# 企業情報システムにおける
# データの抽出の効率化に関する研究

2012 年 1 月

松 本 俊 子

# 企業情報システムにおける
# データの抽出の効率化に関する研究

提出先　　大阪大学大学院情報科学研究科

提出年月　2012 年 1 月

松 本 俊 子

# 研究業績

## A. 学術論文誌論文

1. T. Matsumoto, W. Yukawa, Y. Nozaki, R. Nakashige, M. Shinya, S. Makino, M. Yagura, T. Ikuta, T. Imanishi, H. Inoko, G. Tamiya, and T. Gojobori: Novel Algorithm for Automated Genotyping of Microsatellites, *Nucleic Acids Research*, Vol. 32, No. 20, pp.6069-6077, 2004.

2. G. Tamiya, M. Shinya, T. Imanishi, T. Ikuta, S. Makino, K. Okamoto, K. Furugaki, T. Matsumoto, S. Mano, S. Ando, Y. Nozaki, W. Yukawa, R. Nakashige, D. Yamaguchi, H. Ishibashi, M. Yonekura, Y. Nakami, S. Takayama, T. Endo, T. Saruwatari, M. Yagura, Y. Yoshikawa, K. Fujimoto, A. Oka, S. Chiku, S. E. V. Linsen, M. J. Giphart, J. K. Kulski, T. Fukazawa, H. Hashimoto, M. Kimura, Y. Hoshina, Y. Suzuki, T. Hotta, J. Mochida, T. Minezaki, K. Komai, S. Shiozawa, A. Taniguchi, H. Yamanaka, N. Kamatani, T. Gojobori, S. Bahram, and H. Inoko: Whole Genome Association Study of Rheumatoid Arthritis using 27,039 Microsatellites, *Human Molecular Genetics*, Vol. 14, No. 16, pp.2305-2321, 2005.

3. 松本俊子, 大峡光晴, 小野山隆, 秋吉政徳：ビジネス文書からのメタデータ抽出のためのルール自動生成技術, 電気学会 C 部門論文誌, Vol.131, No.8, pp.1502-1511, 2011.

4. 松本俊子, 小野山隆, 秋吉政徳：業務情報周知のための業務遂行状況に応じた情報提示要否の判別方式, 電気学会 C 部門論文誌, Vol.131, No.10, pp.1819-1827, 2011.

## B. 国際会議

1. T. Matsumoto, K. Sadakane, H. Imai, and T. Okazaki: Can General-Purpose Compression Schemes Really Compress DNA Sequences?, in *Proc. of The Fourth*

*Annual International Conference on Computational Molecular Biology*, poster-40, pp.76-77, 2000.

2. T. Matsumoto, K. Sadakane, and H. Imai: Biological Sequence Compression Algorithms, in *Proc. of The Eleventh Workshop on Genome Informatics*, pp.43-52, 2000.

3. T. Matsumoto, Y. Nozaki, R. Nakashige, M. Shinya, Y. Yoshikawa, S. Mano, T. Imanishi, H. Inoko, G. Tamiya, and T. Gojobori: Investigating the Optimal Dataset Size for SNP Hunting, in *Proc. of The 15th Workshop on Genome Informatics*, P105, 2004 (in CD-ROM).

4. T. Matsumoto and R. Nakashige: Evaluating Robustness of Algorithm for Microsatellite Marker Genotyping, in *Proc. of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp.158-164, 2005.

5. T. Matsumoto, R. Nakashige, T. Watanabe, and Y. Sugimoto: Applying Genotyping Algorithm for Microsatellite Markers to Bovine Data, *in Proc. of the 20th IUBMB International Congress of Biochemistry and Molecular Biology*, 2P-B-036, 2006 (in CD-ROM).

6. T. Matsumoto, R. Nakashige, and S. B. Lee: Expert System for Evaluating Automated Allele Call, in *Proc. of 17th International Symposium on Human Identification*, 43, 2006 (in CD-ROM).

7. T. Matsumoto, Y. Nozaki, and R. Nakashige: SNP Data Consulting Program, in *Proc. of 2006 IEEE Region 10 Conference*, BI1.1, 2006 (in CD-ROM).

8. M. Oba, Y. Nozaki, T. Matsumoto, and T. Onoyama: Underline Removal Method by utilizing Characteristics of Japanese Business Documents, in *Proc. of 2009 IEEE Region 10 Conference*, Tue 4.2.1, 2009 (in CD-ROM).

9. T. Matsumoto, M. Oba, and T. Onoyama: Sample-based Collection and Adjustment Algorithm for Metadata Extraction Parameter of Flexible Format Document, in *Proc. of The 10th International Conference on Artificial Intelligence and Soft Computing*, pp.566-573, 2010.

10. M. Oba, T. Matsumoto, Y. Iwata and T. Onoyama: Generating Hierarchical Virtual Directory by Metadata Frequency Difference, in *Proc. of The fifth Workshop on Human-Computer Interaction and Information Retrieval*, 7, 2011 (in CD-ROM).

## C. 学会講演

1. 松本俊子, 野崎康行, 中重亮, 間野修平, 知久季倫, 今西規, 五條堀孝, 猪子英俊, 田宮元： Clark アルゴリズムによるハプロタイプ解析, 第 26 回日本分子生物学会年会, p.1045, 2003.

2. 松本俊子, 湯川航, 野崎康行, 中重亮, 新屋みのり, 生田智樹, 今西規, 猪子英俊, 田宮元, 五條堀孝： マイクロサテライトマーカーの Stutter ピーク予測を用いた pooled typing 補正, 第 27 回日本分子生物学会年会, p.1021, 2004.

3. 松本俊子, 野崎康行, 中重亮： 連鎖不平衡解析のための SNP データ確認プログラム, 第 28 回日本分子生物学会年会, p.526, 2005.

4. 松本俊子, 大峡光晴, 小野山隆, 薦田憲久：営業文書からのメタデータ自動抽出のためのパラメータ自動生成技術, 電気学会 情報システム研究会, IS-10-046, pp.129-134, 2010.

5. 松本俊子, 大峡光晴, 小野山隆, 薦田憲久：メタデータ抽出用パラメータの自動生成による導入容易な業務文書管理活用支援システム, 平成 22 年度 電気学会 C 部門大会, TC15-3 , pp.546-551, 2010.

6. 大峡光晴, 松本俊子, 岩田泰明, 小野山隆：メタデータの頻度差を利用した階層的仮想フォルダ自動生成, 信学技報, Vol. 110, No. 467, PRMU2010-266, pp.171-176, 2011.

7. 松本俊子, 小野山隆, 秋吉政徳 : 業務情報周知および活用を実現するビジネスレコメンデーション技術, 電気学会 情報システム研究会, IS-11-041, pp.47-52, 2011.

8. 松本俊子, 小野山隆, 薦田憲久 : ファイルサーバのファイルサイズ分布のモデル化に基づくファイル削除可否確認回数と削減量の上限との関係推定方式, 平成 23 年度 電気学会 C 部門大会, GS5-6, pp.1284-1287, 2011.

## D. その他

1. 中重亮, 野崎康行, 松本俊子: だれでも使えるバイオインフォマティクスリソース 連載第 7 回 実践しよう関連解析, 分子精神医学, Vol. 7, No. 3, pp.45-50, 2007.

# 内容梗概

本論文は, 筆者が 2000 年から現在まで日立ソフトウェアエンジニアリング (株) (現 (株) 日立ソリューションズ) ならびに 2010 年から現在まで大阪大学大学院マルチメディア工学専攻在学中に行ってきた，企業情報システムにおけるデータ抽出の効率化に関する研究をまとめたものである。

企業情報システムによる業務効率化の進展に伴い，効率化の対象は，状況に応じて様々な内容が記載される定型性の低いデータへと移りつつある。近年データの大規模化がますます顕著になる中，データの中から，ユーザが種々の業務をこなす中で着目するべき部分を抽出する作業の効率化に対するニーズが高まっている。この点に関し，データに内在する法則性に基づいて抽出を行うことを方針として，企業情報システムにおいて利用されるデータの主なものとして挙げられる数値データおよび文書データにおける課題を解決する手法について提案する。

数値データからの着目するべき箇所の抽出に関しては，データ量の増大に伴い自動処理の重要性が高まるとともに，データの測定原理に由来する内在的法則性を利用する高精度な処理が求められている。内在的法則性について蓄積された専門家の知見は定性的な形で表されることが多いため，法則性が典型的に現れているシンプルなデータを集めて傾向を調べることで知見を定量化し，着目するべきデータを抽出する手法を提案する。

文書データに関しては，内部統制の監査における迅速な提出のため，タイトル，顧客名などのメタデータを文書データから抽出し整理分類して管理するニーズが高まっている。しかし従来のメタデータ抽出技術は抽出にあたり着目するべきキーワードやレイアウトを抽出用ルールとしてあらかじめ設定しておくことを前提としている。そこで，ビジネス文書の記載上の傾向に基づき，サンプル文書における正解メタデータの記載から抽出用ルールを生成する手法を提案する。

さらに，コンプライアンス違反の防止のため，法令，社内規則などの多数の文書データからのユーザの業務遂行上参照が必要なものの抽出効率改善が求められている。そこで，業務上利用するアプリケーションの表示文字列の例とそれぞれの状況における業務情報の参照要否を入力として，参照要否の判別条件を構成・維持する手法を提案する。

本論文は全 5 章から構成される。

第 1 章の序論では，企業情報システムにおいて用いられるデータから業務上着目するべき箇所の抽出の効率化について，数値データおよび文書データのそれぞれについて解決するべき課題を述べ，従来研究を概観するとともに，本論文の目的と位置づけを明らかにする。

第 2 章では，数値データの具体例として，データ量の急速な増大により自動処理の必要性が高まっている DNA(Deoxyribo Nucleic Acid)データを具体例として，真のデータをノイズデータから識別する手法を提案する。入力データのうちノイズデータの判別が容易な実験結果を選んでノイズデータの傾向を計算し，その傾向を用いてノイズデータから真のデータを抽出する。さらに，174 個の DNA 多型データを用いてヒトゲノム全域における平均抽出精度を評価して提案手法の有効性を示すとともに，企業情報システムにおける他の数値データへの適用性について論じる。

第 3 章では，ビジネス文書からのメタデータ抽出用ルールを生成する手法を提案する。メタデータの記載上の特徴を洩れなく集めるため，サンプル文書における正解メタデータの記載からルールの候補を列挙する。さらに，必要性の低いルールにより誤った抽出が行われるのを防ぐため，サンプル文書の中で目的外の文字列にもルール候補があてはまらないか調べることでルール候補の絞り込みを行う。営業文書および週次作業報告書を用いてルール生成に要する時間とメタデータ抽出の再現率を評価し，提案手法の有効性を示す。

第 4 章では，業務情報の周知のための業務遂行状況に応じた提示要否の判別方式を提案する。業務上利用するアプリケーションの表示文字列における特徴に基づき，高頻度な形態素列を用いて判別を行う。提示された業務情報が不要であった場合に簡易かつ確実にフィードバックを反映し，一方で必要な業務情報の提示の再現率を維持するため，形態素列を用いて抑止キーワードを選択する。三種類の業務情報に対して提案手法を適用し，有効性を示す。

最後に第 5 章では，結論として本研究で得られた成果を要約し，今後に残された課題について述べる。

# 目次

# 第 1 章

# 序論

## 1.1 研究の背景

企業活動において，従業員が行うオフィスワークとしての各業務に対して情報システムを用いた効率化が推進される中で, 情報システムに対する要件も次第に変化してきている。この要件の変化は，情報システムによる効率化の対象がより定型性の低いデータを扱う処理へ移っていることに伴って生じていると捉えることができる。

