。また，傾きが大きいほど，ランドマークの事件に対する影響が大きいと判断できる。表 3- 11 に，各事件種別に対する相関係数と傾きを示す。表 3- 11 によれば，神社に近いほどひったくりが多く，また，駅に近いほど声かけが多いことが分かるため，ひったくりと神社，声かけと駅が強い空間的関係を持っていることが推測できる。

表 3- 11　事件とランドマークの関係

事件＼ランドマーク		駅	交番	神社	小学校
ひったくり	相関係数	-0.328	-0.125	**-0.992**	-0.404
	傾き	-0.011	-0.001	**-0.042**	-0.007
公然わいせつ	相関係数	-0.733	0.188	-0.831	-0.442
	傾き	-0.026	0.002	-0.029	-0.007
声かけ	相関係数	**-0.981**	-0.981	-0.978	-0.895
	傾き	**-0.074**	-0.014	-0.029	-0.026

3．5　分析結果の可視化

分析結果は人が見て最終判断することから，犯罪の状況を直感的に把握できる仕掛けは重要である。表やグラフを使うことが一般的であるが，特に，空間的分析結果の表示においては，地図の利用が有効である。

地図利用の目的としては，全体を俯瞰することと詳細状況を見ることの２つがあることから，犯罪の発生状況を俯瞰できる俯瞰的表示と，それよりも細かい単位で事件発生場所を微視的に把握可能な微視的表示の２段階の可視化表示モードを備えるプロトタイプシステムを開発した。プロトタイプシステムでは，俯瞰的表示を市区町村単位で，微視的表示を町丁目の単位で表示した。両方のモードにおいて，利用者が犯罪発生状況を直観的に把握可能にするため，予め様々なパラメータを反映させたアイコンをデータベース化しておき，表示の際に利用する。可視化に利用する地図は，汎用性を考慮し API (Application Programming Interface) が公開されているツールとして，Google Maps™[62] を利用した。以下にそれぞれのモードの概要を示す。

(1) 俯瞰的表示モード

図 3- 13 に示すように，一定期間における犯罪発生件数をアイコンの大きさによって分かり易く提示する。また，特徴的な分析結果や，特異な状況にあると判断される場合は警報を表示する。

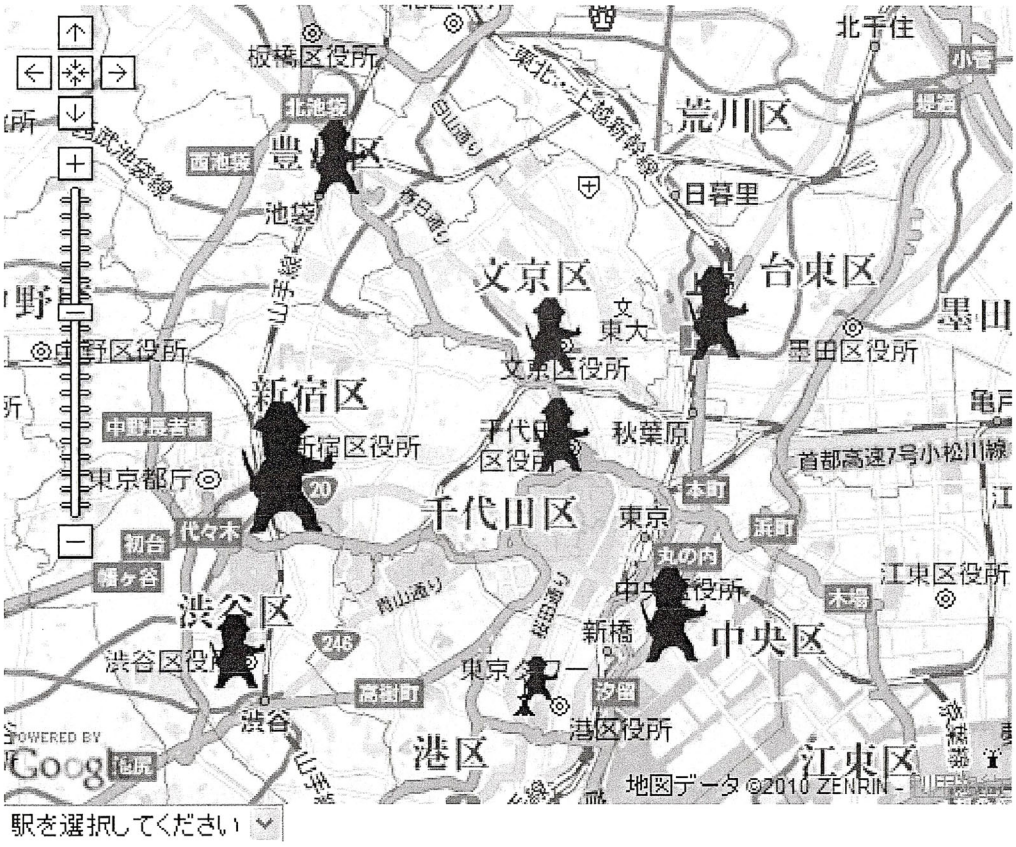


図 3- 13 市区町村レベルでの表示例

⑵ 微視的表示モード

個別犯罪情報を表示する。犯罪種別アイコンにより犯罪の種別や大まかな犯人像を直感的に把握できるほか，ランドマークを中心とする犯罪情報も提示可能である。図 3- 14 の例は，浜松町駅を中心に，架空の強盗発生情報を示したものである。発生からの日時が経過している事件ほど透過度を大きくしている。

図 3- 14　町丁目レベル (駅) での表示例（架空の強盗発生状況）

3．6　結言

本章では，非構造化データを，テキスト解析，データマイニング，統計の技術を用いて分析する手法について提案した。提案手法は，特に，時間的・空間的な分析を目的にしており，以下の３つの特徴を持つ。

(1) テキストからの固有表現抽出

従来から報告されている名前，場所，時間だけでなく，乗物，身長，衣服などの人の属性を抽出し，分析に利用した。

(2) データマイニングによる時間分析

連続値である時刻を離散値に分割してデータマイニングを行い，得られたルールを時間結合させる手法を提案した。

(3) 平均値を用いた空間分析

エリアごとの事件発生状況とランドマークまでの平均距離の相関分析を行い，空間的な特徴を得る手法を提案した。

実際に，提案手法を約５０００件の防犯メールデータに適用することにより，いくつかの有益な知見を得ることができた。この知見は，パトロール等の防犯活動に役立てることができるものと考えられる。

今後の課題としては，テキスト解析の精度向上によるさらに多くの固有表現の高精度抽出がある。固有表現としては，加害者の服装の色等も抽出すれば，本章で得たものとは違う知見が得られるかもしれない。「めーるけいしちょう」は文章が分かりやすいため，テキスト解析は比較的易しいが，より複雑な文章でも正確に固有表現が抽出できる必要がある。また，様々な観点，様々な分析手段，様々な表示インタフェースを工夫して，分かりやすい分析結果を提示することも有効である。

第４章　印刷文書のセキュリティ管理

４．１　緒言

文書管理システムで扱う文書や検索・分析結果は，プリンタで出力して紙媒体の形で利用することが多い。組織内での機密文書であったり公的な文書である場合もあり，改ざん・偽造・漏洩等に対するセキュリティ対策が必須である。３章で例に挙げた防犯メールについてもプライバシ情報を含むことがあり，印刷された紙を紛失したり漏洩したりしないように厳重に管理し，内容を改ざん，偽造したりすることのないように対策する必要がある。

印刷された紙の漏洩を物理的に防ぐことは難しいが，印刷文書の作成元を紙そのものに埋め込むことにより，紙が流出しても発信元が特定できる手段を付加し，漏洩を抑止することが可能である。

また，改ざん，偽造の抑止に対しては，印刷文書中に真贋判定を行うための手段を付加することが有効である。その場合，文書そのものに真正性を保証するデータを埋め込み，必要に応じてそのデータを取り出して評価する方法が一般的であると考えられる。この方法の利用イメージを図 4-1 に示す。このためには，データを埋め込んだ対象を一旦印刷し，それをスキャナで読み込み作成した画像から埋め込まれたデータを取り出すことが可能な手法の開発が必要である。

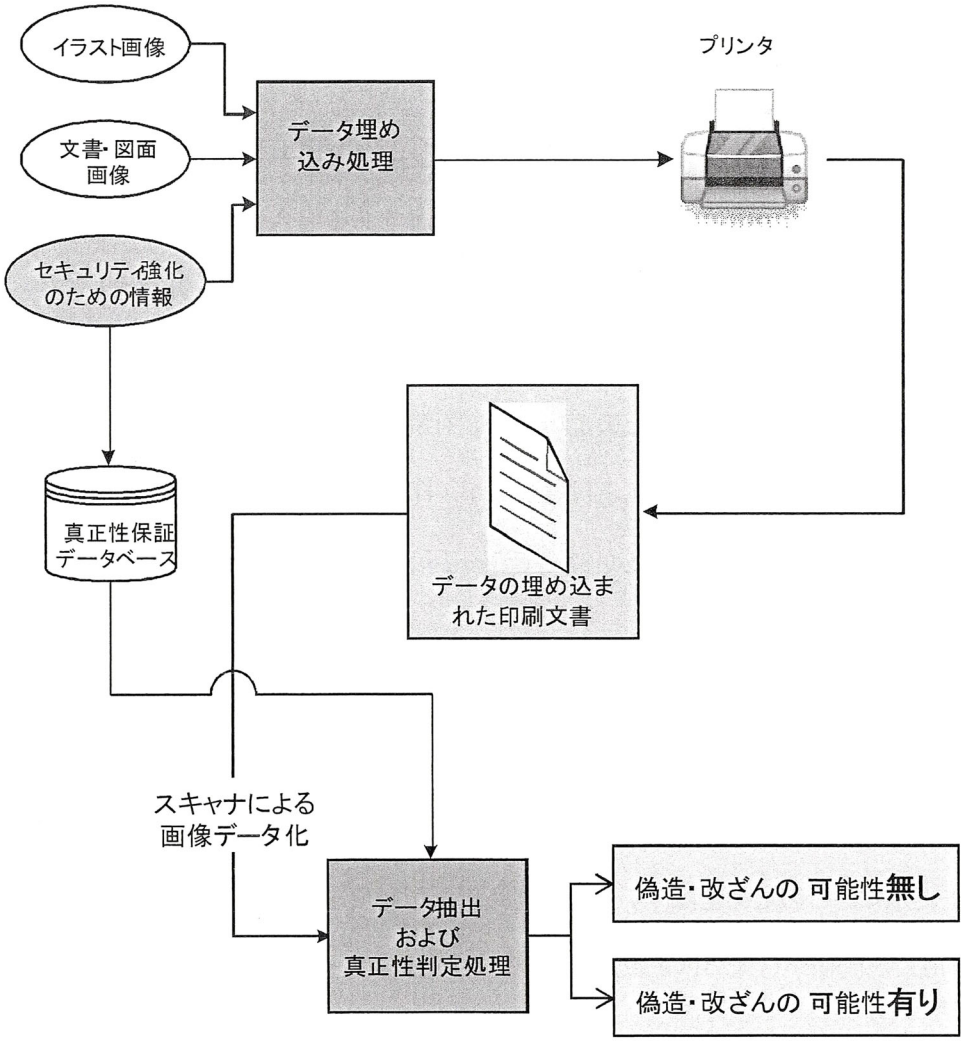


図 4- 1　真正性保証のためのデータを埋め込んだ紙文書の利用イメージ

画像に各種のデータを埋め込む技術として，電子透かし技術があり，カラー画像へのデータ埋め込み手法や２値画像へのデータ埋め込み手法，文書画像へのデータ埋め込み手法が提案されている [39] [40] [41]。しかし，これらは，データの埋め込みから取り出しに至るまでのサイクルにおいて，常に電子データのままであり続けることを想定しており，印刷原稿として一旦出力し，そこからデータを取り出すことは想定されていない。

一方，印刷物にした際にもデータ保持が可能な技術として，カラー画像やグレースケール画像にデータを埋め込むステガノグラフィーがある [42] が，モノクロ２値画像には適用できない。モノクロ２値画像にデータを埋め込む技術としては，プリンタのトナーが赤外線へ示す反応を

利用したもの[43]や，紙面の背景に微細な点を配置することでデータを表現する手法[44][63][64][65][66]が開発されているが，これらの方法により作成される点群や図形自体には意味がなく，利用者に違和感を与えたり，文章が読み難くなることがあると考えられる。

本章では，イラストを地紋に用い，その中にデータを埋め込む地紋透かしを用いることで，データの埋め込まれたモノクロ２値の印刷文書から利用者が受ける違和感を軽減する手法を提案する。なお，これ以降，本章において特に断らない限り，「ドット」とは「黒色のドット」のことを指す。

４．２　提案手法の概要

４．２．１　基本的な考え方

印刷文書のセキュリティを高める方法として，データを，人が視認できない形で埋め込むことが考えられる。他方，データを埋め込むことによって作成された点や図形が，文書の読み易さに悪影響を与えたり，利用者に違和感を与えることは望ましくない。また，印刷文書での使用色を考えた場合，カラーでの使用を前提にしたシステムでは，モノクロプリンタで印刷した際にデータが抽出できない可能性があるため，モノクロプリンタでの使用を前提とするシステムが好ましい。

提案手法では，図 4- 2 に示すように，イラストからドットで構成された地紋(ドット地紋)を作成し，それらのドットの一部を用いてその中にデータを埋め込む。ドット地紋を作成する際，イラスト中の明度が高い部分はドットの密度を低くし，明度が低い部分はドットの密度を高くすることで，イラストの濃淡諧調性を再現する。また，ドットの位置情報のみでデータを表すので，白から黒の濃淡情報しか保持しないモノクロ原稿にもデータの埋め込みが可能である。なお，今回，ドット地紋作成時に利用するイラストは，背景が白であるものを使用する。

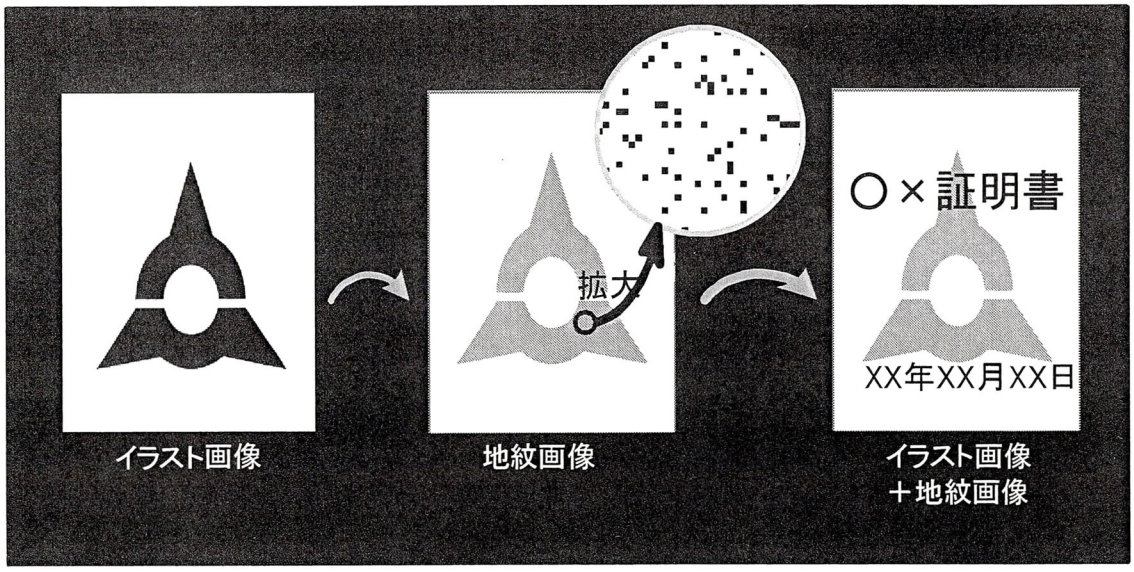


図 4- 2　データ埋め込み手法の概要

データを埋め込んだ印刷文書の作成は，設定する解像度を高くすることで同一スペース内での配置可能な点が増え，埋め込みが可能なデータ量は増加する。その一方で，印刷の際に微小なズレが発生した場合，埋め込んだデータの抽出が困難になることが予想される。このため，適切な解像度を選択することが必要となる。そこで，本手法では現在販売されている一般的なプリンタ(Canon LBP3000：エンジン解像度 600dpi 等)の性能等を考慮し，300dpi の解像度を持った画像を作成し，文書出力に用いることとした。したがって，今後，プリンタの性能が上がるにつれ，解像度を向上させデータの埋め込み量を増やすことが可能になると考えられる。なお，印刷する紙の質によっては，データの埋め込み量に限界があるが，ここでは，通常のオフィスで用いられているコピー用紙(白色度７０％以上，古紙配合率７０％以下)を対象とする。

真正性保証のために埋め込むデータとして，文書作成時刻やシリアルナンバーおよび電子署名等を想定している。これらのデータを印刷文書から取り出し，別の手段で得られる値と照合して突き合わせることにより，セキュリティの確保を実現する。

データの取り出しに際しては，まず，印刷文書をスキャナにより画像データとして読み込んだ後，データの埋め込みがあると判断された領域からデータの取り出しを行う。

スキャナにより印刷文書を画像として読み込む際には，600dpi の解像度でスキャニングすることを想定している。印刷文書作成には解像度 300dpi を用いているので，正確なデータを読

み取るためには，より解像度の高い 600dpi が必要だと考えた。また，読み込む画像のサイズについて考慮した場合，解像度が高くなるにつれ画素数が増え，処理に要する時間が増加するため，読み込み解像度は 600dpi が望ましい。使用するスキャナの性能については，現在市販されているスキャナは，ローエンドモデル(EPSON GT-S600)であっても 3200dpi の読み取り解像度を持つので，600dpi は容易に満たすことができる。

４．２．２　データを埋め込んだ印刷文書の作成処理の流れ

図 4-3 は，提案手法によるデータを埋め込んだ印刷文書を生成する処理の流れを示したものである。提案手法では，印刷文書作成のための入力データとして，データの埋め込み対象とする文書や図面の画像，データを埋め込むための地紋に使用するイラスト画像，そして埋め込むデータを用意する。それらをもとに，以下に示す処理手順により，データが埋め込まれた印刷文書を作成する。

処理(1) 候補位置規定点を配置し，データ埋め込み候補領域を設定

処理(2) データを埋め込む地紋のためのイラストを適切な画素値にするためヒストグラム変換

処理(3) 処理(1)で候補位置規定点を配置しデータ埋め込み候補領域を決定した文書・図面の画像データと，処理(2)で作成したイラスト画像データを基に，データ埋め込み候補領域へのデータ埋め込みの有無を決定