従来，情報システムによる効率化の主な対象は定型性の高いデータであった。基幹システムにより行われる原価計算, 生産管理, 給与計算など[宮川 2004] [吉川 1990] [小谷 1997] [川口 1999]で管理されるデータである，会計，生産，勤怠管理などの数値データおよびその元となる帳票のスキャン画像においては，用紙上の座標位置・データストリーム上のオフセット･タグなどにより各データの記載位置が固定的に定められている。この特徴から，指定した印字位置から文字認識を行う技術[Taylor1992] [Cesarini1998]，帳票の種類を識別する技術[皆川 2009]，バーコード[平本 2001]などによるデータ登録の自動化技術や，電子帳票により帳票そのものを電子データのまま流通させること[日経 BP 企画 2005] [川上 2010]，さらに直接企業間のデータ授受を行う EDI(Electronic Data Interchange)[松野 2002]など，より少ない人的コストでデータを抽出するための提案が行われてきた。これらの取組により定型性の高いデータの処理については十分に効率化が進んでいると考えられる。

定型性の高いデータを対象としたシステム活用が推進されたことに伴い，近年では定型性の低いデータを扱う処理の効率化に焦点が移りつつある。業務の効率改善のためには，必要な情報を適時利用できることが重要であると言われている[西村 2008]が，近年データの大規模化傾向がますます顕著になる中，ユーザが種々の業務をこなす中で重要な部分・着目するべき箇所をデータの中から探し出す作業において，効率化が不十分であることが指摘されている[栗原 2009] [キーマンズネット 2009]。定型性の低いデータでは業務の内容や遂行状況に応じて様々な内容が記載されるため，高度なスキルを用いたデータの解釈

や判断が必要となる。そこで本研究では，定型性の低いデータの抽出に着目する。

専門知識を持つユーザの手作業によって定型性の低いデータの抽出が行われる場合，ユーザはデータに内在する何らかの性質に基づいて判断や解釈の基準を設けて抽出を行っていると考えられる。そこで，定型性の低いデータの抽出に関し，本研究ではデータに内在する性質に基づいて抽出を行うという方針を考える。企業情報システムにおいて利用されるデータの主要なものとして，数値データおよび文書データが挙げられる[川波 1998]。これらのデータからの業務上着目するべき箇所の抽出について，本研究では図 1.1 に示す課題を対象として取り上げる。

まず，数値データから着目するべき部分を抽出する作業については，ビジネスインテリジェンスにおけるドリルダウンなどのデータ絞り込み機能[SAP2008]，データ全体を把握しやすくする表示方法[水野 1993]などが提案されてきた。これらの技術は，着目するべき部分であるかどうかの判断はユーザ自身が行うことを前提としている。また，データマイニングにおいてデータの相関関係を探し出す技術が提案されているが[喜連川 1997]，相関ルールは単純に数値的関係として抽出されるため，自明な相関ルールによる影響をデータ

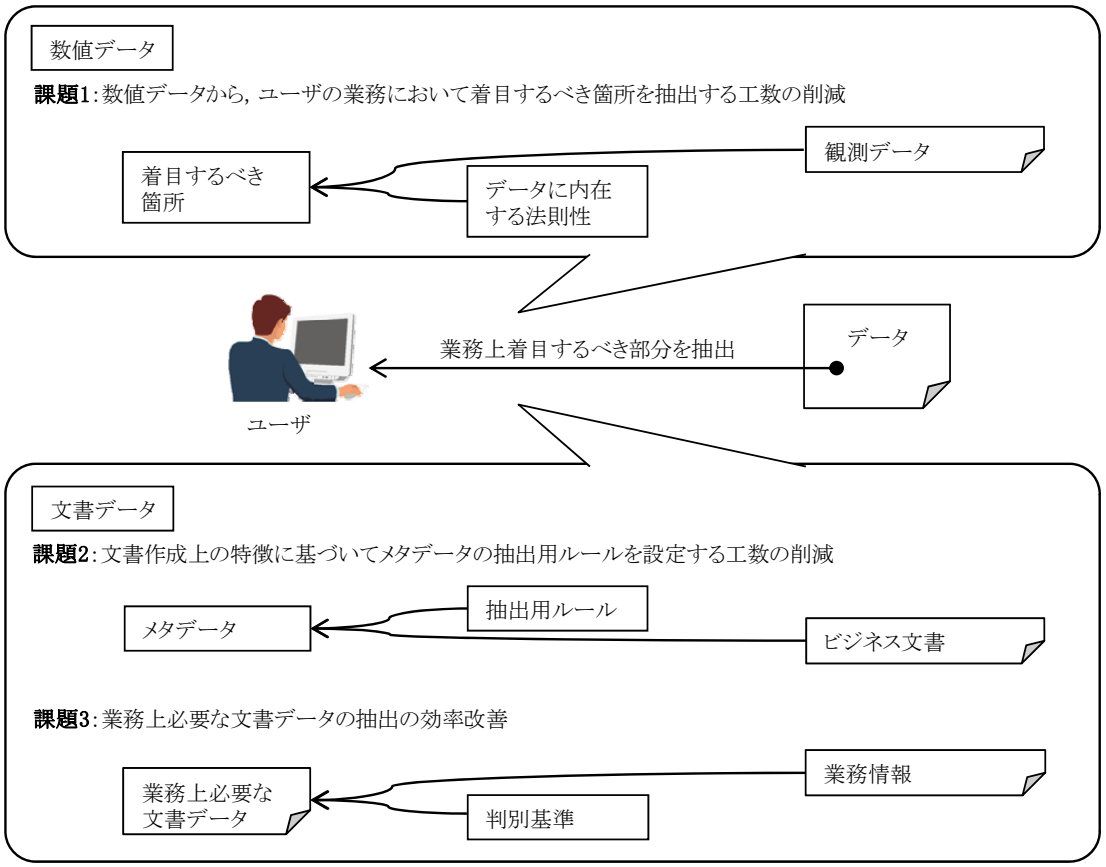



図 1.1　企業情報システムにおけるデータの抽出の効率化における課題

から除くための試行錯誤が必要である。そこで本研究では，数値データに内在する法則性に関する専門家の知見に基づいて，データからユーザの業務において着目するべき箇所を抽出する工数の削減を第一の課題として取り上げる。

次に，文書データの中から着目するべき部分を抽出する処理に対しては，デジタルペンによる記録作成支援技術[古川 2007]，全文検索技術[浅川 1992]などが提案されてきた。これらの技術においても，着目するべき部分であるかどうかの判断は，ユーザ自身が行うことを前提としている。また，閲覧効率を向上させるための文章要約技術[増山 2002]も提案されているが，企業内で用いられる文書データは，主語や述語などが整った文章ではなく名詞句が列挙されて記載される場合が多く，プレーンテキストではなくレイアウトを指定して二次元状に単語が配置される場合が多いため，想定している様式の違いから適用することができない。さらに，内部統制の監査において必要な文書を迅速に提出するための，ビジネス文書を整理分類して登録するための ECM(Enterprise Contents Management)システムへの注目の高まりから[前田 2006]，文書データからタイトル・顧客名・作成日などのメタデータを抽出する作業を自動化することによる効率改善技術が提案されている[Minagawa2006] [Handley2005] [Ishitani1999]。しかし，これらの技術は抽出にあたりメタデータ記載上の特徴として着目するべきキーワードやレイアウトをあらかじめメタデータ抽出用ルールとして定義しておくことを前提としており，組織ごとに異なる文書作成上の特徴を調べてメタデータ抽出用ルールを設定する作業がシステム導入上のボトルネックになっている。そこで本研究では，ビジネス文書の記載上の特徴に基づいてメタデータ抽出用ルールを設定する工数の削減を第二の課題として取り上げる。

さらに，多数の文書データからユーザの業務上重要なものを抽出する処理に対しては，エンタープライズサーチ[日経 BP 企画 2010]，文書クラスタリング技術[後藤 2010]などが提案されてきた。これらの技術では，ユーザが能動的に文書データの抽出を行うことを前提としている。また，イントラネット上にポータルサイトを配置することで，必要な文書データの利用を促す技術[田中 2004]や，改訂版の通知を行うことで最新版の参照を促す技術[日立ソフト 2008]などが提案されてきた。これらの技術は，参照が必要な文書データの想起を促すことができると期待できるが，ユーザがそもそも文書データの参照の必要性に気づいていない場合には効果を及ぼさない。文書データの利用では，特に法令，社内規定や通知などの文書において，必要に応じて参照することができなかった場合はコンプライアンス違反を発生させる可能性がある一方で,過剰な文書の確認は業務効率を低下させる。そこで本研究では，文書データの参照必要性の判断基準に基づいて業務上必要な文書データの抽出効率を改善する技術を対象として取り上げる。

以上述べたように，本論文では，企業情報システムにおいて，数値データからユーザの業務において着目するべき箇所を専門家の知見に基づいて抽出する技術，ビジネス文書の記載上の特徴に基づいてメタデータ抽出用ルールを生成する技術および，文書データの参照必要性の判断基準に基づいて業務上必要な文書データの抽出効率を改善する技術について述べる。

## 1.2 従来研究

### 1.2.1 数値データからの着目するべき箇所の抽出

数値データに対して真に観測するべき値をノイズから識別したり，業務上の判断のきっかけとなる箇所を抽出したりするなどの処理については，あらかじめ数理モデルを仮定しておき，統計値を算出することにより異常値を検出する手法[峰岸 2009] [高橋 2002]，自動分類を用いて推定を行う手法[村田 2004]などが提案されてきた。前者の手法は，適切な数理モデルを仮定でき，着目するべきデータにおいて有意に変動する統計値を用意できる場合に適用範囲が限られる。

後者の手法は，ユーザが正解ラベルを付与した学習データからデータの特徴を抽出し，判別器を構成するものである。学習データの件数を削減したり，ラベルを付与しないデータを用いて学習結果を強化したり，正と負の学習データの比率が均等でない場合でも安定した結果を得るための工夫が提案されてきたが[Tan2005] [Kerdprasop2011]，各データが独立であることを想定しており，近傍のデータ間に関連がある場合には対応していない。

### 1.2.2 メタデータ抽出用ルールの生成

メタデータ抽出用ルールの生成に関しては，大きく二種類の技術が提案されてきた。第一は，ニュース記事から日時・場所・被害者などを抽出したりするためのルールの生成である[Ashish1997] [Riloff1993] [Freitag2000]。これらの手法では，構文解析や HMM (Hidden Markov Model) を用いて，記載内容の順序に関する傾向を調べている。単語の羅列ではなく主語や述語などが揃った文が記載されていることや，入力データとして文字が一次元的に並んだ文字列が与えられることを前提としている。第二の既存技術は，論文の参考文献の段落から書誌情報を抽出するためのルールの生成である。論文誌によって書誌情報の記載のされ方は異なるものの，多くの文献において記載される書誌情報の種類は同じであること，単一の論文においては書誌情報の記載順序が同じであることを前提としている。また，ニュース記事についての既存技術と同じく，入力データが文字列の形で与

えられることを前提としている[薬師 2009]。これらの理由により，既存技術はいずれも，ビジネス文書からのメタデータ抽出用ルールを生成することはできない。

### 1.2.3 業務上必要な文書データの抽出

参照が必要と考えられる業務情報をユーザに提示するニーズに対しては，文書の類似性に基づく手法が提案されてきた[高野 2000]。これらの技術は，ユーザが閲覧したり登録したりしている文書と類似性の高い文書を提示するもので,「ユーザが現在利用している文書に類似した文書は, 参照の必要がある可能性が高い」との仮説に基づいている。このため，提示されるものは同じ業種の過去の成果物である可能性が高く，通達や規則などを参照する場合にはその必要性を検知しにくい。

また，ユーザが業務を遂行しているタイミングで，その業務の遂行に必要な手続きや参照するべき文書を登録させる方式も提案されている[Memmel1997] [鈴木 2006]。これは，参照の要否の判別条件をユーザ自身に指定させるものであるが，指定を行わせるタイミングを調整することでユーザの負担感の減少を図っているものである。ユーザが登録を行う時間を継続的に確保できない場合には，登録件数を充実させられなかったり古い文書がいつまでも残ってしまうことが課題となる。

## 1.3 研究の方針

図 1.2 に企業情報システムにおけるデータからの業務上着目するべき部分の抽出に対し，本論文で対象とする研究課題と方針の関係を示す。

### 1.3.1 数値データからの真のデータの抽出

観測される数値データの中に業務上着目するべき真のデータとそうでないノイズデータが混在し，互いの間に関連がある場合を対象として，真のデータを識別し着目するべき箇所として抽出する技術を提案する。データ間の関連についての法則性は専門知識として知られているが, 法則性の閾値や係数はサンプルや測定状況に依存して変動することが多い。このような数値データの例として，グローバル化と共に激化する新薬開発競争において注目が高まっているゲノム創薬において用いられるとともにヒトゲノム解読以降のバイオテクノロジー市場拡大を牽引している DNA(Deoxyribo Nucleic Acid)実験データ[西村 2003] [医薬産業政策研究所 2007] [特許庁 2007]を取り上げる。

DNA 実験データは実験機器の自動制御技術の発展に伴い大規模化が進み，観測データ

から真のデータを抽出する作業がボトルネックになっている。提案手法では，多様な実験データから着目するべき箇所を効率よく抽出するため，入力データからノイズデータの識別が容易なサンプルを抽出し，データに内在する法則性の計算を行うことを特徴とする。計算した法則性に基づいてノイズデータから真のデータを識別し，着目するべき箇所として抽出する。

### 1.3.2 ビジネス文書からのメタデータ抽出のためのルールの生成

ECM システムでは，営業文書，報告書，技術文書など様々な種類が存在するビジネス

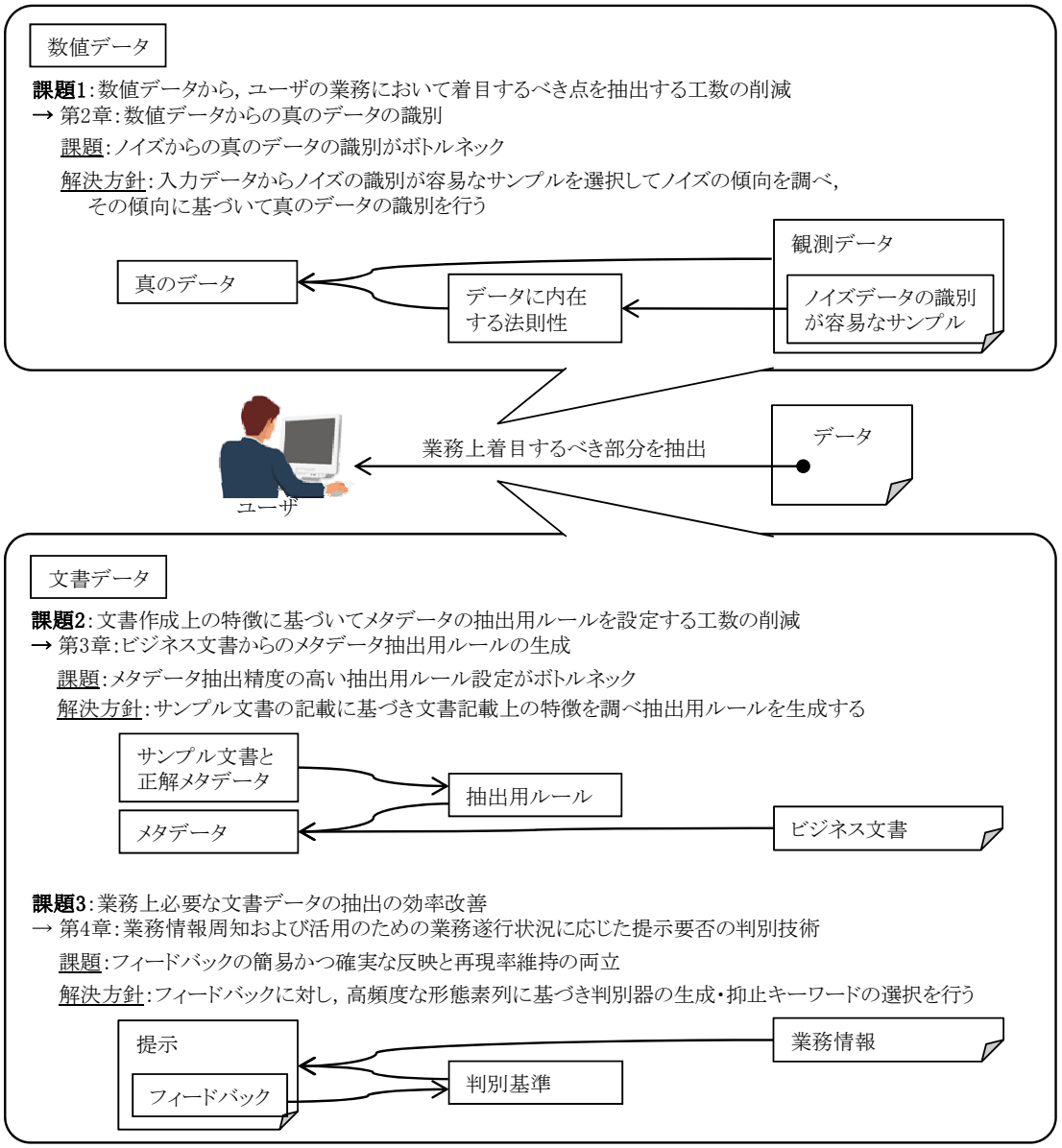



**図 1.2　研究課題と方針**

文書を統括的に管理する。さらにこれらの文書では，二次元状に名詞句が配置される特徴的な構造を持つ場合が多い。ECM システムにおけるメタデータ抽出技術として，メタデータに隣接して記載される文字列，メタデータに含まれる部分文字列および，レイアウトの指定に基づく抽出用ルールを用いるメタデータ抽出技術を対象として，サンプル文書と正解メタデータの組から成る入力をもとにメタデータ抽出用ルールを生成する方法を提案する。提案手法では，メタデータの記載上の特徴を洩れなく集めるため正解メタデータからルールの候補を列挙する。さらに，誤った抽出を起こすルールを除くため，サンプル文書中の正解メタデータ以外の文字列にあてはまるルール候補を除くことでルール候補を絞り込む。

### 1.3.3 業務遂行状況に応じた業務情報の提示要否判別

ユーザ自身による判別条件の登録に依存することなく，業務上参照が必要と考えられる通達や規則などを抽出するため，業務上利用するアプリケーションの表示文字列の例とそれぞれの状況における参照要否を入力として提示要否の判別条件を構成・修正する手法を提案する。提案手法は，表示文字列における「過度に高頻度な典型文字列を持つ」，「業務内容に特徴的な形態素の組で頻度の上昇が見られる」という特徴に基づき，高頻度な形態素列を用いて判別器を構成する。また提案手法は，提示された業務情報が不要であるとのフィードバックに対し，高頻度な形態素列を用いて抑止キーワードを選択することで，提示要否の判別基準を修正する。これにより，簡易かつ確実なフィードバックを実現すると共に，提示が不要な業務上表示文字列を正しく判別できる確率の改善と再現率低下の防止の両立を目指す。

## 1.4 本論文の構成

本論文では，2 章以降を次のように構成する。

第 2 章では，文献[Matsumoto2004] [Matsumoto2005]に基づき，数値データからノイズデータの識別が容易なサンプルを選択してノイズデータの傾向を調べ，観測データから真のデータを識別する技術を DNA 実験データを対象として提案する。そして，174 個の DNA 実験データを用いて提案方式の識別精度の評価を行うとともに，企業内の他の数値データへの適用性について議論する。

第 3 章では, 文献[松本 2010A] [Matsumoto2010] [松本 2010B] [松本 2011B]に基づき，サンプル文書における正解メタデータの記載から，メタデータ抽出のためのルール候補を

列挙し，その後，ルール候補がサンプル文書の中で目的外の文字列にもあてはまらないか調べることで選択および最適化を行う，メタデータ抽出用ルールの生成方式について提案する。さらに，営業文書および週次作業報告書に対し提案方式を適用し，人手で調節した抽出用ルールおよび自動生成した抽出用ルールの比較による評価を行う。

第 4 章では，文献[松本 2011A] [松本 2011C]に基づき，業務上利用するアプリケーションの表示文字列における特徴を活用し，高頻度な形態素列を用いて判別器の構成および抑止キーワードの選択を行う，業務情報周知のための業務遂行状況に応じた文書データ提示要否の判別方式について提案する。そして，三種類の業務情報に対して提案方式を適用し，提示要否判別の精度を評価する。

最後に，第 5 章では，結論として本研究で得られた成果を要約し，今後に残された課題について述べる。

# 第 2 章
# 数値データからの真のデータの識別技術

## 2.1 緒言

企業情報システムで用いられる数値データの増大に伴い，業務を遂行するためにデータから着目するべき部分を探す工数が問題視されつつある。業務の内容や遂行状況に応じて様々な内容が様々な書式で記載される定型性の低いデータでは，着目するべき箇所の抽出には専門知識を用いたデータ解釈が必要である。多くの数値データでは，着目するべき真のデータとその他のデータとの間に関連があるため複数のデータの関係を把握して真のデータを識別する必要があり，統計量により異常値を検出するアプローチは適用できない。専門家は経験的に関連性についての法則性を理解しデータの解釈を行っているが，目視による観察では定性的な傾向という形でしか認識できないため，そのままでは情報システムによる自動抽出に適用することができない。

本章では，グローバル化と共に激化する新薬開発競争において注目が高まっているゲノム創薬において用いられるとともにヒトゲノム解読以降のバイオテクノロジー市場拡大を牽引している DNA 実験データ[西村 2003] [医薬産業政策研究所 2007] [特許庁 2007]を対象とする。真のデータを識別することを対象とした上記の課題の解決策として，データに内在する法則性が典型的に現れているサンプルを用いてノイズデータの傾向を計算し，真のデータをノイズデータから識別する技術を提案する。

ヒトゲノムの完全解読後，遺伝子の機能解析研究が活発に行われている。そのなかでも特定の疾患の有無，薬物の効果の程度，副作用の有無などに関与する主要な遺伝要因を同定するための基盤となる遺伝子の探索研究が特に注目されている。

近年の DNA 配列読み取り分野における進歩は，実験を効率的かつ大規模に実施可能にした。これにより，疾患に影響を与える遺伝子の探索研究で一般に行われているように，数百箇所の DNA 配列における読み取り実験に基づく研究が可能となった。大規模データを効率的に分析するためには，自動処理技術が必要である。分析における主要な問題は，実験中に引き起こされる様々な種類のノイズデータからの真のデータの識別である。従来

は，詳細な目視によりノイズデータと真のデータを識別することがデータの正しい分析を行うための唯一の方法であった。しかし目視による真のデータの識別は時間がかかることから，ノイズデータから真のデータを識別するアルゴリズム[Stoughton1997] [Perlin1994]や，真のデータの自動識別結果が正しい場合にデータ間に成り立つ条件を設定しておくことで自動識別結果の確認件数を削減する技術[Palsson1999]が提案されている。しかしながら，既存の手法はいずれも，DNA 配列の種類ごとに実験条件を詳細に調整してノイズデータを低減したデータを用いることを前提としている[Stoughton1997] [Perlin1994] [Palsson1999]。大規模な遺伝子探索研究においては実験条件の詳細な調整を行うことは不可能であることから，ノイズデータの現れ方の法則性を用いて様々な種類のノイズデータから真の観測結果を精度よく抽出する技術が必要である。

本章では様々なノイズデータの傾向を自動的に見積もることで正確に分析を行うアルゴリズム AutoTyper を提案する。AutoTyper では，ノイズデータの線形性・比率の均一性という法則性を利用する。まずノイズデータを容易に識別できるシンプルなデータを持つ個体を集めることにより，線形回帰直線や比率の値を計算し，ノイズデータの現れ方を見積もる。さらに，様々な DNA 実験データの観察に基づき線形回帰直線の傾きおよび比率の値に閾値を設け，極端な値が計算されることを防ぐ。この結果に基づいて分析を行うことで，少数の個体からノイズデータの現れ方を効率よく見積もり，その他の複雑なデータを示す個体に対しても正確な推定を行うことを目指す。