処理(4) 処理(3)で決定したデータ埋め込み候補領域へのデータ埋め込みの有無に対して，誤り訂正符号を計算し，候補位置規定点近傍へ配置

処理(5) 印刷文書に埋め込むデータを読み込み，誤り訂正符号を計算しデータに付加

処理(6) データを含む地紋作成のため，処理(3)で決定したデータを埋め込む領域への，データを含むドットの配置と，処理(2)で作成したイラスト画像をもとにしたドットの配置

処理(7) 処理(1)で作成した候補位置規定点を配置した書面と，処理(6)で作成した地紋画像を合成し，プリンタで出力

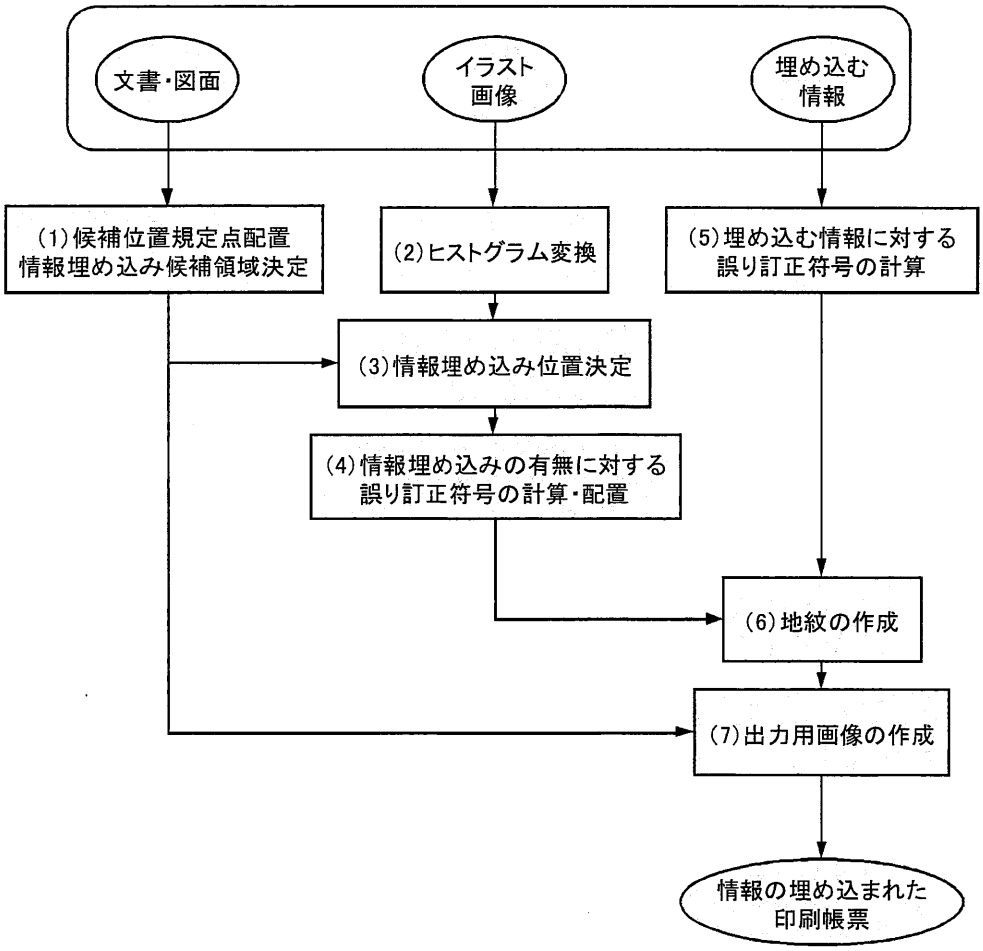


図 4- 3　データを埋め込んだ印刷文書の生成処理の流れ

4．2．3　データ抽出処理の流れ

図 4- 4 はデータを埋め込んだ印刷文書からのデータ抽出手順を示した図である。

データの抽出では，以下の手順で処理を行う。

処理 (1)　印刷文書をスキャナで読み取り，２５６諧調の画像データ化

処理 (2)　処理 (1) で読み込んだ画像データから，一定間隔おきに連続しているという性質を利用しての候補位置規定点を抽出

処理 (3)　処理 (2) で抽出した候補位置規定点を結び合わせ，データ埋め込み候補領域を決定

処理 (4)　処理 (2) で抽出した候補位置規定点の近傍に配置されている，データ埋め込み候補領域へのデータ埋め込みの有無に対して作成された誤り訂正符号用のデータを抽出

処理 (5) 処理 (3) で得られたデータ埋め込み候補領域から，データが埋め込まれていると思われる領域を選び出す。その際，処理 (4) で得られた誤り訂正符号を利用

処理 (6) 処理 (5) で決定したデータ埋め込み位置からデータの抽出

処理 (7) 処理 (6) で抽出したデータに対して，データ自体に付与されている誤り訂正符号を利用し，誤り訂正を実施後，データを出力

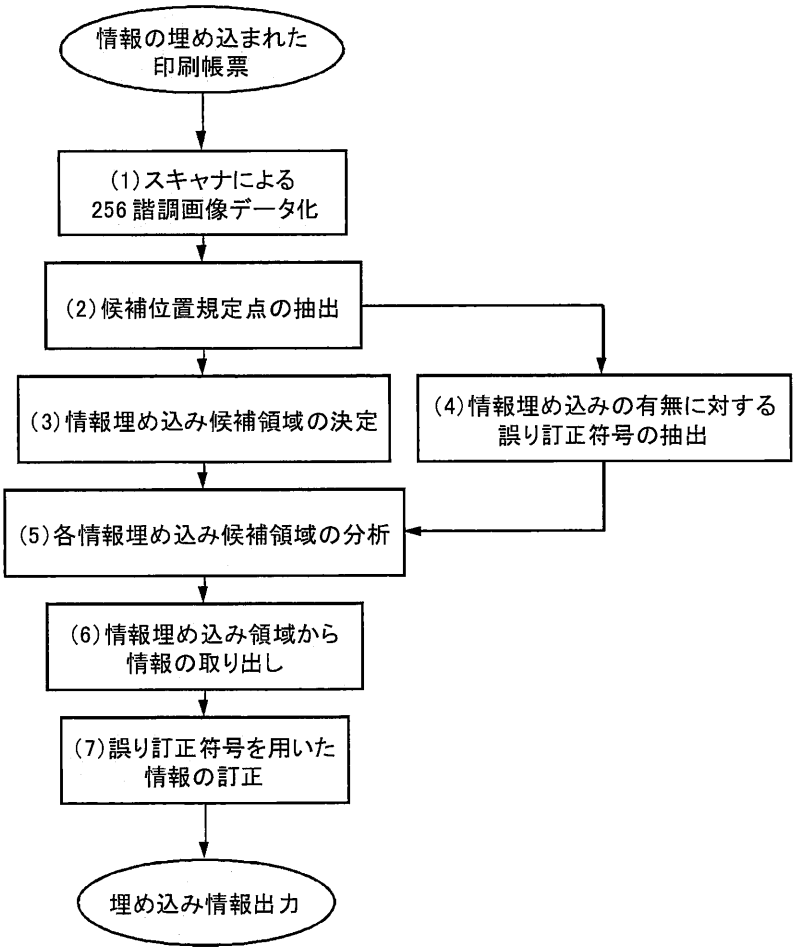


図 4- 4　データが埋め込まれた印刷文書からのデータ抽出処理の流れ

４．３　地紋に適するイラスト生成

４．３．１　ドット地紋に適したイラストの要件

提案手法では，イラスト画像内の画素値により，ドットの密度を決定してそれを地紋として表現している。画素値は０から２５５までの２５６諧調の表現が可能であり，画素値０が黒，画素値２５５が白となる。一般的に，背景部分が明るいイラストは，イラスト内の物体の画素値は低く濃い色で描かれ，背景と物体とのコントラストが強くなるように描かれている。このような画像では，オリジナルのままの画素値を用いて地紋画像を作成すると，イラスト内の物体部分におけるドットの密度が高い画像となることが多く，その画像を用いて印刷文書を作成すると，印刷文書中の文章や図形を読み取る際の障害になるのは自明である。この問題を避けるため，イラストの画素値をオリジナルのものより高くなるように変換し，文章や図形の認識の障害とならない地紋を作成する。

イラストの画素値を上げるためには，全画素の画素値を一定値上げることが考えられる。しかし，画素値を上げすぎると，イラストの絵画的意味が変質してしまう恐れがある。このため，イラスト内の画素値を上げて，なおかつ構成要素が失われることなく，さらに輪郭等の特徴が強調されたイラストに変換することが望ましい。

４．３．２　ヒストグラム変換による地紋用イラスト生成

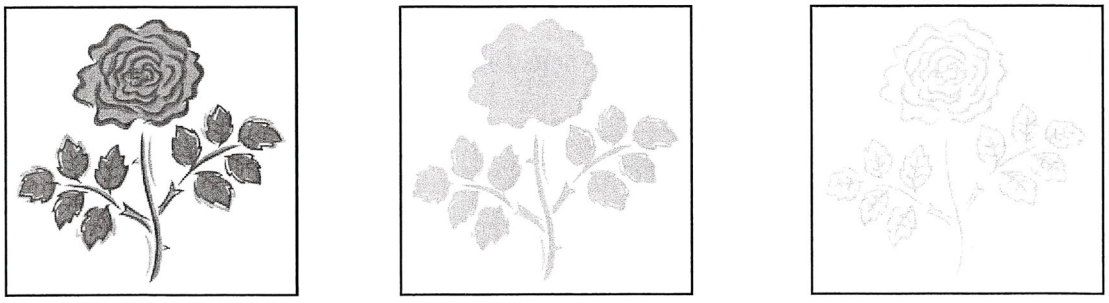
前述のように，一般的にイラストは，背景と物体とのコントラストを強くするため，輪郭線を明度の低い色で描くことが多い。提案手法においてもこの特性を利用し，イラストの物体内のコントラストをできるだけ維持しつつ，明度の高いイラスト画像を作成する。具体的には，イラスト画像の背景部分が画素値２５５に，背景を除いた画素の画素値が高い値（例えば１９２から２５４）に収まるようにヒストグラム変換を利用し，全体的に明るいがイラストの内容がわかりやすい画像を作成する。

提案手法によるイラスト画像のヒストグラム変換手順を以下に示す。まず，イラストがカラーの場合は白黒濃淡画像に変換する。その後，背景（画素値２５５）以外の画素に対して，高い画素値（例えば１９２から２５４）を持つようにヒストグラム変換を行う。ここで，背景も含めすべての画素に対してヒストグラム変換を行うと，背景すなわち白色の画素が非常に多くの画

素を占めるため，背景以外の部分が画素値の狭い範囲に押し込められることになる。したがって，イラストに描かれた物体のコントラストが弱い画像になり，特徴が強調されない。この問題を解決するため，提案手法では，背景以外の画素に対してのみヒストグラム変換を行う。提案手法では，画素値が２５５の画素にはそのまま画素値として２５５を与え，それ以外の画素については次式を用いて変換後の画素値を求めることでヒストグラム変換を行う。

$$w(u) = INT\left(\frac{v(u)\times(L-1)}{t} + V_{\min} - 1\right) \qquad \cdots(4\text{-}1)$$

ここで，$w(u)$ は変換前の画素値 u に対する変換後の画素値を表す。$v(u)$はイラスト画像内での画素値 0 から u までの値を持つ画素数を表し，L は変換後の諧調数である。t はイラスト画像内における画素値 255 の画素以外の画素の数である。$V_{\min}$は変換後に用いる画素値区間の最低画素値である。$INT()$ は整数化関数である。例えば，変換後の画素値区間を１９２から２５４の６３諧調にする場合はL=63，$V_{\min}$=192 を用いる。図 4- 5 (a) は，イラストを２５６諧調の白黒濃淡画像に変換したものである。図 4- 5 (b) は，全画素に対してヒストグラム変換を行ったものである。背景の白い画素が非常に多いため，ヒストグラム変換を施してもイラストで描かれた物体内でコントラストが低下していることがわかる。図 4- 5 (c) は，画素値２５５の画素を除いてヒストグラム変換を行ったものである。物体内でコントラストが高くなり，特徴が分かりやすいイラストに変換されているのがわかる。



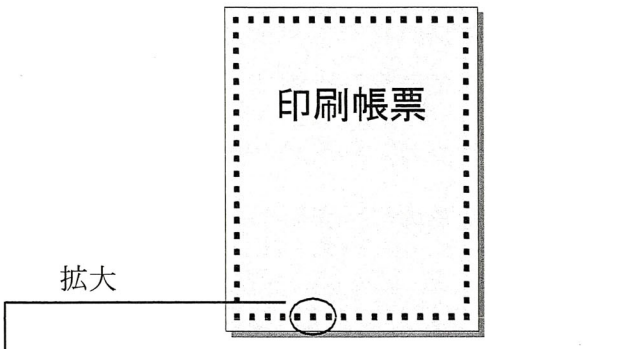
(a) 白黒濃淡画像 (原画)　　(b) 全画素を対象にヒストグラム変換を実施　　(c)提案手法によるヒストグラム変換を実施

図 4- 5　イラスト画像のヒストグラム変換

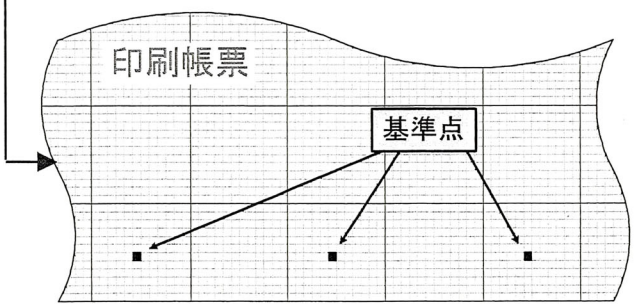
４．４　地紋を用いたデータの埋め込み

４．４．１　　データ埋め込み位置特定に用いる基準点

データが埋め込まれた印刷文書からデータを取り出すためには，データの埋め込み位置を特定するための仕組みが必要である。そこで，提案手法では，印刷文書の周囲を囲むように基準点と呼ぶドットを配置し，これを用いてデータ埋め込み位置を特定する。基準点は，図 4- 6 (a) に示すように印刷文書の左右端および上下端に平行に，それぞれ直線上に一定間隔で配置する。図 4- 6 (b) は印刷文書の下端に配置した基準点の一部を拡大表示した図である。図 4- 7 に示すように，上下および左右の基準点は対になるように配置する。例えば，左端に配置された基準点を１つ取り出し，その点から水平方向を探索すると，右端に対となる基準点が存在する。対になる基準点を直線で結び，できた交点 (候補領域中心点) を中心とする周囲 12×12 ピクセルの領域を，図 4- 8 に示すようにデータ埋め込み候補領域とし，その領域内の点の配置位置によりデータを表す。



(a)　全体の配置図



(b)　(a)の太線内の拡大図

図 4- 6　基準点

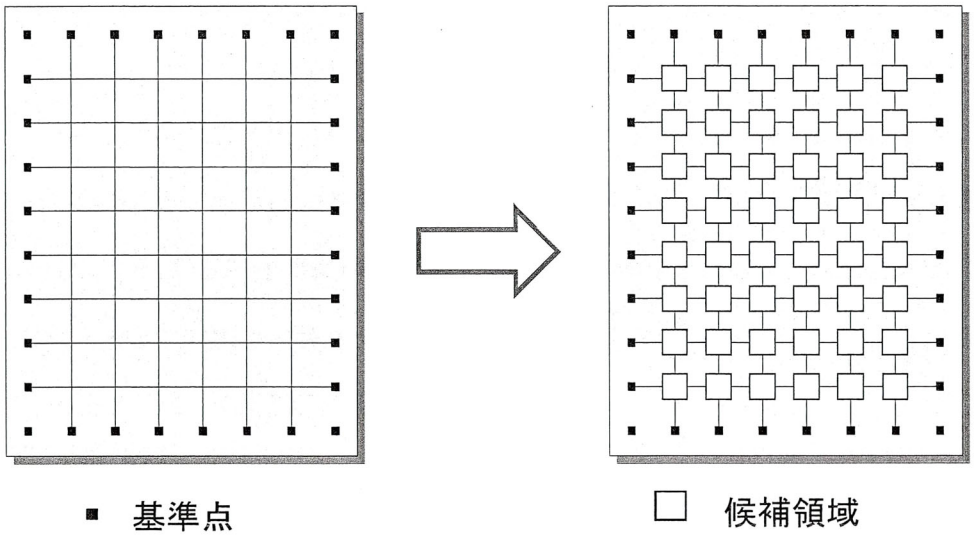


図 4- 7　基準点と候補領域

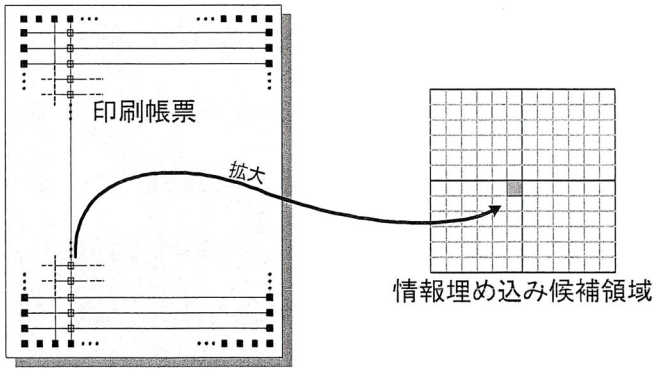


図 4- 8　データ埋め込み候補領域

4．4．2　データ埋め込み領域内でのドット配置方法

基準点を結ぶことによって作成された，多数のデータ埋め込み候補領域の中で，次の２つの条件を満たす領域をデータ埋め込み領域と呼ぶ。

条件１．印刷文書上の文章や図形と重ならない

条件２．変換後イラストの背景部分ではない

提案手法では，12×12 ピクセルからなるデータ埋め込み領域を，図 4- 9(a)のように 6×6 ピクセルの４つの小領域に分割し，その中の１つにだけ黒色のデータを表す点(データドット)を配置する。このように，１つのデータ埋め込み領域で２ビットのデータ量を保持することが可能である。例えば，データ埋め込み領域の左下に位置する小領域にデータドットが配置された場合を２進数表記で“00”とし，右下部分にデータドットが配置された場合を“01”，左上部分に配置された場合を“10”，右上部分に配置された場合を“11”とする。この方法で“00”を埋め込む場合，他の３つの小領域については，各小領域の図 4- 9(b)に示すように太線で囲まれた部分を，イラストの濃淡値の再現に使用する。そのため，小領域に配置するドット数 C を，印刷文書上に配置するイラスト画像内の各小領域と重なる画素から以下の手順を用い決定する。