本章の構成は，以下の通りである。2.2 節では DNA データの構造およびノイズデータの発生原理について述べる。2.3 節では AutoTyper の詳細について述べる。2.4 節では AutoTyper による真のデータ識別精度を評価する実験について述べ，2.5 節では考察を行う。2.6 節ではまとめを述べるとともに，企業情報システムで扱われるその他の数値データへの適用性について論じる。

## 2.2 DNA 実験の手順およびノイズデータの発生原理

図 2.1 に示すように，DNA 断片の長さにおける個体差を PCR（Polymerase Chain Reaction）および電気泳動と呼ばれる実験技術を用いて観測し，横軸に DNA の長さ，縦軸に DNA の量をとった二次元のデータ点として得る実験を対象とする。理想的には，図 2.1 の右上に示すように，父由来・母由来のデータのみ得られるべきところ，実際には図 2.1 の右下に示すように，多くのノイズデータ中に埋もれてしまい，その個体が持つ DNA の状態を正しく反映したデータとノイズデータとを識別できなくなってしまう。このよう

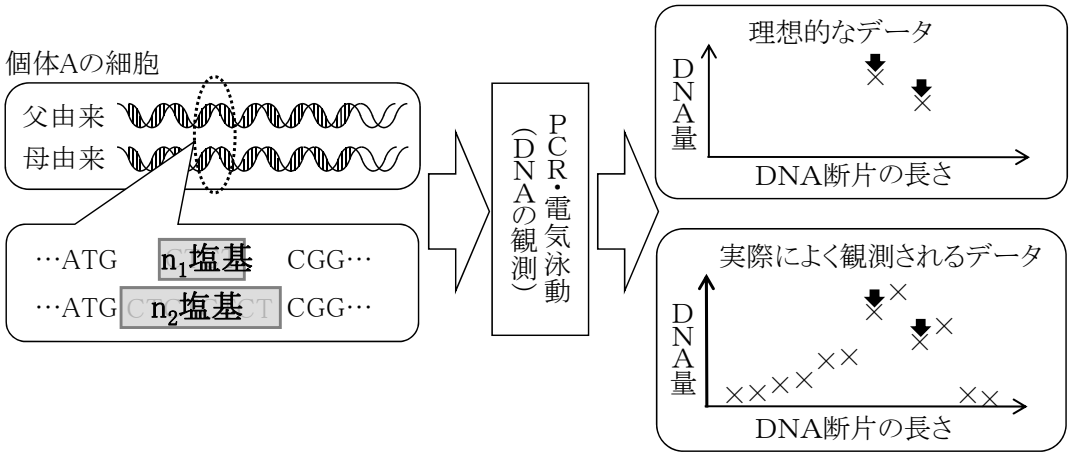



**図 2.1　DNA 実験手順および得られる結果の概要**

なデータをもたらす実験の手順およびノイズデータの発生原理について，以下に述べる。

ヒトゲノム中には，2 塩基から 6 塩基の短い配列パターンが繰り返されて現れるマイクロサテライトと呼ばれる配列パターンが存在する。繰り返し単位の塩基数は unit 長と呼ばれる。マイクロサテライトは，繰り返し回数，すなわちマイクロサテライト部分の長さが個体によって異なる場合がある。

マイクロサテライトにおける個体差の例を図 2.2 に示す。この例では，四種類の DNA 分子，A(Adenin)，T(Thymine)，G(Guanine)および，C(Cytosine)のうち A と T が繰り返されたマイクロサテライトが含まれている。個体 A のように二つの異なる繰り返し回数の配列パターンを持つ個体はヘテロ接合体，個体 B のように同一の繰り返し回数の配列パターンを二つ持つ個体はホモ接合体と呼ばれる。PCR によりゲノム中からマイクロサテライト部分だけを DNA の断片として切り出し，電気泳動により各 DNA 断片の長さごとに DNA 量を測定することで，各個体が持つ配列パターンを調べることができる。

上記したようにマイクロサテライトは個体間で繰り返し回数が異なる場合があるので，ゲノム上で他の DNA 配列と区別がしやすい部分であり，実験的にも検出が容易である

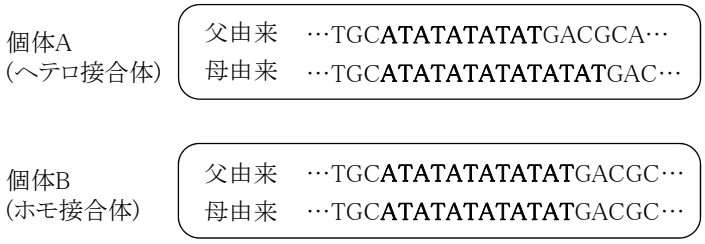



**図 2.2　マイクロサテライトのホモ接合体およびヘテロ接合体の例**

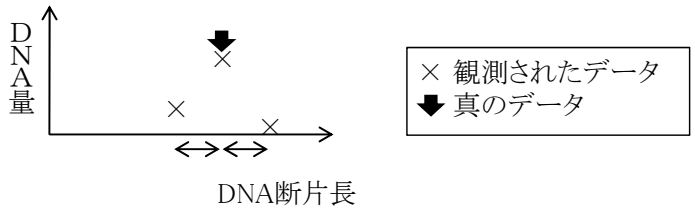
[Tautz1989][LeDuc1995]。これらの特徴により，多くの生物種でマイクロサテライトの解析が行われ，ゲノム上の位置を示すマーカーとして頻繁に利用されてきた[Dietrich1992] [Weissenbach1992] [Gyapay1994] [Dib1996] [Knapik1998] [Shimoda1999]。

PCR および電気泳動によりマイクロサテライトマーカー部分の DNA を観測する際，個体が持つ配列パターンに由来するデータ以外に現れるノイズデータとして, stutter データ[Hauge1993] [Murray1993]および，+A データ[Clark1988]が挙げられる。stutter データは，その個体が持っている配列パターンよりも繰り返し回数が多い DNA 断片や繰り返し回数が少ない DNA 断片ができることにより現れるノイズデータである。元の配列パターンと似て非なる長さの DNA 断片が生じることで，個体が持つ配列パターンの DNA 断片長を示すデータの識別が困難になる。また+A データは，元の DNA 断片のデータよりも 1 塩基だけ長い DNA 断片長に現れるデータである。DNA 断片の片方の端に一つ DNA を付加してしまうことにより発生する。この付加は，個体が持つ配列パターンに由来するデータだけでなく stutter データでも観察される。上記のようなノイズデータだけでなく，複数種類の unit の混在や，繰り返し配列中の別の配列の挿入も，実験結果からの個体が持つ配列パターンに由来する真のデータの識別を複雑で難しいものにする。

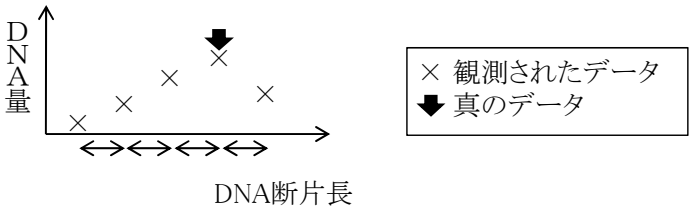
## 2.3 真のデータの識別アルゴリズム

### 2.3.1 アルゴリズムの概要

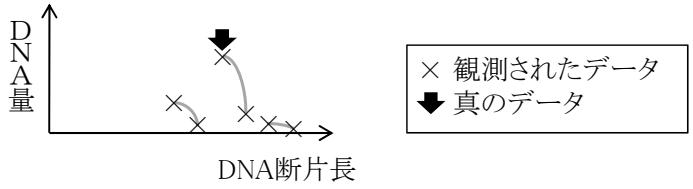
AutoTyper は，stutter データおよび+A データの下記の特徴に基づき，個体が持つ配列パターンに由来する真のデータを識別する。stutter データの現れ方は DNA 配列によって決まり，図 2.3 中の(A)および(B)に示すように，真のデータの周辺に unit 長の間隔で現れる。また，同一のマイクロサテライトマーカーに対して同じ繰り返し回数の配列パターンを持つのであれば，異なる個体間でも stutter データの現れ方は類似していると言われている[Perlin1994]。さらに，真のデータとの DNA 断片長の差が同じ stutter データに着目した場合，stutter データと真のデータの DNA 量の比と，真のデータの DNA 断片長とは線形関係にあることが報告されている[Lipkin1998]。これらの性質により，真のデータを容易に識別できるようなシンプルな実験結果を持つ個体を数個体集めることにより，マイクロサテライトマーカーごとに線形回帰直線を求めることができる。複雑な実験結果を持つ個体では stutter データの中に真のデータが埋没してしまう。そのような個体でも線形回帰直線があれば, 真のデータの DNA 断片長から, 真のデータと stutter データとの DNA 量の比を計算することができる。線形回帰直線から計算した比を，真のデータの候補とそ



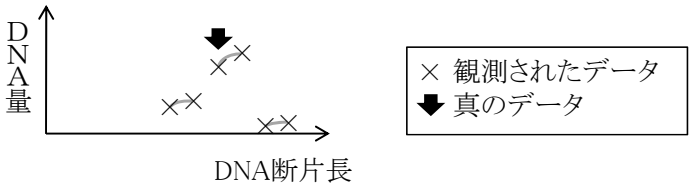

(A)　stutter データが真のデータの左右に等間隔で現れ，DNA 量が少ない例




(B)　sutter データの DNA 量が多く，
真のデータから DNA 断片長が離れた場所にも等間隔で現れ続ける例




(C)　+A データとその元となったデータとの DNA 量の比が一定であり
+A データの方が DNA 量が少ない例




(D)　+A データの方がその元となったデータよりも DNA 量が多い例

図 2.3　stutter データおよび+A データの特徴

の stutter データの DNA 量の比と比較することで，AutoTyper は真のデータとして最も妥当性が高いものはどれか推定することができる。また，+A データとその元となった真のデータまたは stutter データにおける DNA 量の比は, 同一のマイクロサテライトマーカーであっても個体によって変動する[Hu1993] [Magnuson1996] [Smith1995]。さらに，図 2.3 中の(D)に示すように, +A データはその元となったデータよりも DNA 量が多くなる場合もある。しかし，単一の個体に着目した場合は，図 2.3 中の(C)および(D)に示すように，+A データとその元となったデータとの DNA 量の比はおおむね同じ値である。真のデータの識別結果が正しければ, +A データと対応するデータとの DNA 量の比がおおむね一定になることから，複雑な実験結果からも個体が持つ配列パターンに由来すると思われるデータと対応する+A データを識別することができる。stutter データの場合と同じくまず個体が持つ配列パターンに由来する真のデータを容易に識別できるようなシンプルな実験結果を選ぶ。次にシンプルな実験結果における真のデータまたは stutter データと対応する+A データとの DNA 量の比の最大値および最小値を求める。その後，残りの実験結果について真のデータの候補とその+A データの DNA 量の比を求める。もしこの比が上記の範囲外であれば，除去しきれなかった背景ノイズやデータ欠損が影響していると考え，最大値または最小値で置き換える。

AutoTyper は，単一のマイクロサテライトマーカーについての複数の個体の実験結果を一度に処理する入力データとして受け取り, 図 2.4 に示すように大きく二つのステップで, 各個体の実験データから真のデータを識別する。第一のステップでは，入力データにおける stutter データと+A データの現れ方を確認し，真のデータに対し stutter データと+A データを加えて「シンプルな実験結果を示す個体に基づく実験結果の推定値」を計算できるようにする。第二のステップでは，各個体に対して得られた実権結果に基づいて「シンプルな実験結果を示す個体に基づく実験結果の推定値」を求めることで，真のデータを識別する。

### 2.3.2 シンプルな実験結果を示す個体に基づく実験結果の推定値の計算の詳細

「シンプルな実験結果を示す個体に基づく実験結果の推定値」の計算は，図 2.4 に示すように三つの部分から成る。

（1）シンプルな実験結果の選択

（2）stutter データの線形回帰直線および+A データの比の範囲の計算

（3）真のデータごとの「シンプルな実験結果を示す個体に基づく実験結果の推定値」の計算

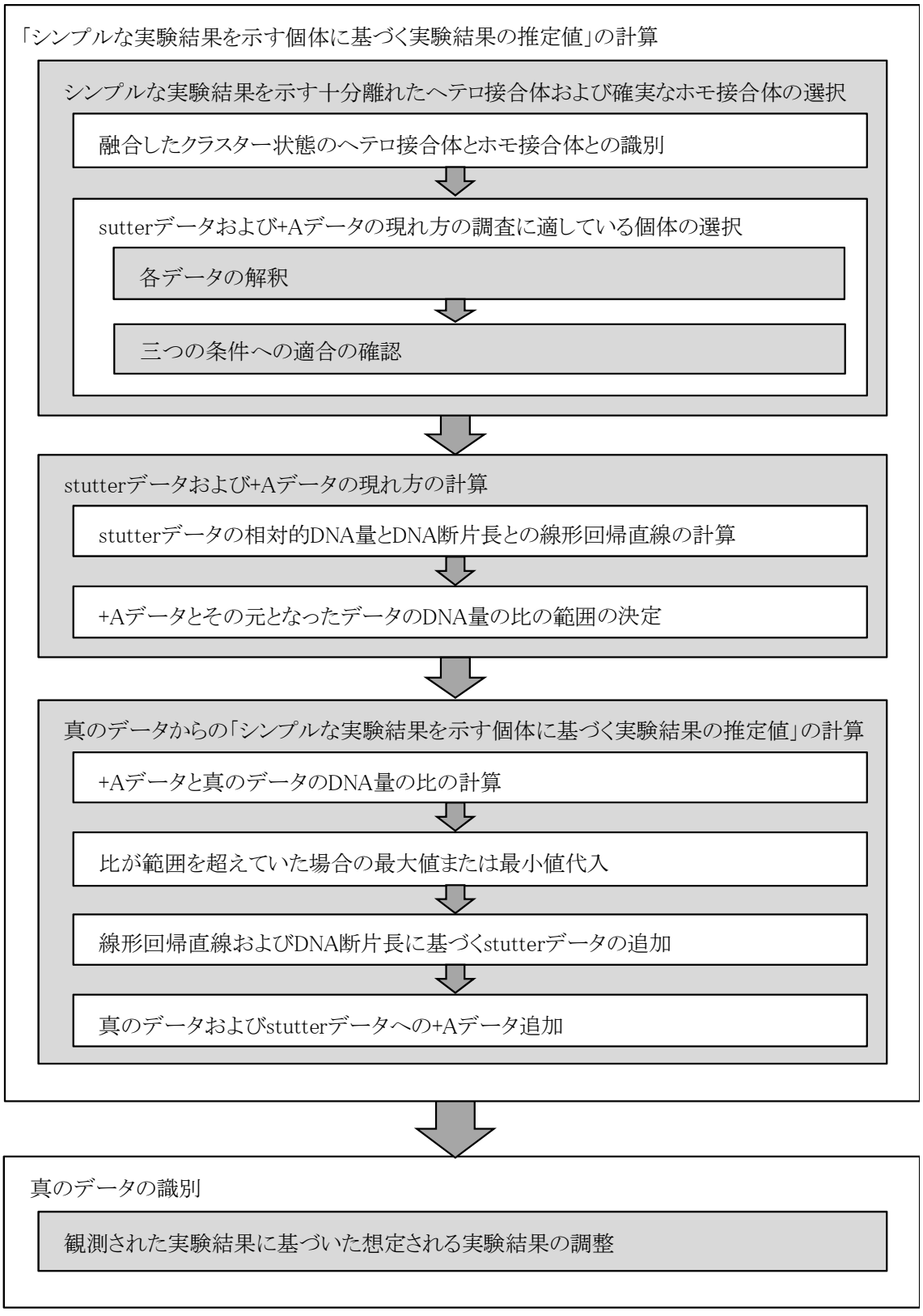



**図 2.4　アルゴリズムの概要**

### （1）シンプルな実験結果の選択

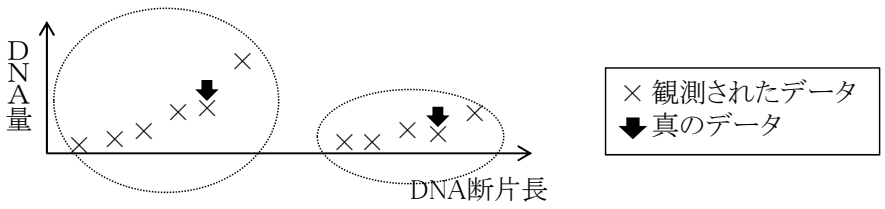
AutoTyper は二つの真のデータが十分離れたヘテロ接合体および確実なホモ接合体の実験結果を選ぶ。これらの個体の実験結果はシンプルなので，比較的容易に解釈できる。ホモ接合体は図 2.5 中の(D)に示すように単一のクラスター状のデータから成る実験結果を示す。二つの真のデータの DNA 断片長が十分離れたヘテロ接合体の実験結果は，図 2.5 中の(A)に例示するような二つのクラスター状のデータを示すため容易に選ぶことができる。二つの真のデータの DNA 断片長が近接するヘテロ接合体は，図 2.5 中の(B)や(C)に例示するような融合したクラスター状のデータから成る，様々な複雑な実験結果となる。シンプルな実験結果の選択は，以下に述べる二つの計算により行う。

### （1-1）融合したクラスター状態のヘテロ接合体とホモ接合体との識別

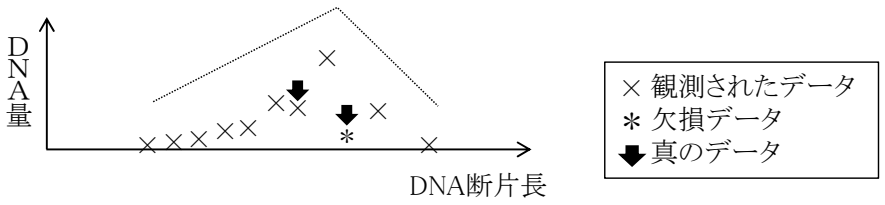
融合したクラスター状態のデータを持つヘテロ接合体とホモ接合体とを識別する。このため図 2.5 中の(A)に示すような完全に分離した二つのクラスター状のデータを持つヘテロ接合体および，図 2.5 中の(C)に示すような融合したクラスター状のデータを持ち DNA 量の変化が二峰性であるヘテロ接合体についての情報を利用する。これらのクラスターにおいて, DNA 量が極大であるデータから DNA 断片長をどれだけ増減させたデータが存在するかを調べる。ホモ接合体においても最も DNA 量が多いデータから同じだけ DNA 断片長を増減させた stutter データおよび+A データとして持つと仮定し, 図 2.5 中の(B)や(D)に例示するような一峰性のデータを持つ個体についてホモ接合体であるかどうか正確に判断する。

### （1-2）stutter データおよび+A データの現れ方の調査に適した個体の選択

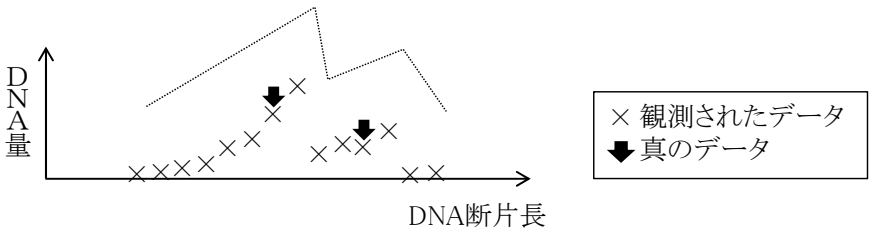
選択した個体における実験結果から stutter データおよび+A データの現れ方の調査に適しているものを選択する。それぞれの個体について三つの条件に適合するか確認することで，stutter データや+A データ以外の様々なノイズデータが実験結果に含まれていないかを調べる。第一に，ヘテロ接合体においては，二つのクラスターに含まれるデータの数はほぼ同じでなくてはならない。もしデータの数の差が 2 より大きいならば，その実験結果はランダムに発生するノイズデータを伴ったホモ接合体であると考える。第二に，真のデータの DNA 量は，対応する+A データの DNA 量の 20%より高くなくてはならない。もしそうでないなら，真のデータと考えたデータは+A データだったと考える。閾値の 20%は予備調査に基づいて設定した。第三に, 真のデータとその+A データとの DNA 量の比は，選択した個体間でほぼ同じでなくてはならない[Magnuson1996] [Smith1995]。



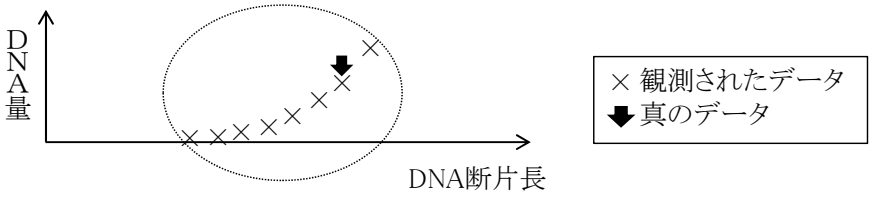

(A)　完全に分離した二つのクラスター状のデータから成る
シンプルな実験結果を示すヘテロ接合体




(B)　欠損データを含む一峰性のヘテロ接合体




(C)　二峰性の融合したクラスター状のデータを持つヘテロ接合体



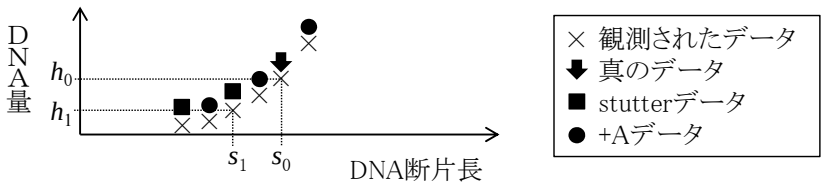

(D)　単一のクラスター状のデータから成るシンプルな実験結果を示すホモ接合体

図 2.5　シンプルな実験結果および複雑な実験結果の例

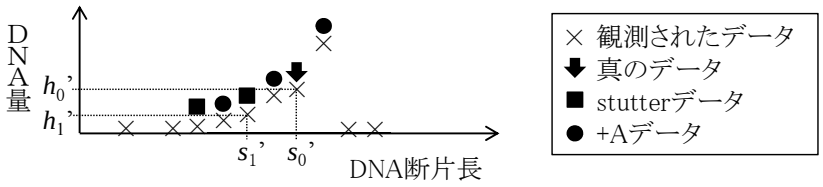
上記の条件を満たしているか調べるためには，それぞれのデータが真のデータ，stutter データおよび+A データのいずれであるか解釈されている必要がある。unit 長が 2 塩基よりも長いマイクロサテライトマーカーでは，クラスターにおいて 1 塩基離れたデータの組を容易に同定できる。最も DNA 量が多いデータを含む組において，DNA 断片長が短い方が真のデータ，長い方がその+A データであると解釈すれば良い。その他にもデータの組があれば， stutter データとその+A データと解釈できる。これに対し，unit 長が 2 塩基のマイクロサテライトマーカーでは， 1 塩基間隔で連なった DNA 断片長を持つデータがクラスター内に現れることが多いため，データの解釈は複雑なものになる。しかし，unit 長が 2 塩基のマイクロサテライトマーカーにおいても，+A データの隣にあるデータは真のデータまたは stutter データであること, および stutter データは真のデータよりも DNA 量が少ないことは仮定できる。これらのことから，最も DNA 量が多いデータは，真のデータまたはその+A データのどちらかであると考えることができる。最も DNA 量が多いデータが真のデータとその+A データのどちらであるかを決めるため, 最も DNA 量が多いデータが真のデータであると仮定した場合とその+A データであると仮定した場合のそれぞれについて, +A データとその元となったデータとの DNA 量の比の分散を求める。ただし，DNA 量が非常に少ないデータは定量性が低いので，最も DNA 量が多いデータの 15%以上の DNA 量を持つデータのみを比の分散の計算対象とする。比は各データの組で一定になる[Magnuson1996]ため，仮定が正しければ分散の値は小さくなる。そこで AutoTyper は分散が 0.01 未満となる仮定を採用する。もし両方の仮定で分散が 0.01 未満になるならば，最も DNA 量が多いデータは真のデータであると考える。これは，+A データが現れないマイクロサテライトマーカーでは両方の仮定で分散が小さくなりやすいという観察に基づく。逆に，両方の仮定で分散が 0.01 以上になった実験結果は，最も DNA 量が多いデータは電気泳動時に発生した気泡などに由来すると思われるため，ノイズデータの現れ方の調査に用いる実験結果から除く。