(1) 小領域と重なる位置にある画素に対して次式を用い各画素の画素値の総和 S を求める。

$$S=\Sigma\{255-(\text{イラスト画像内の少領域と重なる各画素の画素値})\} \quad \cdots(4\text{-}2)$$

(2) INT(S/255)が 28 以上なら C=28 とし，INT(S/255)が 27 以下なら C=INT(S/255)とする。

小領域に配置するドット数 C を決定後，配置位置を決定する。各小領域に１つ以上のドットを配置する場合，図 4- 9(c)において斜線で示すＬ字型領域内に少なくとも１つ以上のドットが配置されるように設定する。これは，データ読み取り時に，イラスト画像の濃淡値を再現するためのドットが小領域に１点だけ配置された際，データ読み取り時にデータドットとして誤判断することを避けるためである。その後，残りのドットを図 4- 9(b)の各小領域内の太線内にランダムに配置する。図 4- 9(d)の点線で囲んだ領域は，データ読み込み時にドットが存在した場合，そのデータ埋め込み候補領域にはデータが配置されていないと判断する。このため，濃淡値を再現するためのドットは配置しない。また，図 4- 9(d)の斜線部分は，データドットとその他のドットの距離が一定間隔以上となることを保証するための領域である。

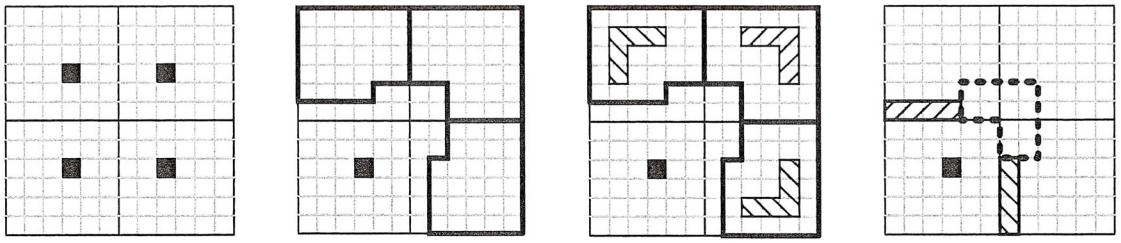


図 4- 9　データ埋め込み領域のドット配置

上述の手順により，データ埋め込み領域中のデータドットを配置しない３つの小領域を用いて，地紋透かしにしたときの画素値の総和をオリジナルの値にできるだけ近づけることにより，イラスト画像の濃淡値を地紋透かしで再現して，絵画性が失われることを防止する。

提案手法では，基準点が配置される間隔を狭め，多数の点を配置することで，データ埋め込み候補領域を増やして，同じ大きさの印刷文書により多くのデータを埋め込むことができる。しかし，データが埋め込まれたデータ埋め込み候補領域は，イラストの諧調性を失うことになるので，その点を考慮して基準点数を選択するように注意する必要がある。

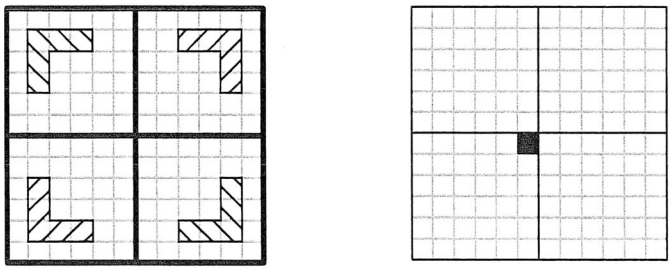
４．４．３　データ埋め込み領域外へのドット配置

印刷文書上で，データ埋め込み領域以外の部分を，前節に示した２つの条件を満たさなかったデータ埋め込み候補領域（埋め込み条件不適合領域）と，その他の領域（非データ埋め込み候補領域）に分類する。前者は，４．４．１で説明した，基準点を結んで多数作成した候補領域中心点の周囲 12×12 ピクセルの領域（データ埋め込み候補領域）のうち，前節で示した２つの条件を満たさなかった領域を指し，データドットは配置されていない。また，後者は，紙面上のデータ埋め込み候補領域を除く箇所を指す。

まず，埋め込み条件不適合領域でのドットの配置について図 4- 10 を用いて説明する。12×12 ピクセルの大きさを持つ，埋め込み条件不適合領域を４分割し，それぞれの小領域でイラスト画像の濃淡値を再現する。各小領域に配置するドットの数は，前節と同様に式 (4-2) により決定する。各小領域に配置するドットの数が１点以上の場合，図 4- 10 (a) に示す各小領域内の斜線

で示した位置に，少なくとも１点以上のドットを配置する。これは，前節同様，データ読み取り時に，濃淡値を再現するためのドットを，データドットと誤判断することを避けるためである。残りのドットは，各小領域内でランダムに配置する。また，ドットの配置対象としている埋め込み条件不適合領域が，印刷文書に配置された文章や図形と重なる場合，図 4- 10 (b) 中の黒色で示した位置にドットを配置する。このドットは，文字や図形を構成する画素を，データドットだと誤って判断しないために配置するものである。

次に，非データ埋め込み候補領域でのドットの配置について説明する。印刷文書上の非データ埋め込み候補領域では，6×6 ピクセルの小領域に分割し，その小領域に対応するイラストの濃度に応じて配置するドット数を決定する。配置するドットの数は，前節と同様の手順で式 (4-2) を使用して決定し，その後，小領域内に求めた数のドットをランダムな位置に配置する。



(a)小領域判別のための
ドット配置位置

(b)文章等回避用
ドット配置位置

図 4- 10　埋め込み条件不適合領域のドット配置

４．４．４　抽出精度向上のための誤り訂正符号導入

提案手法は，証明書の利用者が書類をプリンタにより出力し，受け取り側は作成された印刷文書からスキャナによりデータを読み込むという利用方法を想定している。したがって，文書の印刷や，スキャナによる読み取りにより，データの劣化が生じることが予想される。そこで，誤り訂正符号 [67] を２段階の異なるレベルで導入し，データ抽出精度の向上を図る。

誤り訂正符号は，送りたいデータ (データビット) に，任意の誤り訂正アルゴリズムを用いて作成した検査ビットと呼ばれるデータを付加し，それらを１つのデータとして扱う。誤り訂正能力は，「生成多項式」と呼ばれる多項式 G(x) に依存する。図 4- 11 は，符号化の流れを多項式表現を用いて示したものである。

提案手法では，印刷文書上に埋め込むデータ自体に誤り訂正符号を付与することに加え，印刷文書上のデータ埋め込み候補領域でのデータの有無に対して誤り訂正符号を生成し，基準点近傍へ配置することにより，データ抽出精度の向上を図る。後者は，埋め込むデータ自体の誤りを訂正するためのものではなく，データ抽出時に各データ埋め込み候補領域でのデータ埋め込みの有無を間違えることにより，データ抽出精度が下がることへの対策である。

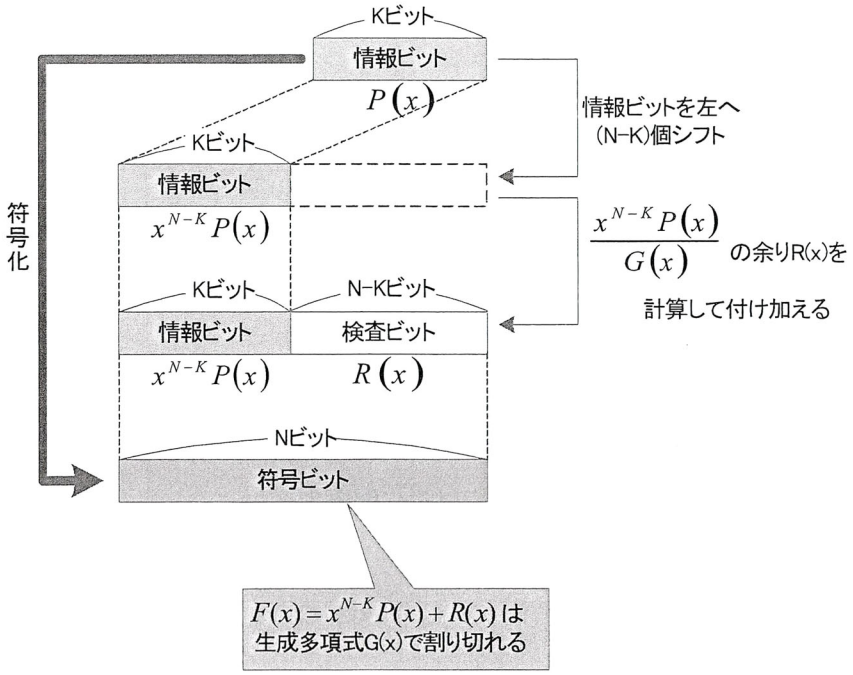


図 4- 11　符号化の流れ

この方法では，まず，垂直または水平に並んだデータ埋め込み候補領域に，データが配置されているかどうかをもとにビット列(データ埋め込みビット列)を作成する。例えば４つの水平または垂直方向に連続したデータ埋め込み候補領域を調べた際に，１つ目から順番にデータの埋め込みが有，無，無，無となっていた場合，データ埋め込みビット列は“1000”となる。このデータ埋め込みビット列に対する検査ビットを作成し符号化する。検査ビットを表すドットは，基準点の近傍に配置する。検査ビットに割り当てるデータ量を多くすると，誤り訂正の対象のデータ埋め込み候補領域数を増加させることができる。その反面，検査ビットを表すドットが多く配置され，印刷文書の見た目に悪影響を及ぼすと考えられる。

そこで，提案手法では，誤り訂正効果の高い一部のデータ埋め込み候補領域に対してのみ，誤り訂正符号を作成することで，基準点近傍に配置するドット数を抑制する。提案手法では，作成した印刷文書からデータを取り出す際に，画像の周辺部に配置された基準点を結び合わせて候補領域中心点を作成する。そのため，基準点からの距離が大きいほど，候補領域中心点が本来の位置からズレて検出される可能性が高まる。大きなズレが発生した場合，本来データが埋め込まれていない領域を，データが埋め込まれている領域と誤判断してしまう。そのため，画像の中心付近のデータ埋め込み候補領域に対してデータ埋め込みビット列を作成し，誤り訂正を行うことで高い効果を得られると考えられる。以上のことから誤り訂正符号作成時に使用するデータ埋め込みビット列を，画像の垂直方向，または水平方向の中心位置から取得していく。これにより，基準点から離れた領域に対して誤り訂正を施すことが可能となり，効果的に精度が向上できると考えられる。図 4- 12 は，上下端の基準点近傍へ配置する検査ビットを作成するためのデータ埋め込み候補領域を示したものである。垂直方向の中心を示す点線よりも下側で，なおかつ中心付近のデータ埋め込み候補領域から，データ埋め込みビット列を作成し，それに対応する検査ビットを下端の基準点付近に配置する。

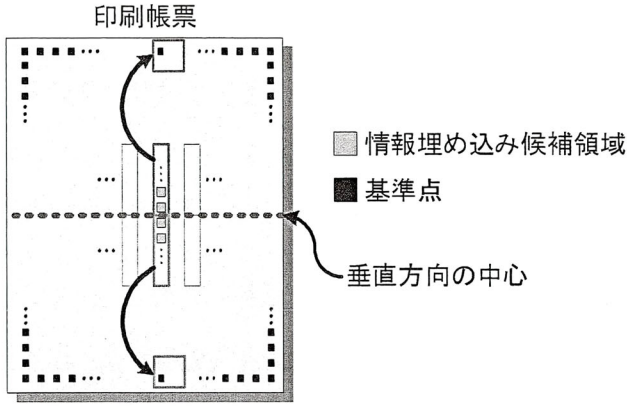


図 4- 12　データ埋め込みの有無による誤り訂正符号

４．５ 印刷文書からのデータ抽出

４．５．１　データ位置規定点の抽出

データを埋め込んだ印刷文書からデータを取り出すために，まず，スキャナにより印刷文書を読み込み電子データ化する。この際，印刷文書が解像度 300dpi の画像から作成されているため，最低でも 600dpi の解像度で読み込むことが必要である。スキャナにより画像ファイルに変換後，データの埋め込み位置を特定する。データの埋め込み位置を特定するために，まず，候補位置規定点を抽出する。候補位置規定点の特定方法は，以下に記す読み取り手法を利用する[68]。

候補位置規定点を特定するために，まず，１つのドットを構成していると思われる画素を１画素分の座標を持つように抽象化し，その座標をドットの位置と定める。抽象化は以下の手順で行う。なお，候補位置規定点は画像の縁に存在するため，読み込んだ画像内の上下左右端に近い画素にのみこの処理を行う。

(1) 読み込んだ画像内の画素の画素値(図 4- 13 (a))で横方向に差分を取る(図 4- 13 (b))。最も端に存在する画素等の差分をとる対象が存在しない画素は計算の対象としない。

(2) 差分値により，画素を＋画素と－画素と０画素の３種類に分類する。差分値が＋の閾値(現在は８)よりも大きければ＋画素とし，－の閾値(－８)よりも小さければ－画素とし，それ以外を０画素とする。

(3) ＋画素と－画素について，隣り合う同じ分類の画素をまとめて領域とする(図 4- 13 (c))。

隣り合う同じ分類の画素がなければ，その画素だけで領域を作成する。

(4) 差分の方向で，－領域と＋領域が連続している，または，２つ以下の画素からなる０領域をはさんで－領域と＋領域が隣り合って並んでいる箇所を取り出し，その箇所をドットを構成する画素だと判断する(図 4- 13(d))。

(5) (4)においてドットを構成する画素だと判断した領域において，領域を構成する画素の数が３つ以下のものは，ドットではなく汚れだと判断して除去する。

(6) ドットが表している画素だと判断した画素中で最も低い画素値を持つ点を探し，その点の座標をドットの位置とする(図 4- 13(e))。

このようにして，複数の画素から構成されていたドットを１画素の大きさに抽象化し，その画素を用いてドットの連続性を評価する。連続性の評価は，以下の手順に示す方法で，１画素の大きさに抽象化したドットを採点し任意の閾値以上の点数を獲得したものを候補位置規定点と判断する。

(7) １画素の大きさに抽象化したドット(ドット a)から X 画素離れた画素を中心とする距離に位置する領域(ドット存在期待領域)内に，一点だけ他のドットが存在している場合にドット a に対する評価点として，点数を加点する(図 4- 14(a))。複数点のドットが存在していた場合は加点しない。ドットが存在しない場合は点数を減点する。

(8) ドット存在期待領域に一点のドットが存在していた場合，そのドット(ドット b)からさらに X 画素離れた距離にある，ドット存在期待領域を探索し手順(7)の方法で採点する。

ドット存在期待領域に複数点のドットが存在していた場合，読み込んだ画像内のドットの位置の画素について画素値が小さいものを選び出し，その位置からさらに X 画素離れた距離にある，ドット存在期待領域を探索し採点するドット存在領域にドットが存在しない場合は，ドット存在期待領域の中心から X 画素離れた距離にあるドット存在期待領域を探索し採点する。得られた採点結果を，ドット a の評価点として加算する。

(9) 手順(7)，(8)の作業をドット a から 5X 画素離れた距離のドット存在期待領域に達するまで繰り返す(図 4- 14(b))。

図 4- 14 は右方向へ探索を行っている図である。読み込んだ画像の左半分に位置する上下端の領域では右方向へ探索を行い，水平方向に並ぶ候補位置規定点を抽出する。また，右半分の

上下端の領域では同様の処理を左方向へ行い水平方向に並ぶ候補位置規定点を抽出する。同様に，上半分の左右端の領域では下方向へ探索を行い垂直方向に並ぶ候補位置規定点を抽出する。下半分の左右端の領域では上方向へ探索を行い垂直方向に並ぶ候補位置規定点を抽出する（図4- 15）。

上記方法では，ドット存在期待領域の大きさにより，スキャナを用い読み込んだ際に発生する画像の傾きへの許容度が決定される。ドット存在期待領域が大きくなれば，画像の傾きや歪み，印刷時に発生したドットのずれ等が候補位置規定点に発生した場合もドットの抽出が可能となるが，候補位置規定点以外の点を候補位置規定点として誤って読み込む可能性が高くなるため，必要なデータ抽出の精度に合わせてドット存在期待領域の大きさを適切なものにする必要がある。

242	240	241	240	240
240	141	119	125	235
235	145	108	122	234
239	148	116	126	235
240	240	242	241	241

(a) 画像内の画素値

	-2	1	-1	0
	-99	-22	6	110
	-90	-37	14	112
	-91	-32	10	109
	0	2	-1	0

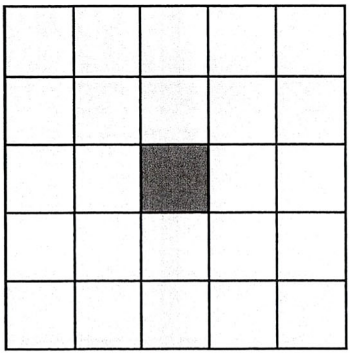
(b) 横方向の差分

	0	0	0	0
	-	-	0	+
	-	-	+	+
	-	-	+	+
	0	0	0	0

(c) 3種類へ分類

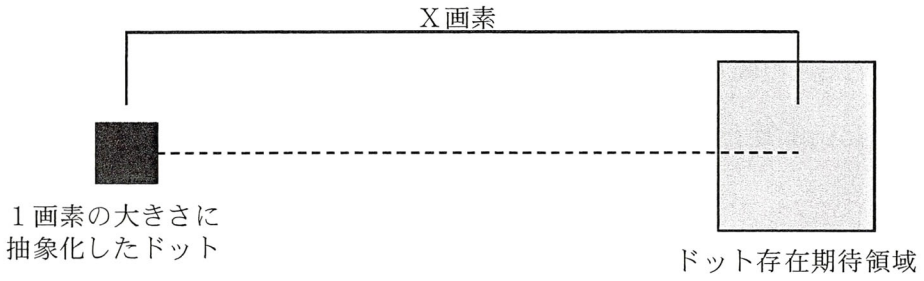
	141	119	125	235
	145	108	122	234
	148	116	126	235