### (2) stutter データの線形回帰直線および+A データの比の範囲の決定

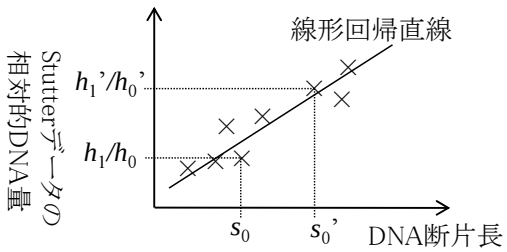
stutter データの現れ方の度合いを調べるため， AutoTyper はシンプルな実験結果を用いて下記の手順で線形回帰直線を求める。図 2.6 中の(A)および(B)に，stutter データの現れ方の計算に用いるとして選択されたホモ接合体を示す。これらの個体は DNA 断片長が異なる配列パターンを持つとする。図 2.6 中の(A)における個体が持つ配列パターンに由来する真のデータの DNA 断片長と DNA 量を $s_0$ および $h_0$ で表す。$k$ 回繰り返し回数が少ない stutter データの DNA 断片長と DNA 量を $s_k$ および $h_k$ で表す。座標$(s_0, h_k/h_0)$を図 2.6




(A)　stutter データの現れ方の計算に用いるため選択されたホモ接合体の一個体目の例




(B)　stutter データの現れ方の計算に用いるため選択されたホモ接合体の二個体目の例




(C)　線形回帰直線の計算

**図 2.6　stutter の相対的高さと DNA 断片長との線形回帰直線の計算**

中の(C)のようにプロットする。選択したシンプルな実験結果の全てから同様にプロットを行う。stutter データの相対的 DNA 量の線形性[Lipkin1998]に基づき，$k$ 回繰り返し回数が少ない stutter データと真のデータの DNA 量の比と DNA 断片長の間の線形回帰直線を次の式のように求める。

$$\frac{k\text{回だけ繰り返し回数が増減したstutterデータのDNA量}}{\text{アリルに由来するデータのDNA量}} = a_k * \text{DNA断片長} + b_k$$

上記の方法で計算した線形回帰直線を使うことで，任意の配列パターンに由来する真の

データに対して $k$ 回繰り返し回数が少ない stutter データの相対的な DNA 量を見積もることができる。図 2.3 中の(B)のように真のデータよりも DNA 断片長が大きい stutter データにも対応するため，$k$ が負の場合に対しても線形回帰直線を求める。さらに，誤った線形回帰直線が計算されるのを防ぐための確認処理を行う。例えば，stutter データの DNA 量が負になる，あるいは真のデータよりも DNA 量が多くなったりした場合，線形回帰直線を破棄し，stutter データの相対的 DNA 量は全ての配列パターンで一定，すなわち，線形回帰直線の傾きは 0 であると修正する。予備実験を行い目視で線形回帰直線を求めたところ傾きは 0.05 未満だったことから，この修正は stutter データの現れ方を見積もるために有効であると考える。特に，線形回帰直線を求めるために集めたシンプルな実験結果において，真のデータの DNA 断片長が狭い範囲に集中していた場合には傾きが過大になりやすいため，この修正は有効である。

+A データと真のデータの DNA 量の比の範囲は，選択したシンプルな実験結果から求めた値によって求める。それぞれの実験結果から比を計算して，比の最大値 $r_{max}$ と最小値 $r_{min}$ を求め，DNA 量の比の範囲とする。

**(3) 各個体・配列パターンごとの「シンプルな実験結果を示す個体に基づく実験結果の推定値」の計算**

上記で述べた stutter データについての線形回帰直線および+A データの比の範囲により，任意の個体および配列パターンに由来する真のデータに対して「シンプルな実験結果を示す個体に基づく実験結果の推定値」を計算することができる。真のデータの DNA 断片長と DNA 量を $s$ および $h$，+A データと真のデータの DNA 量の比を $r$ とする。「シンプルな実験結果を示す個体に基づく実験結果の推定値」における真のデータより $k$ 回繰り返し回数が少ない stutter データの DNA 量は $h*(a_k*s+b_k)$，その+A データの DNA 量は $r*h*(a_k*s+b_k)$ となる。真のデータよりも繰り返し回数が多い stutter データについても同様に計算できる。ただし，$r>r_{max}$ または $r<r_{min}$ である場合は，それぞれ $r$ を $r_{max}$ または $r_{min}$ で置き換える。このように各個体ごとに $r$ を求める処理により，個体によって真のデータの DNA 量が+A データの DNA 量よりも多かったり少なかったりする場合でも，真のデータを正確に識別することを狙う。

### 2.3.3 真のデータを識別するアルゴリズムの詳細

この節では，「シンプルな実験結果を示す個体に基づく実験結果の推定値」を観測された実験結果に合わせて調節し，真のデータを識別する方法について述べる。ホモ接合体にお

いて得られる実験結果は「シンプルな実験結果を示す個体に基づく実験結果の推定値」と類似しており，ヘテロ接合体において得られる実験結果は二つの「シンプルな実験結果を示す個体に基づく実験結果の推定値」を重ね合わせたものに類似しているはずである。AutoTyper ではヘテロ接合体とホモ接合体のどちらであるかおよび，真のデータと+A データの DNA 量の大小関係の可能性を網羅すると共に，図 2.5 中の(B)でアスタリスクで示した欠損データに対応するため，AutoTyper では各個体について，下記の六つの仮説を検討する。それぞれの仮説の下で「シンプルな実験結果を示す個体に基づく実験結果の推定値」と観測された実験結果との差をスコアとして求め，六つの仮説のうちスコアが最も小さいものを選ぶ。

(1) 処理対象の個体はホモ接合体であり，真のデータはその+A データよりも DNA 量が多い
(2) 処理対象の個体はヘテロ接合体であり，二つの真のデータはいずれもその+A データよりも DNA 量が多い
(3) 処理対象の個体はヘテロ接合体であり，二つの真のデータのうち DNA 量が多い方のデータはその+A データよりも DNA 量が多く，もう片方はその+A データよりも DNA 量が少ない
(4) 処理対象の個体はホモ接合体であり，真のデータはその+A データよりも DNA 量が少ない
(5) 処理対象の個体はヘテロ接合体であり，二つの真のデータのうち DNA 量が多い方のデータはその+A データよりも DNA 量が少なく，もう片方はその+A データよりも DNA 量が多い
(6) 処理対象の個体はヘテロ接合体であり，二つの真のデータはいずれもその+A データよりも DNA 量が少ない

図 2.7 を参照して，「シンプルな実験結果を示す個体に基づく実験結果の推定値」と観測された実験結果の差をスコアとして求める手順を説明する。まず，仮説(1)に基づいて，最も DNA 量が多いデータ Pmax を真のデータと考える。図 2.7 のプラス記号は仮説(1)における「シンプルな実験結果を示す個体に基づく実験結果の推定値」を示し，矢印は観測された DNA 量と「シンプルな実験結果を示す個体に基づく実験結果の推定値」での DNA 量の差を示す。$i$ 番目のデータにおける，観測された DNA 量と「シンプルな実験結果を示す個体に基づく実験結果の推定値」での DNA 量の差を $h_i$ とする。スコア $diff_1$ は

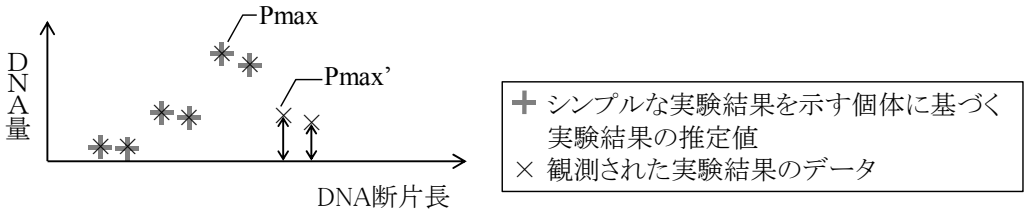



**図 2.7　「シンプルな実験結果を示す個体に基づく実験結果の推定値」と観測された実験結果との比較**

$diff_1 = \sum_i h_i^2$ で定義される。次に，AutoTyper は仮説(2)に基づいて，データ Pmax および DNA 量の差 $h_i$ が最も大きいデータ Pmax'を真のデータであると考える。Pmax と Pmax'の DNA 断片長の差は unit 長の整数倍でなくても構わないとする。これにより，複数種類の繰り返し配列が混在する，あるいは繰り返し配列の中に別の配列が挟まっているマイクロサテライトマーカーにおいて unit 長の整数倍でない間隔で真のデータが存在しても，真のデータを正確に識別することを狙う。差のスコア $diff_2$ は，Pmax および Pmax' から計算された「シンプルな実験結果を示す個体に基づく実験結果の推定値」を二つ重ね合わせたものと，観測された実験結果における各データの DNA 量の差の二乗を足し合わせたものである。$diff_3$ は仮説(3)，すなわち Pmax が真のデータ，Pmax'が真のデータの+A データと考えた場合について同様に計算する。$diff_4$，$diff_5$ および $diff_6$ は同様に仮説(4)，(5)および(6)に対応し，データ Pmax を真のデータの+A データであるとして計算する。

「シンプルな実験結果を示す個体に基づく実験結果の推定値」を観測された実験結果に合わせるため，「シンプルな実験結果を示す個体に基づく実験結果の推定値」に対し下記の三種類の調整を行う。

(1) 実験結果においては DNA 断片長は整数ではなく実数として与えられるため，「シンプルな実験結果を示す個体に基づく実験結果の推定値」を $x$ 軸に沿って調整する。「シンプルな実験結果を示す個体に基づく実験結果の推定値」において stutter データおよびその+A データの DNA 断片長は Pmax および Pmax'の DNA 断片長と unit 長から計算できるので，その±0.5 塩基以内にデータがあれば stutter データまたは+A データであると考えて DNA 断片長を調整する。電気泳動による DNA 断片長の分解能は約 1 塩基であるため，上記の調整により DNA 断片長が最も近いデータ同士を照合することができる。

(2) 図 2.7 における仮説(2)のように二つの真のデータが近接している場合，データ Pmax の DNA 量は, 一番目の真のデータの DNA 量と二番目の真のデータに由来する stutter データの DNA 量の和であると考えられる。このため,「シンプルな実験結果を示す個体に基づく実験結果の推定値」の各データの DNA 量は過大に見積もられている可能性がある。そこで,「シンプルな実験結果を示す個体に基づく実験結果の推定値」に含まれる各データの DNA 量を定数倍し，*diff* の値を最小化する。ヘテロ接合体を仮定する仮説(2)，(3)，(5)および(6)においては，二つの定数を用いる。この調整により，それぞれの仮説が実験結果と適合している度合いをより正確にスコアとして求めることができる。

(3) 背景ノイズが連続したものを真のデータとその stutter データおよび+A データだと解釈するのを防ぐ必要がある。このため，二つの真のデータのうち DNA 量が少ないものがもう片方の DNA 量の 20%以上である場合に限り，その個体はヘテロ接合体であると考える。この調整により，真にヘテロ接合体だと考えるべき実験結果を背景ノイズを伴ったホモ接合体の実験結果から識別することができる。閾値の 20%は予備調査に基づいて設定した。また，この値は既存の研究とも一致する[Stoughton1997]。