(d) ドットを構成する画素だと予想される領域

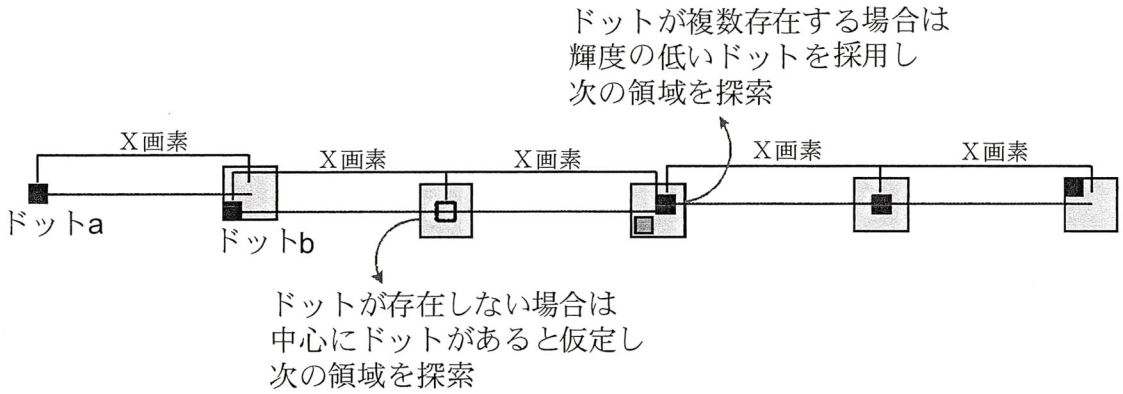


(e) ドットの位置を示す画素

図4- 13 ドットの位置の決定



(a) ドット存在期待領域



(b) ドット存在期待領域の決定

図 4- 14 連続性の評価

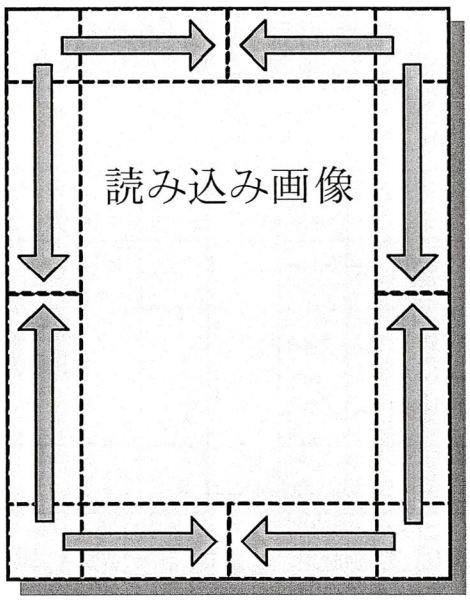


図 4- 15 連続性の評価方向

４．５．２　データ埋め込み位置の決定とデータの抽出

抽出した候補位置規定点を用いてデータ埋め込み候補領域を選び出す。まずはじめに，候補位置規定点を対にする。そのために，下端に位置する各候補位置規定点から左右に任意の画素数分，幅を持たせて垂直方向にドットを探索し，上端に位置する候補位置規定点を，対となるものとして選び出す。同様に左端に位置する各候補位置規定点の対になるドットを右端から選び出す。それぞれ各候補位置規定点を上下，左右で対にしたのち，それらの各２点を直線で結びつける。直線はブレゼンハムのアルゴリズムを利用して作成する[69]。直線を作成後，生成された交点を候補領域中心点とし，その点の周囲を含む画素がデータ埋め込み候補領域であると判断し，その領域内の濃淡値によりデータが埋め込まれているかどうかや，どのようなデータが埋め込まれているかを判断する。印刷文書は作成時に 300dpi の画像により作成され，その文書を 600dpi で読み込んでいる。そのため理想的には，データ埋め込み候補領域は 24×24 ピクセルの大きさに読み取られているはずである。そこで，候補領域中心点として抽出した画素を中心にし，図 4-16 (a) に示すように２４ピクセル四方の領域を設定する。また，その領域を４つに分割し作成される，１２ピクセル四方の領域を小領域とする。そして，設定した領域から埋め込まれているデータを以下の手順で抽出する。

(1) 図 4-16 (b) に示す，中心の４画素四方にドットが配置されているかどうかを調べ，配置されていた場合はその候補領域内にはデータが埋め込まれていないと判断する。

(2) 作成時に候補位置規定点近傍に誤り訂正符号を配置している場合，それを取り出し手順 (1) で得られた結果と照合し，誤りが見つかった場合は訂正する。

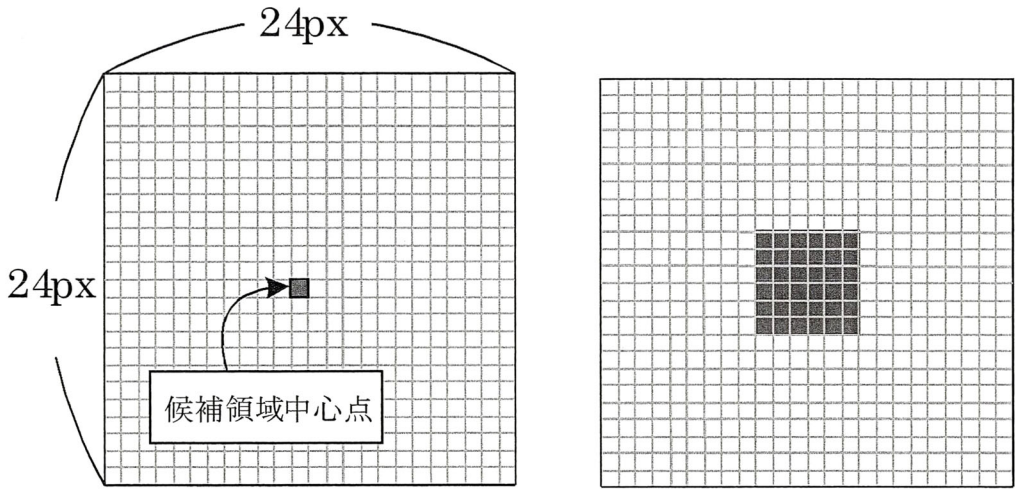
ここまでで，データが埋め込まれていないと判断された領域については以下の処理は行わない。

(3) 図 4-16 (c) の濃く示した位置にドットが配置されているかどうかを調べ，配置されていた小領域に対して次の処理を実行する。

(4) 図 4-16 (d) の濃く示した位置にドットが配置されているかどうかを調べ，配置されていない小領域が存在した場合，その箇所にデータが存在すると考えられる。

(5) データが存在すると判断した小領域の位置が，左下である場合“00”が，右下である場合“01”，左上である場合“10”，右上である場合を“11”としデータを取り出す。

なお，ここでドットが配置されている画素かどうかを調べるためには，2×2 四方の画素の画素値の合計が閾値５１２以下ならドットが存在するとしている。これは，４画素のうち最低２画素は画素値の小さい画素がある場合をドットが存在していると判断するために設定した値である。以上の手順を用い取り出したデータを基に，データ自体に付与されている誤り訂正符号を利用し，データの誤り訂正を行う。



(a) 候補領域中心点とデータ埋め込み候補領域の位置　　(b) 中心付近に配置された探索領域

(c) データドットの探索位置　　(d) 周囲付近に配置された探索領域

図 4- 16 データ取り出しのための探索領域

４．５．３　誤り訂正符合を利用したデータ抽出精度の向上

提案手法では，印刷文書からデータを抽出する際に，誤り訂正符号を用いることでデータ抽出精度の向上を図っている。本節では誤り訂正符号を使用した印刷文書と，使用していない印刷文書のデータ抽出精度を示して比較する。

今回は，誤り訂正符号の利用による抽出精度の変化のみに注目するために，誤り訂正を用いた印刷文書と用いていない印刷文書を複数枚用意し，それぞれを別々に読み込み，抽出精度を比べるという方法は採用せず，以下の方法で比較を行った。

まず，５種類の画像を用意する。用意した画像は，候補領域規定点近傍への誤り訂正符号を含まない印刷文書１枚(印刷文書 a)と，以下に挙げる設定の候補位置規定点近傍への誤り訂正符号を配置した印刷文書４枚(ｂ～ｅ)である。

印刷文書ｂ：上下左右の候補位置規定点近傍に１０ビットのデータを配置した印刷文書

印刷文書ｃ：左右の候補位置規定点近傍に１０ビットのデータを配置した印刷文書

印刷文書ｄ：上下左右の候補位置規定点近傍に２０ビットのデータを配置した印刷文書

印刷文書ｅ：左右の候補位置規定点近傍に２０ビットのデータを配置した印刷文書

なお，印刷文書ａと印刷文書ｂ～ｅを比較した場合，候補位置規定点近傍のドットの配置を除けばそれらに差異はない。

比較は，まず，印刷文書ａをスキャナで読み込み，データを取り出す。次に，印刷文書ｂ～ｅの４枚をスキャナで読み込み，候補位置規定点近傍に配置された誤り訂正のためのデータを取り出す。印刷文書ｂ～ｅから取り出した誤り訂正のためのデータを，印刷文書ａから取り出したデータに適用し，誤り訂正を行う。

このようにして，誤り訂正符号利用の有無によるデータ抽出精度の比較を行う。

本節でデータ抽出精度比較のために利用した画像は，候補位置規定点を２４画素ごとに配置することにより縦１３０点，横８８点の候補位置規定点を作成したものである。出力に使用したプリンタは OKI MICROLINE 9500PS-F(カタログによる最大印刷解像度 1200dpi)であり，印刷品位を「きれい」，カラーモードを「グレースケール」に設定し，モノクロで出力したものであり，それをスキャナ(EPSON GT-9800F)により，600dpi で読み込んだ。この印刷文書には，３５

４箇所のデータ埋め込み領域が存在している。抽出結果は，表 4- 1 に示す結果となった。表 4-1 における横の項目の「余分な点」とは，データの埋め込まれていないデータ埋め込み候補領域を誤ってデータ埋め込み領域として読み込んだ箇所のことである。また，「欠落した点」とはデータ埋め込み領域に配置されたデータを読み逃し，データの埋め込みがその領域には無いと判断した箇所のことである。「誤った点」とはデータ埋め込み領域から取り出したデータが誤ったデータだった場合の箇所のことである。「誤り訂正無」は候補位置規定点付近の誤り訂正符号を利用せずにデータを読み込んだ場合の抽出誤りの個数である。「誤り率」は，「余分な点」，「欠落した点」および「誤った点」の合計数を総情報ドット数で割った値である。表 4- 1 の縦の項目の「上下左右 10bit」は，上下左右の候補位置規定点付近に配置された１０ビットの誤り訂正のためのデータを利用した場合での抽出誤りの個数である。同様に「左右 10bit」は左右に配置された 10 ビットのデータを，「上下左右 20bit」は上下左右に配置された 20 ビットのデータを，「左右 20bit」は左右に配置された２０ビットのデータを利用した場合のものである。

誤り訂正を用いることで，データの読み込み時に発生した誤りの個数が４２～６０％減少している。中でも「余分な点」の誤り個数は大きく低下している。これは，候補位置規定点近傍に配置したデータ埋め込み位置に対する誤り訂正符号による効果が高いと考えられる。また，「誤った点」の個数の低下は，埋め込まれたデータに直接付与された誤り訂正符号によるところが大きいと考えられる。

「欠落した点」が発生する箇所は，候補領域中心点のズレが大きく発生している可能性が高く，データ埋め込み領域内のドットの配置を用いた正確なデータの取り出しは難しい。そのような場合，埋め込まれたデータを正しく復号するために利用することができるのは誤り訂正符号だけとなる。そのため，データ復号時に参考にすることができるデータが少なく誤りの訂正が難しいことが，改善率の低さの原因と考えられる。

表 4- 1　候補位置規定点近傍へ配置された誤り訂正符号の利用による抽出誤り個数の変化

誤り訂正	余分な点	欠落した点	誤った点	誤り率
誤り訂正無	15	20	5	11%
上下左右 10bit	1	14	1	5%
左右 10bit	5	16	2	6%
上下左右 20bit	0	13	1	4%
左右 210bit	4	13	1	5%

４．６ 印刷文書の被験者実験による評価

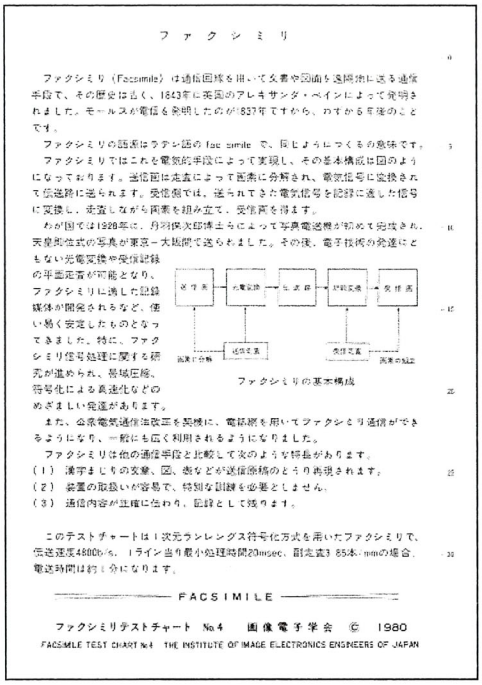
４．６．１　実験の概要

提案手法により作成した印刷文書に対して，データの埋め込みが違和感を与えるかどうか調査する。具体的には，基準点の間隔，誤り訂正用検査ビットの配置量等を変化させた印刷文書を作成し，それらを用いて被験者実験を行った。また，背景部分が文書の読み易さに与える影響も併せて調査した。実験は，２０歳から２５歳の男女２０人に対して，表 4- 2 に示す 7 種類の方式で作成した画像をランダムに提示し，質問項目について評価させた。使用する画像はすべて同じ文書画像を用い，データを埋め込む地紋の元となるイラストは図 4- 17 に示す画像を使用した。方式１は，文書画像にイラストを白黒濃淡化し重ねたものであり，データは埋め込んでいない。また方式２～７は，イラストをデータを含む地紋に変換し，文書画像に付加したものである。図 4- 18 (a) は方式５による画像全体を縮小表示したものであり，図 4- 18 (b) はその一部を実寸大となるよう拡大したものである。方式１から７による画像の例を図 4- 19 に示す。なお，実物大の方式１から７による画像は付録Ｃに示す。

表 4- 2　実験使用画像の種類

方式	基準点近傍の誤り訂正用検査ビット量	基準点の間隔
1	配置無し	配置無し
2	配置無し	２３画素
3	配置無し	４７画素
4	１０ビット	２３画素
5	１０ビット	４７画素
6	２０ビット	２３画素
7	２０ビット	４７画素

図 4- 17　地紋に用いたイラスト画像



ファクシミリ

ファクシミリ（Facsimile）は通信回線を用いて文書や図面を遠隔地に送る通信手段で、その歴史は古く、1843年に英国のアレキサンダ・ベインによって発明されました。モールスが電信を発明したのが1837年ですから、わずか6年後のことです。

ファクシミリの語源はラテン語の fac simile で、同じようにつくるの意味です。

ファクシミリではこれを電気的手段によって実現し、その基本構成は図のようになっております。送信面は走査によって画素に分解され、電気信号に変換されて伝送路に送られます。受信側では、送られてきた電気信号を記録に適した信号に変換し、走査しながら画素を組み立て、受信面を得ます。

わが国では1928年に、丹羽保次郎博士らによって写真電送機が初めて完成され、天皇即位式の写真が東京－大阪間で送られました。その後、電子技術の発達にともない光電変換や受信記録の平面走査が可能となり、ファクシミリに適した記録媒体が開発されるなど、使い易く安定したものとなってきました。特に、ファクシミリ信号処理に関する研究が進められ、帯域圧縮、符号化による高速化などのめざましい発達があります。

ファクシミリの基本構成

また、公衆電気通信法改正を契機に、電話線を用いてファクシミリ通信ができるようになり、一般にも広く利用されるようになりました。

ファクシミリは他の通信手段と比較して次のような特長があります。

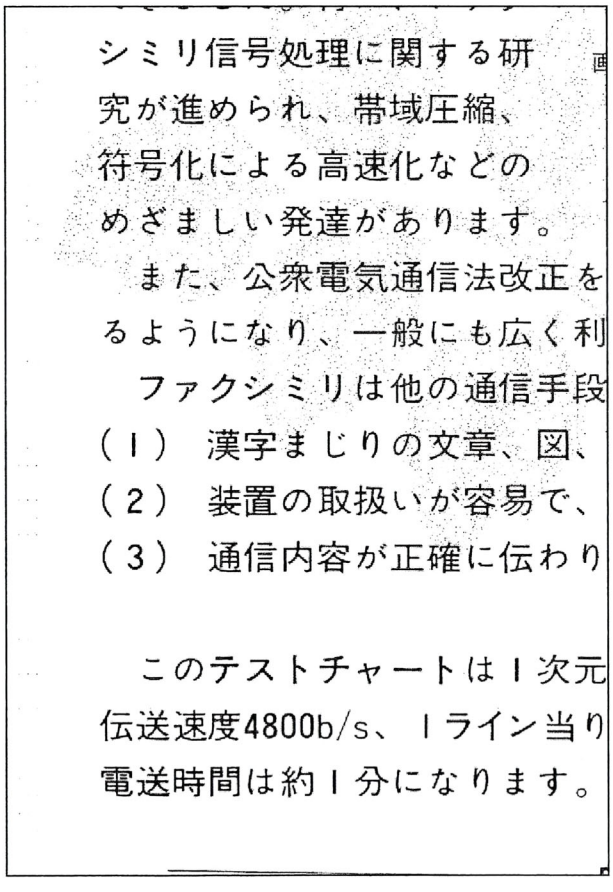
（1） 漢字まじりの文章、図、表などが送信原稿のとうり再現されます。
（2） 装置の取扱いが容易で、特別な訓練を必要としません。
（3） 通信内容が正確に伝わり、記録として残ります。

このテストチャートは1次元ランレングス符号化方式を用いたファクシミリで、伝送速度4800b/s、1ライン当り最小処理時間20msec、副走査3.85本/mmの場合、電送時間は約1分になります。

FACSIMILE

ファクシミリテストチャート No.4　画像電子学会 ©　1980
FACSIMILE TEST CHART No.4　THE INSTITUTE OF IMAGE ELECTRONICS ENGINEERS OF JAPAN

(a)　方式5による画像



シミリ信号処理に関する研
究が進められ、帯域圧縮、
符号化による高速化などの
めざましい発達があります。
また、公衆電気通信法改正を
るようになり、一般にも広く利
ファクシミリは他の通信手段
（1） 漢字まじりの文章、図、
（2） 装置の取扱いが容易で、
（3） 通信内容が正確に伝わり