## 2.4 実験結果

### 2.4.1 実験に用いたデータ

評価用データとして，250 人の日本人サンプルで観測された，174 個のマイクロサテライトマーカーを用いた[Tamiya2005]。174 個のマイクロサテライトマーカーそれぞれに対して三個体ずつ実験を行い，図 2.8 に示す四種類のクラスに分類した。

(A) 全ての個体において真のデータは対応する+A データよりも DNA 量が多く(または全ての個体において少なく)，真のデータより短い DNA 断片長の stutter データのみが現れるマイクロサテライトマーカー

(B) 全ての個体において真のデータは対応する+A データよりも DNA 量が多く(または全ての個体において少なく)，真のデータより長い DNA 断片長および短い DNA 断片長の stutter データの両方が現れるマイクロサテライトマーカー

(C) 個体によって真のデータが対応する+A データより DNA 量が多い場合も少ない場合もあるマイクロサテライトマーカー

(D) データの現れ方が複雑であるため，どれが真のデータ，stutter データ，+A データな

のか専門家にも識別できないマイクロサテライトマーカー

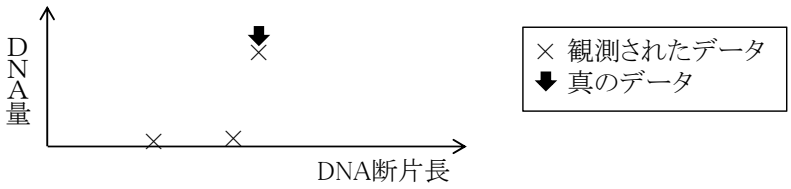
それぞれのクラスに対し174個のマイクロサテライトマーカーからそれぞれ典型的なマイクロサテライトマーカー例を選択した。クラス(A)に対しては, 図2.8中の(Aa)および(Ab)に示す2通りのマイクロサテライトマーカーを選択した。図2.8中の(Aa)に示すマイクロサテライトマーカーはほとんど+Aデータもstutterデータも発生しないものである。これらの5つの典型的なマイクロサテライトマーカーにおいて250人の個体に対して実験を行ったデータを用いて，AutoTyper による真のデータを識別する精度を評価した結果を表2.1 に示す。専門家の目視による識別結果との比較により，正しく識別できた個体の割合を精度として求めた。なお，クラス(D)のマイクロサテライトマーカーについては，実験条件を調整しての再実験および配列パターンの頻度の遺伝法則[Crow1988]への適合度の確認に基づく専門家の試行錯誤により，専門家の目視による識別結果を与えた。

### 2.4.2 実験結果

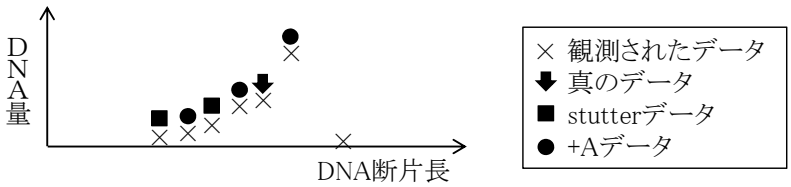
AutoTyper および，既存の商用ソフトウェア Genotyper software [Tereba1999]，GeneMapper [Leibelt2003]による真のデータを識別する精度を表2.1に示す。これらのツールはマイクロサテライトマーカーの実験で広く利用されているため，比較対象として選択した。Genotyper softwareは，stutterは真のデータより短いDNA断片長のもののみが現れ, +Aデータは真のデータよりDNA両が少ないもののみ現れると仮定して真のデータの識別を行う。GeneMapperはGenotyper softwareの後継製品であるが，アルゴリズムの詳細は公開されていない。真のデータの識別機能に加え，個体ごとに実験の制度を推定して“pass”, “check”, および“low-quality”のいずれかのラベルを付与する機能を持つ。その他の既存技術 [Perlin1994] [Stoughton1997] [Palsson1999]は現在利用できないため比較できなかった。クラス(A)，(B)および(C)について，AutoTyperは GeneMapper および Genotyper software よりも高い精度を達成した。しかし，クラス(D)においてはGeneMapperおよびGenotyper softwareの方がAutoTyperよりも高精度であった。

GeneMapperにより“pass”とラベル付けされたサンプルにおける精度を表2.2に示す。クラス(B)および(D)に対しては，数個体しか“pass”が割り当てられなかったため，精度計算の対象外とした。

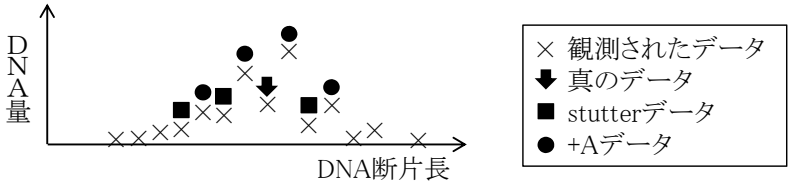
次に，AutoTyper，GeneMapperおよびGenotyper softwareの平均精度の見積もりを下記の手順で行った。表2.3に, 174個のマイクロサテライトマーカーをクラスおよびunit長で分類した結果を示す。ヒトゲノム全体に存在する 20,431 のマイクロサテライトマー



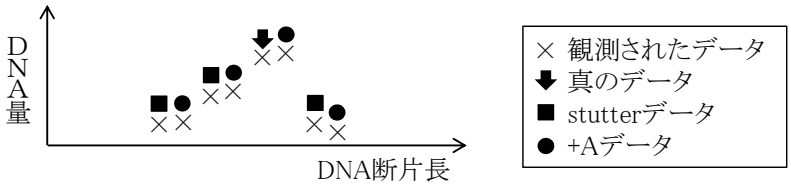

(Aa)　+A データも stutter データもほとんど現れない例



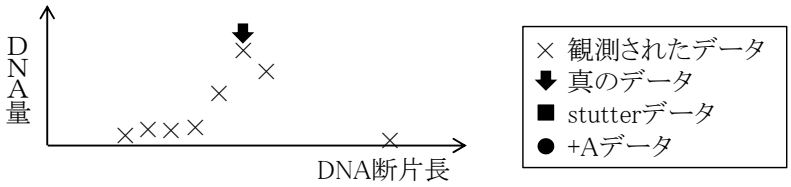

(Ab)　+A データは常に元となるデータより DNA 量が少なく，
真のデータより短い DNA 断片長の stutter データのみが現れる例




(B)　真のデータより長い DNA 断片長および短い断片長の stutter データの両方が現れる例




(C)　+A データが元となるデータよりも DNA 量が多い場合も少ない場合もある例




(D)　どれが真のデータ，stutter データ，+A データなのか専門家にも識別できない例

図 2.8　評価に用いたマイクロサテライトマーカーの実験結果の例

表 2.1　真のデータの識別精度

| クラス | 名称 | unit 長 | 精度(%) | | |
|---|---|---|---|---|---|
| | | | Genotyper | GeneMapper | AutoTyper |
| (Aa) | D6S1038 | 4 塩基 | 96 | 97 | 99 |
| (Ab) | D2S374 | 2 塩基 | 83 | 85 | 93 |
| (B) | D4S1087i | 2 塩基 | 44 | 75 | 86 |
| (C) | D20S910 | 2 塩基 | 61 | 67 | 86 |
| (D) | HUMUT1988 | 4 塩基 | 98 | 91 | 32 |

表 2.2　"pass"とラベルされた個体における真のデータの識別精度

| クラス | 名称 | "pass"とラベルされた個体における GeneMapper 精度 | "pass"とラベルされた個体数 |
|---|---|---|---|
| (Aa) | D6S1038 | 98% | 232 |
| (Ab) | D2S374 | 86% | 224 |
| (B) | D4S1087i | ―― | 5 |
| (C) | D20S910 | 84% | 156 |
| (D) | HUMUT1988 | ―― | 34 |

表 2.3　クラスおよび unit 長ごとのマイクロサテライトマーカー数

| クラス | unit 長 | | | | |
|---|---|---|---|---|---|
| | 2 塩基 | 3 塩基 | 4 塩基 | 5 塩基 | 合計 |
| (A) | 62 | 9 | 65 | 7 | 143 |
| (B) | 4 | 1 | 11 | 0 | 16 |
| (C) | 7 | 0 | 2 | 0 | 9 |
| (D) | 0 | 0 | 5 | 1 | 6 |
| 合計 | 73 | 10 | 83 | 8 | 174 |

カー[Tamiya2005]において, 73%の unit 長が 2 塩基, 7%の unit 長が 3 塩基, 17%の unit 長が 4 塩基，そして 3%の unit 長が 5 塩基であった。このことから，表 2.3 の比率を用いて，84%のマイクロサテライトマーカーがクラス(A)，7%がクラス(B)，7%がクラス(C)，1%がクラス(D)と見積もることができる。この見積もりの下で，Genotyper software およ

び GeneMapper との精度比較に基づき, ヒトゲノム全体のマイクロサテライトマーカーに対する AutoTyper の精度は下記のように考えられる。

- ヒトのマイクロサテライトマーカー全体の 7%+7%=14%を占めるクラス(B)および(C)のマイクロサテライトマーカーにおいて Genotyper software および GeneMapper よりも明確に高い精度を達成する
- 84％を占めるクラス(A)のマイクロサテライトマーカーにおいて，Genotyper software および GeneMapper の精度をやや上回る
- 1％を占めるクラス(D)のマイクロサテライトマーカーにおいて，Genotyper software および GeneMapper の精度を下回る

各クラスのマイクロサテライトマーカーのヒトゲノム全体における頻度の見積もりと，それぞれのアルゴリズムがそれぞれのクラスのマイクロサテライトマーカーに対して真のデータの識別から，ヒトゲノム上のマイクロサテライトマーカー全体に対する平均精度を求めた。Genotyper software は 85%，GeneMapper は 88%の精度でノイズデータから真のデータを識別できる。AutoTyper は Genotyper software および GeneMapper の両方を上回り，94%の精度で識別を行える。

## 2.5 考察

### 2.5.1 ノイズデータの傾向の正確性

AutoTyper は以下の三種類の特徴を持つため，入力データからノイズピークの傾向を正確に計算し，真のデータの識別精度を改善させることができた。三つの特徴はいずれも DNA データの測定原理に由来するノイズデータの法則性に関する専門家の知見[Lipkin1998] [Hu1993] [Magnuson1996] [Smith1995]を利用するものであり，強化学習[木村 1999]，過学習の防止[銅谷 2005]および，属性選択[Yang1997]など機械学習分野で提案されてきた技術とは異なるアプローチである。

第一の特徴として，2.3.2 節の(2)で述べたように，AutoTyper は真のデータより DNA 断片長が短い stutter データだけでなく DNA 断片長が長い stutter データも考慮に入れる。このため，クラス(B)のマイクロサテライトマーカーについても高い精度を達成できた。これに対し，Genotyper software は真のデータよりも短い DNA 断片長を持つ stutter データのみ識別するため，クラス(B)のマイクロサテライトマーカーに対する推定では多くの個

体で false positive を生じてしまった。すなわち，真のデータよりも 1 回繰り返し回数が多い stutter データを二つ目の真のデータと推定し，ホモ接合体を誤ってヘテロ接合体と解釈してしまった。また，2.4 節での実験結果から，評価に用いたバージョンの GeneMapper では真のデータより長い DNA 断片長を持つ stutter データを識別できないことが示された。

第二の特徴として，AutoTyper は 2.3.2 節の(3)で述べたように個体ごとに+A データの現れ方を調べることが挙げられる。この特徴により，クラス(C)のような，個体によって+A データの DNA 量が元となるデータの DNA 量よりも多かったり少なかったりするマイクロサテライトマーカーでも，より正確に真のデータを識別できた。これに対し，Genotyper は+A データは元となるデータよりも必ず DNA 量が少ないと仮定しているため，+A データが真のデータより DNA 量が多いサンプルにおいて，真のデータの識別を正しく行うことができなかった。GeneMapper は，+A データが真のデータより DNA 量が少ない場合のみ，“pass”というラベルを付与していると思われ，これにより AutoTyper と同等の精度を達成した。