このテストチャートは1次元
伝送速度4800b/s、1ライン当り
電送時間は約1分になります。

(b)　方式5による画像の一部

図4- 18　実験使用画像の例

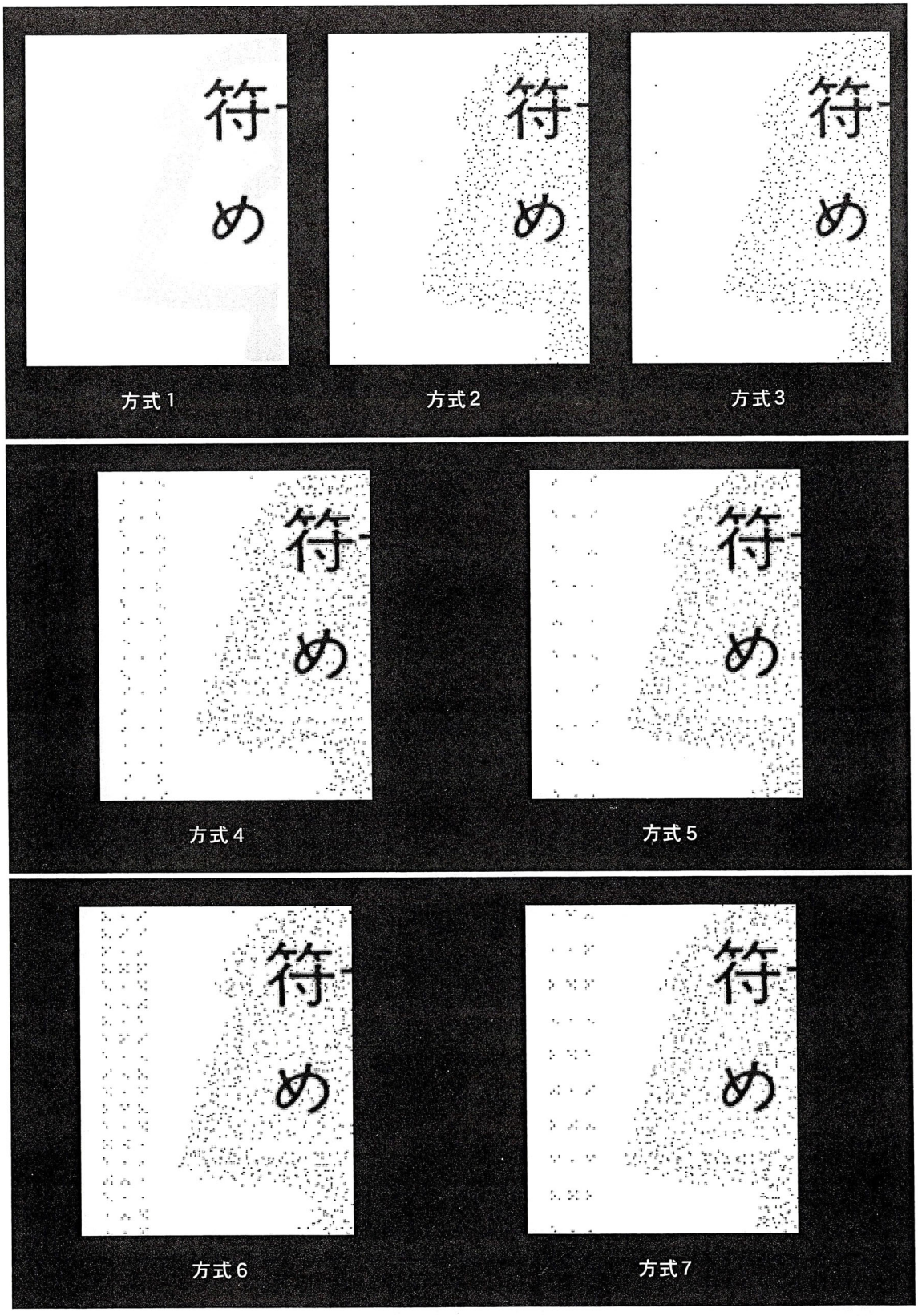


図 4- 19　実験使用画像の例

評価は以下の３項目に対して行った。

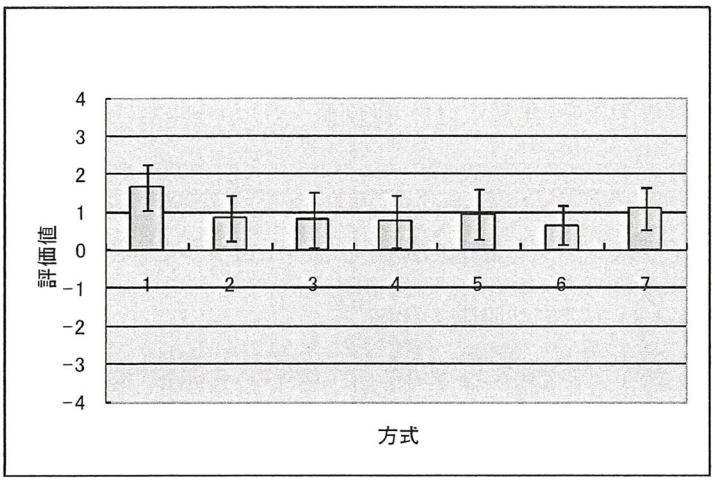
(1) 紙面全体から受ける印象を，非常に自然（評価値＋３）から非常に不自然（－３）までの７段階で評価

(2) 背景に存在するイラストが文章認識の障害となるかどうかを，全く気にならない（＋３）から非常に気になる（－３）までの７段階で評価

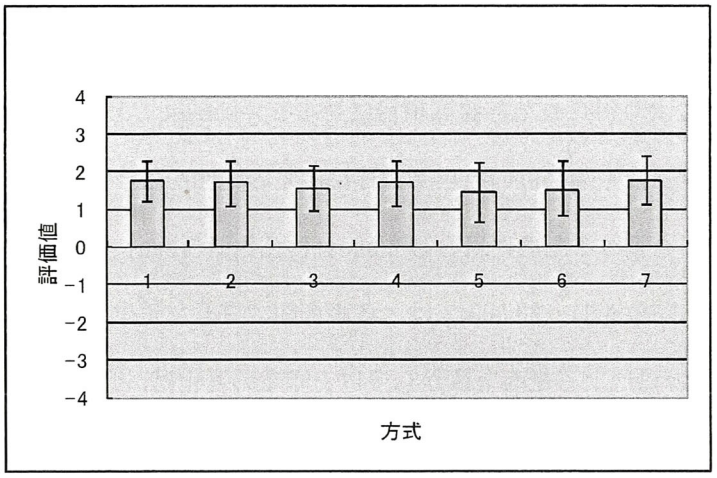
(3) 印刷文書の縁に沿うように配置されたドットが気になるかを，全く気にならない（＋３）から非常に気になる（－３）までの７段階で評価

４．６．２　実験結果と考察

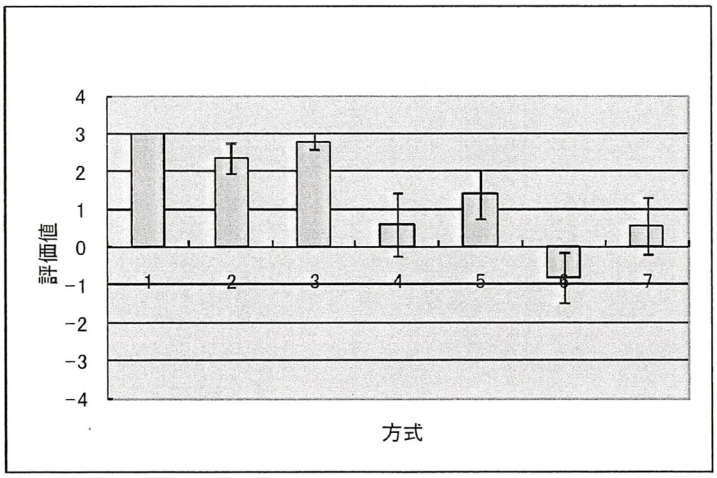
図4-20は，実験により得られた各質問項目に対する評価結果である。棒グラフは評価値の平均値を表し，棒グラフの先端に付いたヒゲは標準偏差の範囲を表す。各画像の評価値間で有意な差があるかどうかを調べるためにＨＳＤ（Honestly Significant Difference）検定を行った。紙面全体から受ける印象（図4-20(a)）では，各画像間で有意な差は見られなかった。したがって，データの埋め込み量による紙面全体から受ける印象の変化は，ほとんどないと考えてよい。背景のイラストが文章認識の障害となるかどうかについても，各画像間で有意な差は表れず，全体的に高い評価が得られた（図4-20(b)）。



(a) 「紙面全体から受ける印象」の評価結果



(b) 「背景が文章認識の障害となるかどうか」の評価結果



(c) 「帳票周囲のドットが気になるか」の評価結果

図 4- 20 被験者実験の結果

評価結果から，今回の例については，背景に配置された地紋により文章が読みにくくなるという問題は発生していないと考えてよい。印刷文書の周囲に配置されたドットについては，質問項目３(図4-20(c))において，周囲にドットを配置していない方式１と，方式２，３の間には有意な差が見られなかった。印刷文書の周囲に誤り訂正用検査ビットを配置すると評価値が低下するが，方式２と方式５の間では有意な差が見られなかった。また，紙面全体の印象の変化(図4-20(a))では，誤り訂正用検査ビットの量が変化した場合も，有意な差は見られなかった。検査ビットの増減により，紙面から受ける印象は若干変化するが，大きなものではない。以上のことから，周囲のドットによる影響を，誤り訂正用のデータを配置していない方式２程度に収め，データ抽出精度の向上を目指す場合は，方式５程度の誤り訂正符号用データ量が適切であると考えられる。

４．７　総合評価

被験者実験において，提案手法により作成した文書で，背景にドット地紋を配置しデータを埋め込んだ場合も，印刷文書上の文章や図形の認識に障害とはならないことが示された。また，紙面周囲に配置される誤り訂正用検査ビットを表すドットは，ドットの数が多くなるほど，見た目の評価が低下する。しかし，画像全体を見た場合に受ける印象については，評価値の低下は小さい。誤り訂正符号を用いることで，データ抽出時に発生した誤りのうち５０％程度を訂正可能であり，抽出精度向上の効果を考慮すると，誤り訂正用検査ビットを配置することが望ましいと考える。

実験に使用した画像は，基準点の間隔が２３画素のものは，紙面上の３５４か所にデータドットを配置できる。すなわち，７０８ビットのデータを保持できる。そのうち半分は，誤り訂正用に利用するため，実際に埋め込み可能なデータ量は３５４ビットである。

以上のことから，埋め込むデータが１０～２０バイトに収まる場合に関しては，印刷文書の見た目と，誤り訂正のバランスを考え，実験時の方式５の設定を用いることが望ましいと考える。それより多くのデータ量が必要な場合は，データ埋め込み画像の中で方式５の次に評価が高い方式４の設定を使用して，印刷文書の作成をすることが望ましいと思われる。

４．８　結言

本章では，印刷文書の内容に影響を与えることの少ないイラストを地紋に用いその中へデータを埋め込むことで，利用者がデータの埋め込まれた印刷文書から受ける違和感を軽減する手法を提案した。また，印刷文書からのデータ抽出精度を向上させるための，誤り訂正符号の導入方法を提案した。今後の課題として，地紋に用いたイラスト細部の絵画的な特徴の再現性の改善や，実用に向けて抽出精度１００％を実現する方法，より多量のデータを埋め込み可能な手法，違和感がより少ない誤り訂正用データの配置方法の開発等が挙げられる。

第５章　結論

５．１　本研究の成果

本論文では，文書のライフサイクル管理における収集，分析，表示・印刷の各フェーズに関する以下の３つの課題を解決することを目的とした。

(1) 外国語文書の収集

グローバル化の進展に伴い，外国語で書かれた文書の重要度が増している。より多くの文書を収集し，活用するためには，今後，外国語を含んだ文書を収集し，共有できることが必要である。外国語の中でも，特に，近年，文化経済の面での交流が飛躍的に伸びているアジア圏の言語の重要性が増している。

(2) 非構造化文書の分析

従来から，ＯＬＡＰを主体とする構造化文書向けの分析手法が使用されてきたが，多くの文書が非構造化文書として未活用のまま眠っていると言われている。大量の非構造化文書を対象とした新たな分析手法が必要である。特に，企業，団体，官公庁等の組織では，空間的，時間的な観点での分析ニーズが高い。

(3) 印刷文書の管理

検索・分析結果を活用する際に，セキュリティ面から文書の管理が必要である。電子的な文書のセキュリティ管理はこれまでいろいろな技術が提案されているが，印刷文書の管理に関しては充分な対策がなされていない。実際，漏えいの媒体としては，紙文書がインシデント件数の大多数（７２．６％）を占めており，対策が必要である。印刷文書の中に管理データを埋め込むことが有効な対策であると考えられる。

外国語の文書の収集・共有に関しては，機械翻訳技術の活用が有効である。従来から，日本語－英語間の機械翻訳が研究されており，同じ方式でアジア圏の言語も機械翻訳することが試みられていた。本論文では，アジア圏の言語と日本語の部分的な類似性を反映できる翻訳方式を提案した。具体的には，中国語を例にとり，トランスファ方式をベースに３段階のトランスファレベルを設ける方式を提案した。これにより，翻訳ルールを減少させ，効率的な翻訳が可能となる。また，本方式は，ユーザ自身が辞書，熟語等のデータを追加して質を向上できる仕掛けも設けており，簡単に翻訳誤りを訂正し，次回以降の翻訳時に同じ誤りを起こさないこと

も可能とした。

非構造化文書の分析に関しては，大量のテキストデータから固有表現を自動抽出して，データマイニングおよび統計処理を行う方式を提案した。具体的には，約５０００件の防犯メールデータから，事件種別，場所，時間，等の情報を抽出し，アソシエーションルールおよび決定木から事件と時間の関係ルールを導いたり，ランドマークまでの距離と事件との関係性を統計処理により見出すことができた。ベテランの分析者であれば，データを眺めていると分かることもあるが，自動的に事件の傾向を得ることができず，分析者によるばらつきが発生する。提案手法を用いることにより，分析者によるばらつきを防ぎ，大量のデータを短時間で分析することが可能になると期待される。

印刷文書の管理に関しては，印刷文書の内容に影響を与えることの少ないイラストを地紋に用いその中へデータを埋め込むことで，利用者がデータの埋め込まれた印刷文書から受ける違和感を軽減する手法を提案した。また，印刷文書からのデータ取り出し時の精度を向上させるための，誤り訂正符号の導入方法を提案した。

５．２　今後の課題

本論文で述べた内容は，いわゆる人工知能と呼ばれる分野にも含まれる翻訳や分析等の人間が行う作業の半自動化を目指すものが含まれており，課題は尽きることがない。企業活動等の実社会で役立つかどうかは，これらの技術の導入によりコスト削減や利益増大が実現できるかという経済合理性の観点でのみ判断されるものであり，技術が完全でなくとも活用できる部分は少なくないと考えている。

比較的近未来で取り組めそうな今後の課題としては，以下を挙げることができる。

(1) 機械翻訳について

究極的には翻訳の質の向上があるが，ユーザが自分で翻訳誤りを直して学習させる手法がまず必要である。人間と同レベルの翻訳を機械が行うのは，当面，不可能であるので，少なくともユーザがシステム開発者の手を借りることなく，各種データを更新して翻訳の質を向上させることが有効である。本論文では，一部のデータを更新できる手法を実現したが，他にもユーザが自由に更新して翻訳の質を向上できるような手法がないか，今後の課題として検討したい。

また，本論文では，中国語を例にしたが，他の言語での有効性検証も必要である。近年，ベトナムやインドネシア等のアジア諸国の経済発展が著しく，これらの国々の言語での評価を優先的に行うべきではないかと思われる。

(2) 分析について

分析についても，精度向上により，プロの分析者に近づくことが究極の課題である。具体的には，テキスト解析の精度向上によるさらに多くの固有表現の高精度抽出，地図等を使った分析結果の分かりやすい表示方法，分析による予測等がある。

(3) 印刷管理について

地紋に用いたイラスト細部の絵画的な特徴の再現性の改善や，実用に向けて抽出精度をさらに向上させる方法，より多量のデータを埋め込み可能な手法，見た目へ与える影響がより小さい誤り訂正用データの配置方法の開発等がある。

少子高齢化，人口減少時代を迎え，日本が今までと同等以上に成長していくためには，様々な組織に属する一人一人が，今まで以上の知識を持ち，その知識を活用することにより，国際競争力を維持していくことが必要である。組織における文書管理，知識管理の仕組みをより高度にしていくことが，そのための一助となると信じている。

謝辞

本論文は，筆者が㈱日立製作所および山口大学大学院理工学研究科で研究した成果をまとめたものである。山口大学大学院理工学研究科多田村克己教授には，論文作成を薦めていただくと共に，本研究をまとめるに当たり，終始懇切丁寧にご指導くださり深く感謝いたします。また，山口大学大学院理工学研究科三池秀敏教授，大林正直教授，山口大学大学情報機構メディア基盤センター市川哲彦教授ならびに山口大学大学院理工学研究科水上嘉樹准教授には，論文に対して多くの有益なご示唆をいただきました。深く感謝いたします。

多言語処理に関しては，㈱日立製作所限井裕之氏に有益な議論をいただきました。また，地紋透かしの方式設計，プログラム開発，評価に関しては，三輪智也氏（当時山口大学大学院博士前期課程）に多大なるご協力をいただきました。非構造化文書の分析に関しては，山崎竜平氏（山口大学大学院博士前期課程）にツール作成，表示プログラムの開発にご協力いただきました。深く感謝いたします。ここに記載していない多くの方々にも，これまで研究開発の進め方をご教示いただいたり，論文執筆を激励していただき，感謝いたします。

最後に，社会人にもかかわらず，仕事とは別に論文をまとめたいというわがままを聞いてもらい理解して見守ってくれた妻裕子といつも元気に声援を送ってくれた子供たち純奈，理央奈に感謝します。

参考文献

[1] 吉岡真治, 矢入郁子：情報爆発時代に向けた新しい IT 基盤の研究, 人工知能学会誌, Vol. 22, No. 2, pp. 208-240 (2007)

[2] 柴山悦哉, 鳥澤健太郎, 田浦健次朗, 河野健二：情報爆発時代におけるわくわくする IT の創出を目指して, 情報処理, Vol. 49, No. 8, pp. 880-955 (2008)

[3] R. H., Jr. Sprague, R. H.: Electronic Document Management: Challenges and Opportunities for Information Systems Managers, *MIS Quarterly*, Vol. 19, No. 1, pp. 29-49 (1995)

[4] 野中郁次郎, 竹内弘高, 梅本勝博：知識創造企業, 東洋経済新報社 (1996)

[5] 社団法人ビジネス機械・情報システム産業協会：文書管理システム導入のすすめ, 東洋経済新報社 (2006)

[6] 石井啓豊：一次情報の収集･整理･保管･廃棄, 情報管理, Vol. 35, No. 2, pp. 129-143 (1992)

[7] 山下貞麿：ナレッジマネジメントと記録管理, 情報管理, Vol. 49, No. 3, pp. 132-141 (2006)

[8] 木谷強, 相原理, 高木徹：新時代における情報提供術:全文データベースの事例紹介, 情報管理, Vol. 41, No. 6, pp. 460-470 (1998)

[9] ＮＰＯ日本ネットワークセキュリティ協会：情報セキュリティインシデントに関する調査報告書 (2009)

[10] Incept Inc. : ＩＴ用語辞典, http://e-words.jp/

[11] AIIM. : What is Enterprise Content Management (ECM) ?, http://www.aiim.org/What-is-ECM-Enterprise-Content-Management

[12] FUJI XEROX. : DocuWorks, http://www.fujixerox.co.jp/product/software/docuworks/

[13] Hitachi, Ltd. : 文書管理基盤 uCosminexus DocumentBroker, http://www.hitachi.co.jp/Prod/comp/soft1/docbro/index.html

[14] RICOH: Ridoc Document System, http://www.ricoh.co.jp/ridoc_ds/rds/

[15] Open Text Corporation: OPEN TEXT, http://www.opentext.jp/

[16] OSK Co., LTD.: 文書管理システム Visual Finder, http://www.evalue.jp/pro/vf/default.asp

[17] Weblio, Inc.: ＩＴ用語辞典, http://www.sophia-it.com/

[18] 清兼義弘, 関口宏司, 田澤孝之, 松野良蔵：エンタープライズサーチ　技術と導入, アスキー・メディアワークス(2008)

[19] ファスト サーチ&トランスファ株式会社：FAST Search Server 2010 for Internet Sites, http://sharepoint.microsoft.com/ja-jp/product/capabilities/search/Pages/fast.aspx

[20] Hitachi, Ltd.: 全文検索エンジン「HiRDB Text Search Plug-in」, http://www.hitachi.co.jp/Prod/comp/soft1/textsearch/product/component/hirdb_ts/index.html

[21] JustSystems Corporation: ConceptBase Enterprise Search, http://just-enterprise.com/product/cbes/cbes01.html

[22] Autonomy Inc.: エンタープライズ検索, http://www.autonomy.co.jp/enterprise-search

[23] アクセラテクノロジ株式会社：検索システム Accela BizSearch　概要, http://www.accelatech.com/products/BS/index.html

[24] 中野康明：文字認識・文書理解の最新動向［Ⅱ］, 電子情報通信学会誌, Vol.83, No.2, pp.143-148(2000)

[25] 田町常夫：機械翻訳の概要と歴史, 情報処理, Vol.26, No.10, pp.1140-1147(1985)

[26] 長尾真：機械翻訳はどこまで可能か, 岩波書店(1986)

[27] 野美山浩：事例の一般化による機械翻訳, 情報処理学会論文誌, Vol.34, No.5, pp.905-912(1993)

[28] CICC.: 近隣諸国間の機械翻訳システムに関する研究協力, http://www.cicc.or.jp/japanese/kyoudou/mt.html?PHPSESSID=8e1d0e19ac37d4d5293c5e6a31b3d604

[29] Koehn P., Och F.J. and Marcu D.: Statistical Phrase-Based Translation, *Proc. of HLTNAACL*, pp. 127-133 (2003)

[30] Moses -statistical machine translation system: http://www.statmt.org/moses/

[31] Rakuten, Inc.: Infoseekマルチ翻訳, http://translation.infoseek.co.jp/

[32] Google: Google翻訳, http://www.google.co.jp/language_tools?hl=ja

[33] Excite Japan Co., Ltd.: excite翻訳, http://www.excite.co.jp/world/

[34] 竹本義美, 山田洋志, 福島俊一：日本語テキストからの固有表現抽出システムの開発と評価, 情報処理学会第５９回全国大会, pp.″2-323″-″2-324″ (1999)

[35] Feldman R. and Sanger J., The Text Mining Handbook, Cambridge University Press (2007)

[36] Agrawal R. and Srikant R.: Fast Algorithms for Mining Association Rules, *VLDB*, pp. 487-499 (1994)

[37] 鈴木英之進, 鹿島久嗣：最新！データマイニング手法, 情報処理, Vol. 6, No. 1, pp. 2-51 (2005)

[38] 高阪宏行, 関根智子：ＧＩＳを利用した社会・経済の空間分析, 古今書院 (2005)

[39] 中村高雄, 小川宏, 高嶋洋一：ディジタル画像の著作権保護の為の周波数領域における電子透かし方式, 暗号と情報セキュリティシンポジウム, SCIS97-26A (1997)

[40] 萩原剛志, 金田悠紀夫:画像の２値化過程で情報埋め込みを行う手法の改良について, SCIS99-T4-2.5 (1999)

[41] 阿部悌, 井上浩一：２値画像への電子透かし, Ricoh Technical Report, Vol. 26, pp. 8-16 (2000)

[42] 画像電子学会：電子透かし技術―ディジタルコンテンツのセキュリティ, 東京電機大学出版局 (2004)

[43] XEROX Co.: InfraredMark Specialty Imaging Font, http://www.xerox.com/go/xrx/template/inv_rel_newsroom.jsp?app=Newsroom&ed_name=NR_2007Oct4_InfraredMark_Specialty_Printing&format=article&view=newsrelease&Xcntry=USA&Xlang=en_US

[44] 沖電気工業株式会社：Val-Code, http://www.oki.com/jp/FSC/valcode/

[45] 中山順子, 宮増フラミニア, 宮尾真理子：英語とインターネット, 東京家政学院筑波女子大学紀要第３集, pp. 77-93 (1999)

[46] 西垣通：多言語時代を迎えたインターネット, 世界, No. 10 (1997)

[47] 三浦信孝：多言語主義とは何か, 藤原書店 (1997)

[48] Miniwatts Marketing Group: INTERNET WORLD USERS BY LANGUAGE, http://www.internetworldstats.com/stats7.htm

[49] CICC: Joint Development Research on International Standardization: Multilingual Information Technology (1999)

[50] 三野昭一：中国語文法の基礎, 三修社 (1978)

[51] 加藤直樹, 羽室行信, 矢田勝俊：データマイニングとその応用, 朝倉書店 (2008)

[52] 瀧澤重志, 佐伯研, 加藤直樹：京都市伏見区におけるひったくりを中心とした犯罪空間分析, 日本建築学会学術講演梗概集, Vol. A-2, pp. 441-442 (2007)

[53] 玉田慶太：名古屋市の街頭犯罪についての要因分析, http://www.seto.nanzan-u.ac.jp/msie/gr-thesis/ms/2006/matsuda/03mm106.pdf

[54] 市村信・岡部篤行：ひったくりの空間分布と都市の諸要因との関連性についての時空間分析, 地理情報システム学会講演論文集 14, pp. 85-88 (2005)

[55] 島田貴仁, 鈴木護, 原田豊：ローカルな空間的自己相関を用いた犯罪多発地区の分析, 日本行動計量学会第 30 回大会発表論文集, pp. 88-91 (2002)

[56] 大下祐樹, 垂水共之：川口市犯罪データの空間分析, Journal of the Faculty of Environmental Science and Technology, Okayama University, Vol. 13, No. 1, pp. 17-22 (2008)

[57] 高阪宏行：板橋区における犯罪発生の空間分析,
http://nihonims.chs.nihon-u.ac.jp/f15.pdf

[58] Sekine S. and Isahara H.: IREX: IR and IE evaluation-based project in Japanese, *LREC*, pp. 1475-1480 (2000)

[59] Machine Learning Group at University of Waikato: WEKA,
http://www.cs.waikato.ac.nz/~ml/weka/index.html

[60] Frank E. and Witten I.H.; Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann (1999)

[61] 福田剛志, 徳山豪, 森本康彦：データマイニング, 共立出版 (2001)

[62] Google: GoogleMap, http://maps.google.co.jp/

[63] 日立公共システムエンジニアリング株式会社： 紙の番人,
http://www.gp.hitachi.co.jp/eigyo/product/bannin/

[64] 日立製作所：印刷用媒体への情報埋め込み装置, 情報読み取り装置および情報を埋め込んだ媒体, 公開特許公報 特開 2005-286963 (2005)

[65] 吉田健治：Drive to Web を容易に実現する新たな自動認識システム, 月刊自動認識, Vol. 18, No. 8, pp. 32-35 (2005)

[66] 吉田健治：ドットパターンを用いた情報入出力方法, 特許 第 3709385 号 (2005)

[67] 三谷政昭： やり直しのための工業数学, CQ 出版社 (2001)

[68] 高橋由泰, 青島弘和, 野山英郎：印刷用媒体への情報埋め込み装置, 情報読み取り装置および情報を埋め込んだ媒体, 特開 2005 - 286963 (2005)

[69] ＣＧ標準テキストブック編集委員会：ＣＧ標準テキストブック：技術編, 財団法人画像情報教育振興協会 (1999)

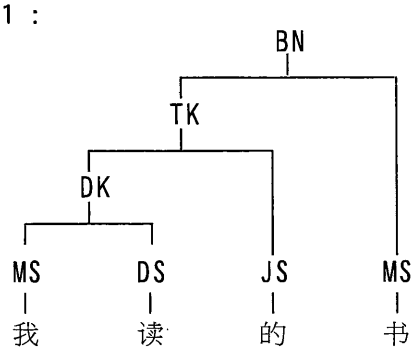
付録A　文中で引用した中国語の解説

２章で例示した中国語文について，構文と単語の意味を説明する。使用する構文ラベルは，下記のとおりである。

ラベル	記号	ラベル	記号	ラベル	記号
名詞	MS	介詞(前置詞)	KS	連体修飾句	TK
数詞	SS	動詞	DS	連用修飾句	YS
量詞	RS	形容詞	KY	熟語	JG
名詞句	MK	動詞句	DK	文	BN
助詞	JS	助動詞	JD		

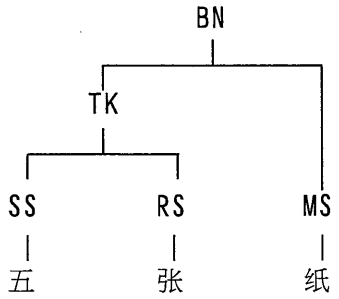
以下，各例文の構文と辞書を示す。

例１：



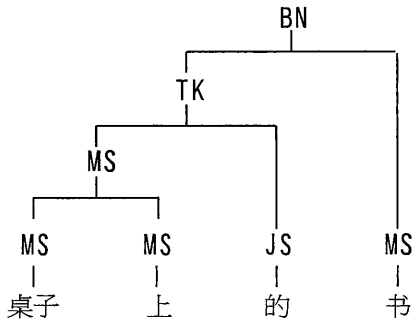
単語	意味
我	私
读	読む
的	(連体修飾助詞)
书	本

例２：



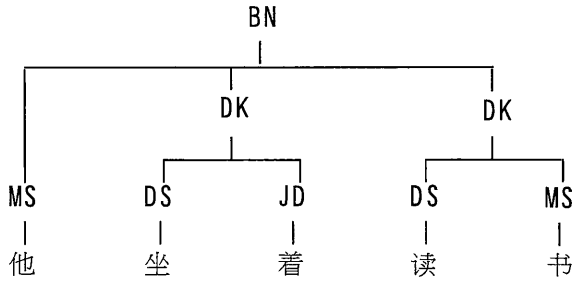
単語	意味
五	5
张	枚
纸	紙

例３：



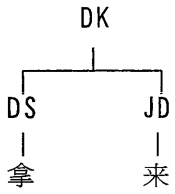
単語	意味
桌子	机
上	上
的	(連体修飾助詞)
书	本

例４：



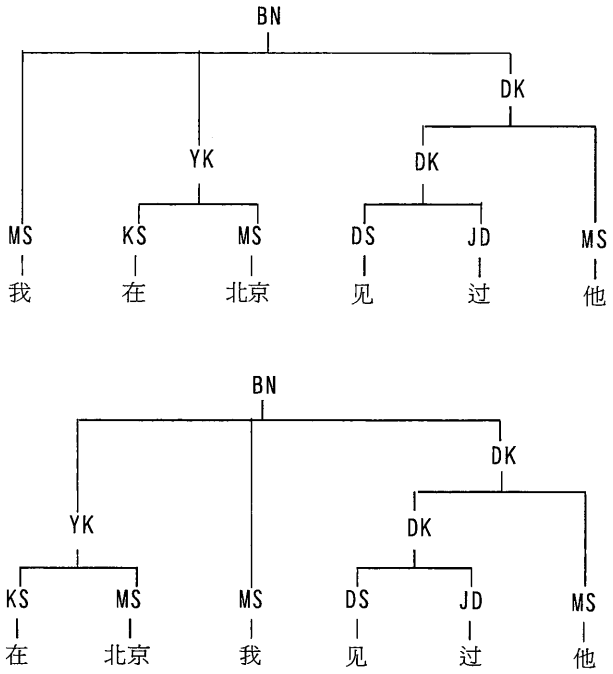
単語	意味
他	彼
坐	座る
着	「～して」を意味する助動詞
读	読む
书	本

例５：



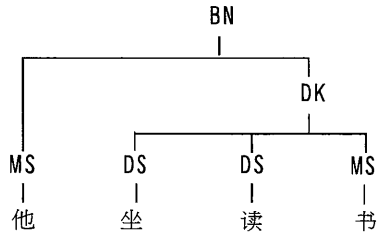
単語	意味
拿	持つ
来	「～してくる」を意味する助動詞

例６および例７：



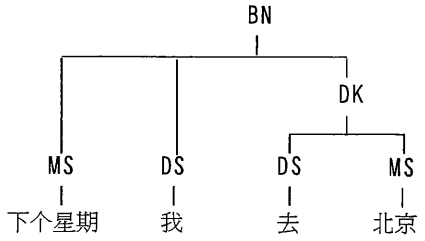
単語	意味
我	私
在	場所を意味する介詞
北京	(連体修飾助詞)
见	会う
过	完了を意味する助動詞
他	彼

例８：



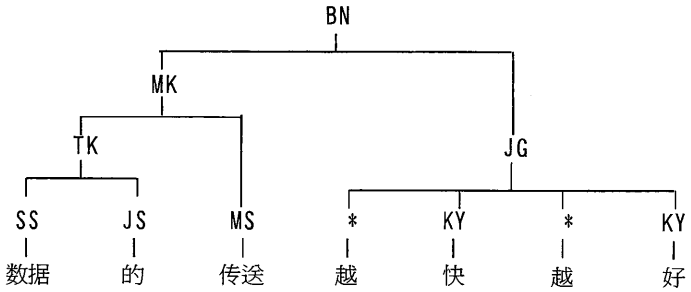
単語	意味
他	彼
坐	座る
读	読む
书	本

例９：



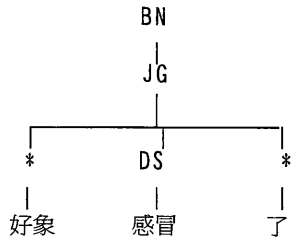
単語	意味
下个星期	来週
我	私
去	行く
北京	北京

例１０：



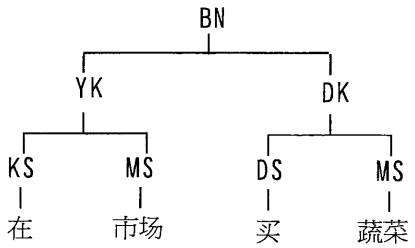
単語	意味
数据	データ
的	連体修飾助詞
传送	伝送
快	速い
好	よい

例１１：



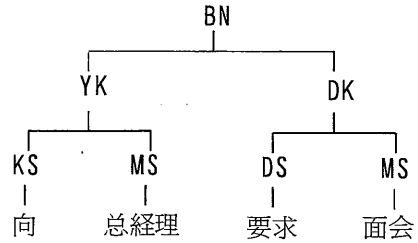
単語	意味
感冒	風邪をひく

例１２：



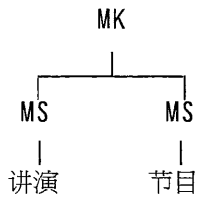
単語	意味
在	場所を意味する介詞
市场	市場
买	買う
蔬菜	野菜

例１３：



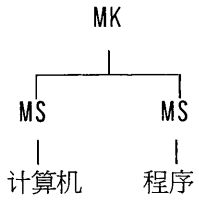
単語	意味
向	相手を意味する介詞
总経理	社長
要求	要求する
面会	面会

例１４：



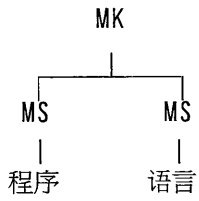
単語	意味
讲演	講演
节目	プログラム

例１５：



単語	意味
计算机	**計算機**
程序	**プログラム**

例１６：



単語	意味
程序	**プログラム**
语言	**言語**

付録B　データマイニングの結果リスト

３章で述べた「めーるけいしちょう」の分析で得られた全結果を示す。大部分は自明の結果であったが，表3- 6で述べた有効な知見として採用したのは，以下のルールである。

表3- 6　アソシエーションルールによる分析結果（再掲）

分析結果（得られた知見）	対応するルール番号
徒歩の加害者が女性に対する場合は声かけがほとんどである	1
２０/３０歳代および「若い感じ」の加害者が女性に対する場合は声かけがほとんどである	2, 4, 10
午後３時/４時に４０歳代の加害者が声かけをする相手は子どもがほとんどである	21, 22
オートバイの午前３時/午前１１時の事件はひったくりが多い	83, 91

1. higaisha=女性 vehicle=徒歩 196 ==> kind=声かけ 196　　acc:(0.99499)
2. higaisha=女性 age=２０歳代 183 ==> kind=声かけ 183　　acc:(0.99499)
3. higaisha=女性 vehicle=自転車 147 ==> kind=声かけ 147　　acc:(0.99498)
4. higaisha=女性 age=３０歳代 100 ==> kind=声かけ 100　　acc:(0.99494)
5. higaisha=女性 yohbi=火 81 ==> kind=声かけ 81　　acc:(0.9949)
6. higaisha=女性 yohbi=木 78 ==> kind=声かけ 78　　acc:(0.99489)
7. higaisha=女性 time=午前0時 75 ==> kind=声かけ 75　　acc:(0.99488)
8. higaisha=女性 yohbi=水 72 ==> kind=声かけ 72　　acc:(0.99487)
9. higaisha=女性 yohbi=金 55 ==> kind=声かけ 55　　acc:(0.99476)
10. higaisha=女性 age=若い感じ 53 ==> kind=声かけ 53　　acc:(0.99474)
11. yohbi=金 age=若い感じ vehicle=オートバイ 53 ==> kind=ひったくり 53 acc:(0.99474)
12. higaisha=女性 time=午前1時 42 ==> kind=声かけ 42　　acc:(0.99458)
13. kind=声かけ yohbi=木 age=４０歳代 39 ==> higaisha=子ども 39　　acc:(0.99451)

14. higaisha=女性 vehicle=オートバイ 38 ==> kind=声かけ 38 acc: (0.99448)

15. yohbi=日 age=若い感じ vehicle=オートバイ 38 ==> kind=ひったくり 38 acc: (0.99448)

16. time=午後6時 age=若い感じ vehicle=オートバイ 34 ==> kind=ひったくり 34 acc: (0.99435)

17. higaisha=女性 time=午後7時 33 ==> kind=声かけ 33 acc: (0.99431)

18. time=午前0時 age=若い感じ vehicle=オートバイ 29 ==> kind=ひったくり 29 acc: (0.99412)

19. higaisha=女性 time=午後11時 71 ==> kind=声かけ 70 acc: (0.994)

20. kind=声かけ yohbi=木 time=午後4時 26 ==> higaisha=子ども 26 acc: (0.99391)

21. kind=声かけ time=午後3時 age=４０歳代 26 ==> higaisha=子ども 26 acc: (0.99391)

22. kind=声かけ time=午後4時 age=４０歳代 24 ==> higaisha=子ども 24 acc: (0.99373)

23. kind=声かけ yohbi=木 time=午後3時 vehicle=徒歩 24 ==> higaisha=子ども 24 acc: (0.99373)

24. kind=声かけ yohbi=金 time=午後5時 23 ==> higaisha=子ども 23 acc: (0.99363)

25. higaisha=女性 age=４０歳代 22 ==> kind=声かけ 22 acc: (0.99351)

26. higaisha=女性 yohbi=土 61 ==> kind=声かけ 60 acc: (0.99344)

27. kind=声かけ yohbi=金 time=午後3時 21 ==> higaisha=子ども 21 acc: (0.99338)

28. yohbi=金 time=午後11時 vehicle=オートバイ 21 ==> kind=ひったくり 21 acc: (0.99338)

29. time=午前4時 vehicle=オートバイ 19 ==> kind=ひったくり 19 acc: (0.99305)

30. kind=声かけ yohbi=月 time=午後3時 19 ==> higaisha=子ども 19 acc: (0.99305)

31. kind=声かけ yohbi=水 time=午後3時 19 ==> higaisha=子ども 19 acc: (0.99305)

32. kind=声かけ yohbi=水 time=午後2時 18 ==> higaisha=子ども 18 acc: (0.99284)

33. kind=声かけ time=午後5時 age=５０歳代 18 ==> higaisha=子ども 18 acc: (0.99284)

34. kind=声かけ time=午前11時 17 ==> higaisha=子ども 17 acc: (0.99261)

35. kind=声かけ yohbi=火 age=５０歳代 17 ==> higaisha=子ども 17 acc: (0.99261)

36. yohbi=月 time=午後3時 vehicle=徒歩 17 ==> higaisha=子ども 17 acc:(0.99261)

37. time=午後5時 age=若い感じ vehicle=オートバイ 17 ==> kind=ひったくり 17 acc:(0.99261)

38. higaisha=女性 time=午後9時 51 ==> kind=声かけ 50 acc:(0.99241)

39. higaisha=女性 time=午前2時 16 ==> kind=声かけ 16 acc:(0.99233)

40. kind=声かけ time=午後5時 age=１０歳代 16 ==> higaisha=子ども 16 acc:(0.99233)

41. yohbi=日 time=午後10時 vehicle=オートバイ 16 ==> kind=ひったくり 16 acc:(0.99233)

42. yohbi=火 time=午後7時 age=若い感じ 16 ==> kind=ひったくり 16 acc:(0.99233)

43. higaisha=女性 time=午後10時 49 ==> kind=声かけ 48 acc:(0.99212)

44. higaisha=女性 time=午後6時 15 ==> kind=声かけ 15 acc:(0.992)

45. kind=声かけ yohbi=火 time=午前0時 15 ==> higaisha=女性 15 acc:(0.992)

46. kind=声かけ time=午後5時 age=若い感じ 15 ==> higaisha=子ども 15 acc:(0.992)

47. kind=声かけ time=午後1時 vehicle=徒歩 15 ==> higaisha=子ども 15 acc:(0.992)

48. kind=声かけ time=午前1時 age=２０歳代 15 ==> higaisha=女性 15 acc:(0.992)

49. kind=声かけ age=４０歳代 vehicle=自動車 15 ==> higaisha=子ども 15 acc:(0.992)

50. yohbi=日 time=午後7時 vehicle=オートバイ 15 ==> kind=ひったくり 15 acc:(0.992)

51. kind=声かけ time=午後4時 vehicle=自動車 14 ==> higaisha=子ども 14 acc:(0.9916)

52. time=午後3時 vehicle=自動車 13 ==> higaisha=子ども 13 acc:(0.99111)

53. kind=声かけ yohbi=木 time=午前0時 13 ==> higaisha=女性 13 acc:(0.99111)

54. kind=声かけ yohbi=水 time=午後1時 13 ==> higaisha=子ども 13 acc:(0.99111)

55. kind=声かけ time=午前1時 43 ==> higaisha=女性 42 acc:(0.9909)

56. time=午後3時 age=６０歳代 12 ==> kind=声かけ higaisha=子ども 12 acc:(0.99049)

57. time=午後0時 age=若い感じ 12 ==> kind=ひったくり 12 acc:(0.99049)

58. kind=声かけ yohbi=火 age=１０歳代 12 ==> higaisha=子ども 12 acc:(0.99049)

59. kind=声かけ yohbi=木 time=午後3時 41 ==> higaisha=子ども 40 acc:(0.99035)

60. kind=声かけ time=午前3時 11 ==> higaisha=女性 11 acc:(0.9897)

61. higaisha=女性 time=午前3時 11 ==> kind=声かけ 11 acc:(0.9897)

62. kind=声かけ yohbi=土 time=午後0時 11 ==> higaisha=子ども 11 acc:(0.9897)

63. kind=声かけ time=午後3時 vehicle=徒歩 80 ==> higaisha=子ども 78 acc:(0.98948)

64. kind=声かけ time=午後1時 38 ==> higaisha=子ども 37 acc:(0.98933)

65. higaisha=女性 time=午後8時 38 ==> kind=声かけ 37 acc:(0.98933)

66. kind=声かけ yohbi=火 time=午後4時 37 ==> higaisha=子ども 36 acc:(0.98892)

67. kind=声かけ time=午前0時 age=２０歳代 37 ==> higaisha=女性 36 acc:(0.98892)

68. higaisha=女性 age=５０歳代 10 ==> kind=声かけ 10 acc:(0.98865)

69. kind=声かけ yohbi=土 time=午後11時 10 ==> higaisha=女性 10 acc:(0.98865)

70. kind=声かけ yohbi=日 time=午後5時 10 ==> higaisha=子ども 10 acc:(0.98865)

71. kind=声かけ yohbi=木 time=午後2時 10 ==> higaisha=子ども 10 acc:(0.98865)

72. kind=声かけ yohbi=木 age=５０歳代 10 ==> higaisha=子ども 10 acc:(0.98865)

73. kind=声かけ time=午後4時 vehicle=自転車 35 ==> higaisha=子ども 34
acc:(0.98798)

74. kind=声かけ time=午後3時 age=３０歳代 34 ==> higaisha=子ども 33 acc:(0.98744)

75. yohbi=月 time=午後9時 vehicle=オートバイ 34 ==> kind=ひったくり 33
acc:(0.98744)

76. kind=声かけ time=午前5時 9 ==> higaisha=女性 9 acc:(0.9872)

77. higaisha=女性 time=午後4時 9 ==> kind=声かけ 9 acc:(0.9872)

78. higaisha=女性 time=午前5時 9 ==> kind=声かけ 9 acc:(0.9872)

79. higaisha=女性 vehicle=マウンテンバイク 9 ==> kind=声かけ 9 acc:(0.9872)

80. kind=声かけ yohbi=月 time=午前0時 9 ==> higaisha=女性 9 acc:(0.9872)

81. kind=声かけ time=午後0時 32 ==> higaisha=子ども 31 acc:(0.98618)

82. kind=声かけ time=午前0時 vehicle=徒歩 32 ==> higaisha=女性 31 acc:(0.98618)

83. time=午前11時 vehicle=オートバイ 30 ==> kind=ひったくり 29 acc:(0.98463)

84. kind=声かけ yohbi=水 time=午後4時 30 ==> higaisha=子ども 29 acc:(0.98463)

85. higaisha=タクシー 7 ==> kind=強盗 7 acc:(0.98206)

86. higaisha=女性 time=午後2時 7 ==> kind=声かけ 7 acc:(0.98206)

87. time=午後5時 age=60歳代 7 ==> higaisha=子ども 7 acc:(0.98206)

88. yohbi=土 age=若い感じ vehicle=オートバイ 61 ==> kind=ひったくり 59
acc:(0.98199)

89. kind=声かけ yohbi=木 time=午後5時 27 ==> higaisha=子ども 26 acc:(0.98156)

90. time=午後5時 age=50歳代 26 ==> higaisha=子ども 25 acc:(0.98028)

91. time=午前3時 vehicle=オートバイ 26 ==> kind=ひったくり 25 acc:(0.98028)

92. kind=声かけ age=70歳代 6 ==> higaisha=子ども 6 acc:(0.97724)

93. higaisha=子ども age=70歳代 6 ==> kind=声かけ 6 acc:(0.97724)

94. higaisha=女性 age=10歳代 6 ==> kind=声かけ 6 acc:(0.97724)

95. yohbi=火 age=50歳代 24 ==> higaisha=子ども 23 acc:(0.97718)

96. higaisha=女性 yohbi=月 54 ==> kind=声かけ 52 acc:(0.97672)

97. kind=声かけ yohbi=火 time=午後3時 22 ==> higaisha=子ども 21 acc:(0.97319)

98. higaisha=女性 time=午後5時 5 ==> kind=声かけ 5 acc:(0.96932)

99. higaisha=女性 time=午前9時 5 ==> kind=声かけ 5 acc:(0.96932)

100. higaisha=女性 time=午前10時 5 ==> kind=声かけ 5 acc:(0.96932)

付録C　実験に用いた印刷文書

４章の実験に用いた印刷文書７枚を次ページ以降に順に示す。

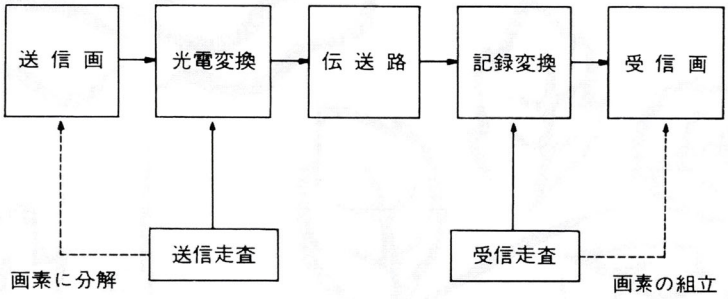
フ ァ ク シ ミ リ

ファクシミリ（Facsimile）は通信回線を用いて文書や図面を遠隔地に送る通信手段で、その歴史は古く、1843年に英国のアレキサンダ・ベインによって発明されました。モールスが電信を発明したのが1837年ですから、わずか6年後のことです。

ファクシミリの語源はラテン語の fac simile で、同じようにつくるの意味です。

ファクシミリではこれを電気的手段によって実現し、その基本構成は図のようになっております。送信画は走査によって画素に分解され、電気信号に変換されて伝送路に送られます。受信側では、送られてきた電気信号を記録に適した信号に変換し、走査しながら画素を組み立て、受信画を得ます。

わが国では1928年に、丹羽保次郎博士らによって写真電送機が初めて完成され、天皇即位式の写真が東京－大阪間で送られました。その後、電子技術の発達にともない光電変換や受信記録の平面走査が可能となり、ファクシミリに適した記録媒体が開発されるなど、使い易く安定したものとなってきました。特に、ファクシミリ信号処理に関する研究が進められ、帯域圧縮、符号化による高速化などのめざましい発達があります。



ファクシミリの基本構成

また、公衆電気通信法改正を契機に、電話網を用いてファクシミリ通信ができるようになり、一般にも広く利用されるようになりました。

ファクシミリは他の通信手段と比較して次のような特長があります。

（1） 漢字まじりの文章、図、表などが送信原稿のとうり再現されます。
（2） 装置の取扱いが容易で、特別な訓練を必要としません。
（3） 通信内容が正確に伝わり、記録として残ります。

このテストチャートは1次元ランレングス符号化方式を用いたファクシミリで、伝送速度4800b/s、1ライン当り最小処理時間20msec、副走査3.85本/mmの場合、電送時間は約1分になります。

FACSIMILE

ファクシミリテストチャート No.4　画像電子学会　©　1980

FACSIMILE TEST CHART No.4　THE INSTITUTE OF IMAGE ELECTRONICS ENGINEERS OF JAPAN

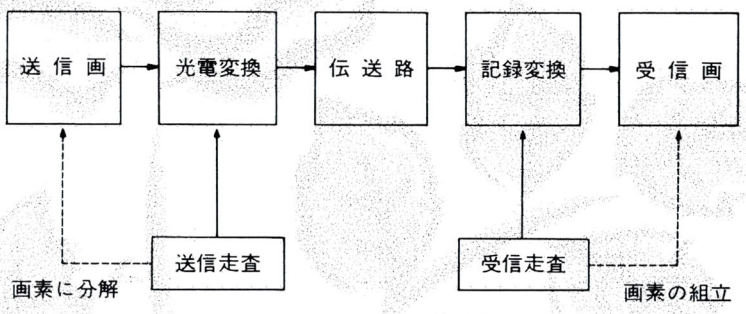
フ ァ ク シ ミ リ

ファクシミリ（Facsimile）は通信回線を用いて文書や図面を遠隔地に送る通信手段で、その歴史は古く、1843年に英国のアレキサンダ・ベインによって発明されました。モールスが電信を発明したのが1837年ですから、わずか6年後のことです。

ファクシミリの語源はラテン語の fac simile で、同じようにつくるの意味です。

ファクシミリではこれを電気的手段によって実現し、その基本構成は図のようになっております。送信画は走査によって画素に分解され、電気信号に変換されて伝送路に送られます。受信側では、送られてきた電気信号を記録に適した信号に変換し、走査しながら画素を組み立て、受信画を得ます。

わが国では1928年に、丹羽保次郎博士らによって写真電送機が初めて完成され、天皇即位式の写真が東京－大阪間で送られました。その後、電子技術の発達にともない光電変換や受信記録の平面走査が可能となり、ファクシミリに適した記録媒体が開発されるなど、使い易く安定したものとなってきました。特に、ファクシミリ信号処理に関する研究が進められ、帯域圧縮、符号化による高速化などのめざましい発達があります。



ファクシミリの基本構成

また、公衆電気通信法改正を契機に、電話網を用いてファクシミリ通信ができるようになり、一般にも広く利用されるようになりました。

ファクシミリは他の通信手段と比較して次のような特長があります。

（1） 漢字まじりの文章、図、表などが送信原稿のとうり再現されます。
（2） 装置の取扱いが容易で、特別な訓練を必要としません。
（3） 通信内容が正確に伝わり、記録として残ります。

このテストチャートは1次元ランレングス符号化方式を用いたファクシミリで、伝送速度4800b/s、1ライン当り最小処理時間20msec、副走査3.85本/mmの場合、電送時間は約1分になります。

FACSIMILE

ファクシミリテストチャート No.4　画像電子学会 © 1980

FACSIMILE TEST CHART No.4　THE INSTITUTE OF IMAGE ELECTRONICS ENGINEERS OF JAPAN

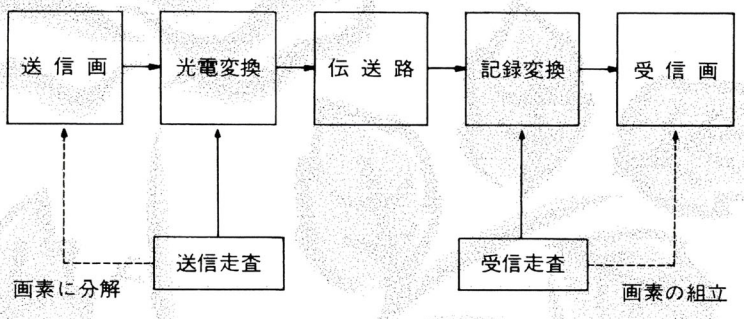
フ ァ ク シ ミ リ

ファクシミリ（Facsimile）は通信回線を用いて文書や図面を遠隔地に送る通信手段で、その歴史は古く、1843年に英国のアレキサンダ・ベインによって発明されました。モールスが電信を発明したのが1837年ですから、わずか6年後のことです。

ファクシミリの語源はラテン語の fac simile で、同じようにつくるの意味です。

ファクシミリではこれを電気的手段によって実現し、その基本構成は図のようになっております。送信画は走査によって画素に分解され、電気信号に変換されて伝送路に送られます。受信側では、送られてきた電気信号を記録に適した信号に変換し、走査しながら画素を組み立て、受信画を得ます。

わが国では1928年に、丹羽保次郎博士らによって写真電送機が初めて完成され、天皇即位式の写真が東京－大阪間で送られました。その後、電子技術の発達にともない光電変換や受信記録の平面走査が可能となり、ファクシミリに適した記録媒体が開発されるなど、使い易く安定したものとなってきました。特に、ファクシミリ信号処理に関する研究が進められ、帯域圧縮、符号化による高速化などのめざましい発達があります。



ファクシミリの基本構成

また、公衆電気通信法改正を契機に、電話網を用いてファクシミリ通信ができるようになり、一般にも広く利用されるようになりました。

ファクシミリは他の通信手段と比較して次のような特長があります。

（1） 漢字まじりの文章、図、表などが送信原稿のとうり再現されます。
（2） 装置の取扱いが容易で、特別な訓練を必要としません。
（3） 通信内容が正確に伝わり、記録として残ります。

このテストチャートは1次元ランレングス符号化方式を用いたファクシミリで、伝送速度4800b/s、1ライン当り最小処理時間20msec、副走査3.85本/mmの場合、電送時間は約1分になります。

FACSIMILE

ファクシミリテストチャート No.4　画像電子学会 © 1980

FACSIMILE TEST CHART No.4　THE INSTITUTE OF IMAGE ELECTRONICS ENGINEERS OF JAPAN

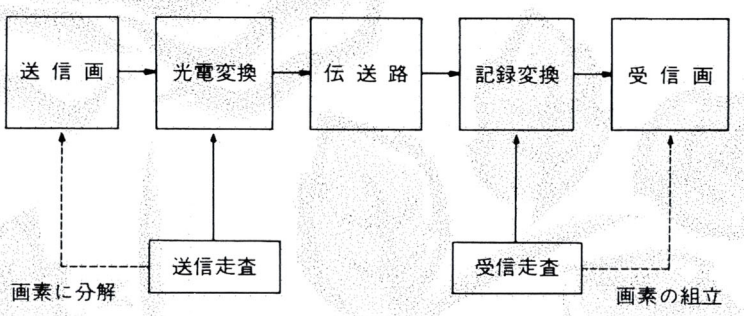
フ ァ ク シ ミ リ

ファクシミリ（Facsimile）は通信回線を用いて文書や図面を遠隔地に送る通信手段で、その歴史は古く、1843年に英国のアレキサンダ・ベインによって発明されました。モールスが電信を発明したのが1837年ですから、わずか6年後のことです。

ファクシミリの語源はラテン語の fac simile で、同じようにつくるの意味です。

ファクシミリではこれを電気的手段によって実現し、その基本構成は図のようになっております。送信画は走査によって画素に分解され、電気信号に変換されて伝送路に送られます。受信側では、送られてきた電気信号を記録に適した信号に変換し、走査しながら画素を組み立て、受信画を得ます。

わが国では1928年に、丹羽保次郎博士らによって写真電送機が初めて完成され、天皇即位式の写真が東京－大阪間で送られました。その後、電子技術の発達にともない光電変換や受信記録の平面走査が可能となり、ファクシミリに適した記録媒体が開発されるなど、使い易く安定したものとなってきました。特に、ファクシミリ信号処理に関する研究が進められ、帯域圧縮、符号化による高速化などのめざましい発達があります。



ファクシミリの基本構成

また、公衆電気通信法改正を契機に、電話網を用いてファクシミリ通信ができるようになり、一般にも広く利用されるようになりました。

ファクシミリは他の通信手段と比較して次のような特長があります。

（1） 漢字まじりの文章、図、表などが送信原稿のとうり再現されます。

（2） 装置の取扱いが容易で、特別な訓練を必要としません。

（3） 通信内容が正確に伝わり、記録として残ります。

このテストチャートは1次元ランレングス符号化方式を用いたファクシミリで、伝送速度4800b/s、1ライン当り最小処理時間20msec、副走査3.85本/mmの場合、電送時間は約1分になります。

FACSIMILE

ファクシミリテストチャート No.4　画像電子学会　©　1980

FACSIMILE TEST CHART No.4　THE INSTITUTE OF IMAGE ELECTRONICS ENGINEERS OF JAPAN

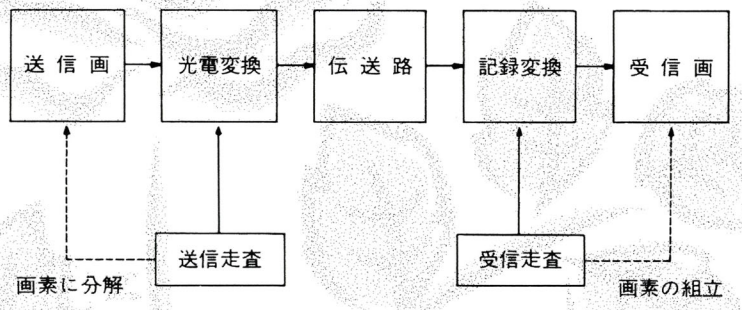
フ ァ ク シ ミ リ

ファクシミリ（Facsimile）は通信回線を用いて文書や図面を遠隔地に送る通信手段で、その歴史は古く、1843年に英国のアレキサンダ・ベインによって発明されました。モールスが電信を発明したのが1837年ですから、わずか6年後のことです。

ファクシミリの語源はラテン語の fac simile で、同じようにつくるの意味です。

ファクシミリではこれを電気的手段によって実現し、その基本構成は図のようになっております。送信画は走査によって画素に分解され、電気信号に変換されて伝送路に送られます。受信側では、送られてきた電気信号を記録に適した信号に変換し、走査しながら画素を組み立て、受信画を得ます。

わが国では1928年に、丹羽保次郎博士らによって写真電送機が初めて完成され、天皇即位式の写真が東京－大阪間で送られました。その後、電子技術の発達にともない光電変換や受信記録の平面走査が可能となり、ファクシミリに適した記録媒体が開発されるなど、使い易く安定したものとなってきました。特に、ファクシミリ信号処理に関する研究が進められ、帯域圧縮、符号化による高速化などのめざましい発達があります。



ファクシミリの基本構成

また、公衆電気通信法改正を契機に、電話網を用いてファクシミリ通信ができるようになり、一般にも広く利用されるようになりました。

ファクシミリは他の通信手段と比較して次のような特長があります。

（1） 漢字まじりの文章、図、表などが送信原稿のとうり再現されます。
（2） 装置の取扱いが容易で、特別な訓練を必要としません。
（3） 通信内容が正確に伝わり、記録として残ります。

このテストチャートは1次元ランレングス符号化方式を用いたファクシミリで、伝送速度4800b/s、1ライン当り最小処理時間20msec、副走査3.85本/mmの場合、電送時間は約1分になります。

FACSIMILE

ファクシミリテストチャート No.4　画像電子学会 © 1980

FACSIMILE TEST CHART No.4　THE INSTITUTE OF IMAGE ELECTRONICS ENGINEERS OF JAPAN

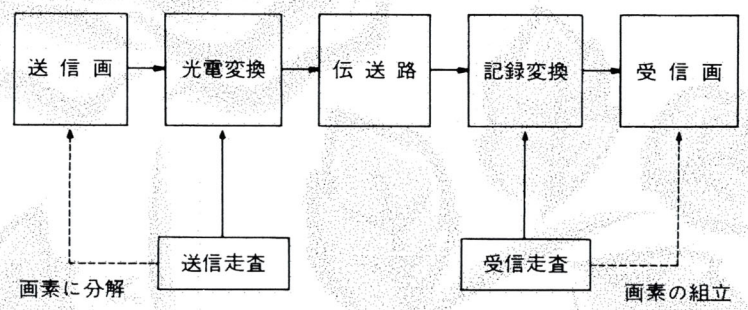
フ ァ ク シ ミ リ

ファクシミリ（Facsimile）は通信回線を用いて文書や図面を遠隔地に送る通信手段で、その歴史は古く、1843年に英国のアレキサンダ・ベインによって発明されました。モールスが電信を発明したのが1837年ですから、わずか6年後のことです。

ファクシミリの語源はラテン語の fac simile で、同じようにつくるの意味です。

ファクシミリではこれを電気的手段によって実現し、その基本構成は図のようになっております。送信画は走査によって画素に分解され、電気信号に変換されて伝送路に送られます。受信側では、送られてきた電気信号を記録に適した信号に変換し、走査しながら画素を組み立て、受信画を得ます。

わが国では1928年に、丹羽保次郎博士らによって写真電送機が初めて完成され、天皇即位式の写真が東京－大阪間で送られました。その後、電子技術の発達にともない光電変換や受信記録の平面走査が可能となり、ファクシミリに適した記録媒体が開発されるなど、使い易く安定したものとなってきました。特に、ファクシミリ信号処理に関する研究が進められ、帯域圧縮、符号化による高速化などのめざましい発達があります。



ファクシミリの基本構成

また、公衆電気通信法改正を契機に、電話網を用いてファクシミリ通信ができるようになり、一般にも広く利用されるようになりました。

ファクシミリは他の通信手段と比較して次のような特長があります。

（1） 漢字まじりの文章、図、表などが送信原稿のとうり再現されます。

（2） 装置の取扱いが容易で、特別な訓練を必要としません。

（3） 通信内容が正確に伝わり、記録として残ります。

このテストチャートは1次元ランレングス符号化方式を用いたファクシミリで、伝送速度4800b/s、1ライン当り最小処理時間20msec、副走査3.85本/mmの場合、電送時間は約1分になります。

FACSIMILE

ファクシミリテストチャート No.4　画像電子学会 © 1980

FACSIMILE TEST CHART No.4　THE INSTITUTE OF IMAGE ELECTRONICS ENGINEERS OF JAPAN

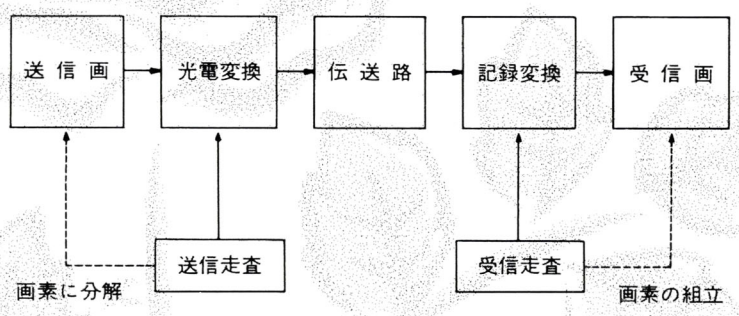
フ ァ ク シ ミ リ

ファクシミリ（Facsimile）は通信回線を用いて文書や図面を遠隔地に送る通信手段で、その歴史は古く、1843年に英国のアレキサンダ・ベインによって発明されました。モールスが電信を発明したのが1837年ですから、わずか6年後のことです。

ファクシミリの語源はラテン語の fac simile で、同じようにつくるの意味です。

ファクシミリではこれを電気的手段によって実現し、その基本構成は図のようになっております。送信画は走査によって画素に分解され、電気信号に変換されて伝送路に送られます。受信側では、送られてきた電気信号を記録に適した信号に変換し、走査しながら画素を組み立て、受信画を得ます。

わが国では1928年に、丹羽保次郎博士らによって写真電送機が初めて完成され、天皇即位式の写真が東京－大阪間で送られました。その後、電子技術の発達にともない光電変換や受信記録の平面走査が可能となり、ファクシミリに適した記録媒体が開発されるなど、使い易く安定したものとなってきました。特に、ファクシミリ信号処理に関する研究が進められ、帯域圧縮、符号化による高速化などのめざましい発達があります。



ファクシミリの基本構成

また、公衆電気通信法改正を契機に、電話網を用いてファクシミリ通信ができるようになり、一般にも広く利用されるようになりました。

ファクシミリは他の通信手段と比較して次のような特長があります。

（1） 漢字まじりの文章、図、表などが送信原稿のとうり再現されます。
（2） 装置の取扱いが容易で、特別な訓練を必要としません。
（3） 通信内容が正確に伝わり、記録として残ります。

このテストチャートは1次元ランレングス符号化方式を用いたファクシミリで、伝送速度4800b/s、1ライン当り最小処理時間20msec、副走査3.85本/mmの場合、電送時間は約1分になります。

FACSIMILE

ファクシミリテストチャート No.4　画像電子学会 ©　1980

FACSIMILE TEST CHART No.4　THE INSTITUTE OF IMAGE ELECTRONICS ENGINEERS OF JAPAN

著者学術研究論文等研究業績一覧

関連論文

・Junichi MATSUDA, Ryuhei YAMAZAKI, Yoshiki MIZUKAMI and Katsumi TADAMURA: A Method for Analyzing Crime Data from the Viewpoint of Temporal and Spatial Perspectives, *ITC-CSCC2010*, pp. 162-165 (2010)

・松田純一, 三輪智也, 松田憲, 水上嘉樹, 多田村克己：認証情報を埋め込んだ印刷帳票用地紋の生成方法, 画像電子学会誌, Vol. 38, No. 5, pp. 599-607 (2009)

・Junichi MATSUDA, Tomoya MIWA, Ken MATSUDA, Yoshiki MIZUKAMI and Katsumi TADAMURA: A Method for Embedding Information into Printed Documents using Dot Pattern Watermarking, *ITC-CSCC2009*, pp. 1487-1490 (2009)

・Junichi MATSUDA and Hiroyuki KUMAI: Transfer-Based Japanese-Chinese Translation Implemented on an E-Mail System, *Machine Translation Summit VII*, pp. 476-480 (1999)

参考論文

・Haru ANDO, Sachiko HORI and Junichi MATSUDA: Animated Chatting -Universal access by Converting Text Information into Animation, Symbols, and Background Pictures, *HCI International 2003*, pp. 507-511 (2003)

・Yoshinori MUSHA, Atsushi HIROIKE, Yasutsugu MORIMOTO and Junichi MATSUDA: Image Rating System for Filtering Web Pages with Inappropriate Contents, *MVA2002*, pp. 518-521 (2002)

・Masaru TAKEUCHI, Haru ANDO, Hirohiko SAGAWA, Atsuko KOIZUMI, Junichi MATSUDA and Hiromichi FUJISAWA: Sign Language Translation Technology and Its Applications, *eBusiness and eWork conference and exhibition*, Session 9F (2000)

・大淵康成, 北原義典, 小泉敦子, 松田純一, 畑岡信夫：マイコン向け音声認識技術を用いた携帯型音声通訳機, 電子情報通信学会論文誌, Vol. J83-D-II, No. 11, pp. 2309-2317 (2000)

査読なし論文

・山崎竜平，松田純一，水上嘉樹，多田村克己：テキストからの情報抽出及びその可視化表現手法の開発，画像電子学会２４６回研究会予稿集，pp. 69-74 (2009)

・三輪智也，松田純一，松田憲，水上嘉樹，多田村克己：印刷帳票用地紋への情報埋め込み手法の開発，画像電子学会２３７回研究会予稿集，pp. 171-178 (2008)

・松田純一：自治体システム最適化の実現に向けた日立のご提案，第５回都道府県ＣＩＯフォーラム (2007)

・松田純一：複数市町村等共同アウトソーシングシステムのご紹介－財務会計・福祉業務（介護保険）システム，財団法人地方自治情報センター共同アウトソーシング推進セミナー（平成19年度）愛知県開催 (2007)

・松田純一：複数市町村等共同アウトソーシングシステムのご紹介－福祉業務・財務会計システム，財団法人地方自治情報センター共同アウトソーシング推進セミナー（平成19年度）福岡県開催 (2007)

・高橋英孝，佐川浩彦，松田純一，田中英之，柏倉賢一，小野口敦，中館俊夫：上部消化管Ｘ線検査における聴覚障害者向け情報提供システムの評価，2003 年人間ドック学会，2-5-22 (2003)

・佐川浩彦，松田純一，楠貴晴，田中英之，高橋英孝：胃部レントゲン検査における高齢者・聴覚障害者向け情報提供システムの開発，情報処理学会第６５回全国大会講演論文集 (5)，pp. 279-282 (2003)

・柴田親男，松田純一，小泉敦子，森本康嗣：企業における非定形文書の活用促進事例－営業日報へのテキスト分析技術の適用－，情報処理学会誌，Vol. 44, No. 10, pp. 1022-1027 (2003)

・松田純一：文書管理システムを利用した情報活用促進事例，NEDO ワークショップ「情報マネージメント技術の戦略的活用Ⅱ」(2003)

・安藤ハル，堀佐知子，松田純一：対話型アニメーションメールシステムの開発，電子情報通信学会総合大会講演論文集 2003 年_情報・システム (1)，p51 (2003)

・伊藤泰樹，冨永雅介，松田純一：ナレッジの循環を支える知識管理ソリューション (特集 進化する企業を支えるＥビジネスミドルウェア)，日立評論 2002 年 9 月号，pp. 595-598 (2002)

・武者義則，広池敦，森本康嗣，松田純一：WWW 有害情報のフィルタリングのための画像判別手法，情報科学技術フォーラム講演論文集 (FIT2002)， I-82 (2002)

・安藤ハル，松田純一：VoiceXML を用いた構造化文型音声入力ガイダンス方式の開発，電子情報通信学会総合大会講演論文集 2002 年_基礎・境界，p. 311 (2002)

・竹内勝，小泉敦子，松田純一，佐川浩彦：実世界における手話認識技術，人工知能学会誌，Vol. 17, No2, pp. 138-141 (2002)

・眞利裕之，伊藤泰樹，松田純一：情報公開を見据えた文書情報マネジメントシステム (特集 21 世紀の豊かな社会を支える電子行政ソリューション)，日立評論 2000 年 9 月号，pp. 49-52 (2000)

・松田純一，本城信輔：情報通信不適正利用対策技術の研究開発，通信・放送機構研究発表会 (2000)

・大淵康成，小泉敦子，松田純一，北原義典：マイコン向け音声認識技術を用いた携帯型音声通訳機 (音声処理技術のデモの紹介)，情報処理学会研究報告. SLP， 音声言語情報処理 2000 (54)， p. 87 (2000)

・大淵康成， 小泉敦子， 北原義典， 松田純一， 塚田俊久， 北爪吉明， 田中誠， 内館秀樹：音声による単語入力機能を持つ携帯型通訳機の開発，電子情報通信学会総合大会講演論文集 1999 年_情報・システム (1)， p. 248 (1999)

・河野勝也， 松田純一， 隈井裕之：情報化社会における多言語解析とインターネット：日中，日韓における翻訳メールシステム，情報処理学会研究報告. DD，デジタル・ドキュメント 99 (10)， pp. 9-15 (1999)

・Kang Yong Hee, Kouichi Tanaka and Junichi Matsuda: The Korean Analysis System by The Using of The Korean/Japanese Machine Translation's Dictionary, Proceedings of the 11th Korean Language Computing Conference and the 1st Workshop for Morphological Analyzer and Tagger Evaluation Contest, Korea Information Science Society, pp. 106-116 (1999) (in Korean)

・河野勝也，松田純一，隈井裕之：日中・日韓双方向間の機械翻訳メールシステム，情報処理学会第５７回全国大会講演論文集 (2)，pp. 265-266 (1998)

・河野勝也，　隈井裕之，松田純一：ピン音表記を用いた文法解析型複数文節中国語文章入力システム，人文学と情報処理，(1998)

・松田純一，河野勝也：構文ダイレクト方式による日韓機械翻訳システム，情報処理学会第４４回全国大会講演論文集(3)，pp. 139-140(1992)

・Junichi Matsuda and Teiichi Kashiwagi，Electronic Dictionaries for Machine Translation using Interlingua，BPPT Natural Language Processing Seminar(1990)

・松田純一，寺田和人，村上孝也：ＣＩＣＣインドネシア語翻訳システムの構成，情報処理学会第４０回全国大会講演論文集(1)，pp. 428-429(1990)

・松田純一：対訳関係を用いた語義の分類方法，人工知能学会第3回全国大会(1989)

・Teiichi Kashiwagi and Junichi Matsuda: CICC Machine Translation Project，ガジャマダ大学４０周年記念セミナー(1989)

・松田純一，梶博行：英文生成における修飾語句の語順決定方式，情報処理学会第36回全国大会講演論文集，pp. 1233-1234(1988)

・松田純一，梶博行，臼井孝雄：機械翻訳における文法評価，情報処理学会第34回全国大会講演論文集，pp. 1275-1276(1987)