第三の特徴として，stutter データの現れ方を調べるために，AutoTyper はシンプルな実験結果を示す個体を正確に選ぶ。Perlin らのアルゴリズム，Stoughton のアルゴリズム[Stoughton1997]および AutoTyper はいずれも，二つの真のデータが十分離れたヘテロ接合体または確実なホモ接合体を stutter データの現れ方を調べるために利用する。AutoTyper は 2.3.2 節の(1-1)で述べた処理により，二つの真のデータが近接したヘテロ接合体とホモ接合体とをより正確に見分けることができる。Perlin らのアルゴリズムは真のデータより短い DNA 断片長を持つ stutter データのみ存在すると想定する。このため，真のデータより長い DNA 断片長の stutter データが現れているホモ接合体をヘテロ接合体と過って解釈する傾向にある。また，Stoughton のアルゴリズムは stutter データの現れ方を調べるためのホモ接合体を下記の手順で選ぶ。まず各配列パターンに対しその配列パターンに由来する真のデータを持ち，かつ二つの真のデータが十分離れたヘテロ接合体を選ぶ。そのようなヘテロ接合体を持たない配列パターンに対し，その配列パターンに由来する真のデータを持つ中で最も幅の狭い波形を持つ個体がホモ接合体であると仮定する。これにより，最も幅の狭いヘテロ接合体をホモ接合体と誤って解釈してしまう可能性がある。特に，その配列パターンを持つ個体が全てヘテロ接合体であり，二つの近接した真のデータを持つ場合に誤解釈が生じやすいと考えられる。

### 2.5.2 多様な実験結果への対応

AutoTyper は下記の三種類の特徴により，多様な実験結果において真のデータをロバストに識別することができる。第一に，AutoTyper は 2.3.3 節で述べたように二つの真のデータの DNA 断片長の間隔が unit 長の整数倍だとの仮定を置かないため, unit 長の整数倍でない間隔で真のデータが現れた場合にも対応できる。この点において，AutoTyper は，unit 長の整数倍の間隔を前提としているアルゴリズム[Perlin1994]より多くのマイクロサテライトマーカーに適用できる。

AutoTyper が持つ第二の特徴は，2.3.3 節の調整(2)で述べたようにヘテロ接合体において二つの真のデータの DNA 量が等しくない可能性を考慮することである。既存アルゴリズムではヘテロ接合体は二つの真のデータの DNA 量が同じであると仮定する場合があるが[Perlin1994]，実際の実験結果においては 30%以上のヘテロ接合体は片方のデータの DNA 量はもう片方より 2 倍以上多い。このため，二つの真のデータの DNA 量は常に等しいという仮定の下では，ヘテロ接合体の実験結果を正しく解釈できない可能性がある。

さらに，AutoTyper は，データが欠損する可能性についても考慮している。Perlin らのアルゴリズムと Genotyper software は図 2.5 中の(B)でアスタリスクで示したデータのように，左右に隣接するデータの方が DNA 量が多いため埋没してしまうと，真のデータの識別を間違えることがある。しかし，AutoTyper では実験結果では常に「真のデータと対応する+A データ」という組で現れるという考えに基づき，もしデータが一つだけ現れたなら欠損データがあるとして処理を行うので，このような実験結果に対しても正しく解釈することができる。欠損したデータの DNA 量は分からないので，2.3.3 節の調整(1)で述べたように，観測可能な値よりも少なかったと考え，AutoTyper は下記の二つの可能性を評価する。第一の可能性は，真のデータは認識できたが，対応する+A データは認識できなかったとするものである。これを仮説 $H_0$ と呼ぶ。第二の可能性は+A データは認識できたがその元となった真のデータは認識できなかったとするものである。これを仮説 $H_1$ と呼ぶ。2.3.3 節で述べた真のデータを識別するアルゴリズムにおいて, 差のスコア $diff_1$, $diff_2$ および，$diff_3$ を求める処理は仮説 $H_0$ に，$diff_4$，$diff_5$ および，$diff_6$ を求める処理は仮説 $H_1$ に基づいている。+A データとその元となるデータの関係は，同一の個体の実験結果の中では再現性があり，stutter データとその+A データであっても真のデータとその+A データであってもほぼ等しい比を持つ[Magnuson1996]。この関係により，AutoTyper は認識できなかったデータの存在を仮定して推定を行うことができる。

## 2.6 結言

DNA 実験データにおいて stutter データおよび+A データの現れ方を自動的に見積もることで, 真のデータをノイズデータから正確に抽出するアルゴリズム AutoTyper を提案した。AutoTyper は，stutter データの現れ方には線形性があり+A データの比率はマーカーごとに概ね一定であるという専門家の知見に基づく。まず，真のデータを容易に解釈できるようなシンプルな実験結果を持つ個体を集め, stutter データについての線形回帰直線および+A データの相対的 DNA 量の範囲を求めることで専門家の知見をマイクロサテライトマーカーごとに定量化する。そして，これらに基づいて真のデータの候補から「シンプルな実験結果を示す個体に基づく実験結果の推定値」を求めて実際に観測された結果と比較し，真のデータの識別を行う。174 個のマイクロサテライトマーカーを用いた実験により, ヒトゲノム上のマイクロサテライトマーカー全体に対する平均で 94%の精度を達成した。

2.4 節で述べた通り，AutoTyper はクラス(D)以外のマイクロサテライトマーカーで Genotyper および GeneMapper よりも高い識別精度を達成した。識別精度をさらに改善するためには，マイクロサテライトマーカーごとに最適なアルゴリズムを選択する方法がある。例えば，DNA 配列に A の繰り返しが含まれている場合はクラス(D)の波形が現れる可能性が高いことから，自動的に最適なアルゴリズムを選ぶためには DNA 配列情報が有効であると考えられる。また，GeneMapper や Palsson らの技術[Palsson1999]にあるように，真のデータの自動識別技術に識別結果の品質評価技術を組み合わせることで，真のデータの識別をさらに効率化することができる。現在の AutoTyper アルゴリズムでは，実験エラーのため様々な状況での真のデータの識別の失敗が避けられない。品質によって真のデータの識別結果を分類し，いくらかでも確実に正しい識別結果を取り出すことができれば，目視確認を要するサンプルの数を減らすことができる。

AutoTyper は，個体数およびマイクロサテライトマーカーの数の両面において大規模化傾向のある DNA 実験データにおいて有効である。個体数によらず stutter データの相対量の線形性[Lipkin1998]および+A データの相対量の均一性により，個体数によらずデータの現れ方を正確に見積もることができる。また，2.5 節で述べたとおり様々なノイズデータの多いマイクロサテライトマーカーでも高精度な識別が可能であり，多くのマイクロサテライトマーカーの実験結果を解釈する場合において特に有効である。さらに, AutoTyper は個体数およびマイクロサテライトマーカー数の両方に対して線形の計算時間を持ち，スケーラビリティが高い。これらの特徴から，多くのマイクロサテライトマーカーの実験結

果を解釈する場合において特に有効であり，リウマチの発症に影響を及ぼす遺伝子のヒトゲノム全域からの探索[Tamiya2005]および，食用牛の肉質に影響を及ぼす遺伝子のウシゲノム全域からの探索[Matsumoto2006]にて利用されている。

また，AutoTyper における，実際に観測したサンプルのうち典型的傾向を示す解釈が容易なものを選ぶことにより専門家の知見に基づく定性的な法則性の定量化を行う特徴は，企業内の様々な定型性の低い数値データから情報を抽出するアルゴリズムに対しても有効である。定型性の低いデータでは業務の内容や遂行状況に応じて様々な内容が様々な書式で記載され，内在的法則性について蓄積された専門家の知見は定性的には成立しても，具体的閾値や係数は変動する場合が多い。時間的・コスト的な問題から予備実験を行って法則性の定量化を行うことが困難である，あるいは固定的な条件の下でのデータを多数集めることが不可能である場合でも，入力サンプルの一部を選んで法則性の定量化を行うことが可能であれば高精度な抽出が可能となる。ただし，このような抽出方式は，専門家が目視で抽出を行う場合の手順と相同性があると考えられることから，個人情報保護や機密情報漏洩防止などの上での問題がなく，専門家による詳細な目視の場合と同等の情報をアルゴリズムが利用できることが前提となる。この前提を満たす数値データからの抽出例として，株や為替の取引業務におけるテクニカル分析が挙げられる。テクニカル分析では，様々な経験則を観測データに当てはめて投資判断を行う[Murphy1986]が，経験則では価格などの尺度やその増減速度の大小を定性的にしか定義していない場合が多い。このため，対象としている金融商品や市場の傾向に合わせて経験則を定量化し，観測データに適用することが正確な解釈に必要である[橋本 2003]。世界経済における金融市場の影響力が拡大し，アルゴリズム取引のように情報システムを活用する取引形態が進展する中で，テクニカル分析はますます重要になりつつある。AutoTyper の上記の特徴を活かし，テクニカル分析における数値データからの高精度な情報抽出技術を開発することは，今後の課題である。

# 第 3 章
# ビジネス文書からのメタデータ抽出用ルールの自動生成技術

## 3.1 緒言

本章では，組織ごとの文書作成上の特徴に基づいたビジネス文書からのタイトル・顧客名・作成日などのメタデータ抽出の導入工数削減を実現するため，メタデータ抽出用ルールを生成する方式について提案する。ビジネス文書をメタデータで整理分類して登録する ECM システムに対するニーズの高まりに伴い，メタデータ抽出手法による文書登録効率化技術が提案されている[Minagawa2006] [Handley2005] [Ishitani1999]。

メタデータ抽出手法においては，抽出用ルールは重要であり必ず設定する必要がある。しかし，抽出用ルールの人手での設定は煩雑であり，ECM システムの導入の障害となっている。メタデータ抽出用ルールとして，自動的にレイアウト特徴を探す手法が提案されている[Esposito2004] [Wnek2002]が，メタデータの相対位置や記載領域が固定的であることを仮定しており，オフィスソフトで作成され柔軟な書式を持つことを特徴とするビジネス文書からのメタデータ抽出用ルールは生成できない。また，論文の参考文献を対象として単語間の遷移確率をもとに著者・論文誌名などを抽出する手法も提案されている[Kramer2007] [Parmentier1997]が，文字の一次元配置を対象としており，数～十文字程度の短い文字列がページ内に二次元的な広がりを持って記載されることが多い営業文書・設計文書・報告書などのビジネス文書からのメタデータ抽出用ルールを生成することはできない。

そこで本章では，サンプル文書と正解メタデータの組を入力として，メタデータ抽出用ルールを生成する Sample-based Collection and Adjustment（以下，SCA 法）と呼ぶ手法を提案する。ビジネス文書は可読性を向上させるための共通的特徴を持つことから，サンプル文書を用いてメタデータの記載の特徴を効率的に調べることができる。SCA 法は二段階の処理方式を用いる。まずサンプル文書でのメタデータの記載のされ方を元にメタデ

ータ抽出用ルールの候補を列挙する。その後候補に対しサンプル文書での記載のされ方を元にした選択および重みの最適化を行う。これにより，最適化の対象を効果が期待できる候補に絞ることで計算時間を削減し，現実的な時間内での抽出用ルールの生成を目指す。

本章の構成は，以下の通りである。まず 3.2 節では対象とする ECM システムおよびメタデータ抽出用ルールの設定の課題について説明する。次に 3.3 節で SCA 法について述べる。3.4 節で SCA 法の有効性を評価するための実験を行い，3.5 節で実験結果の安定性について評価を行う。3.6 節ではまとめを述べる。

## 3.2 ECM システムにおけるメタデータ抽出

### 3.2.1 ECM システムの概要

ECM システムは，企業が保有する文書の統合的な登録，保存，管理，利用を実現し，文書の作成から廃棄に至までのライフサイクル管理を提供するためのシステムである。ECM システムにおいて文書はそれぞれ，内容を示すメタデータと紐付けて管理される。ECM への文書登録におけるデータの流れの概要を図 3.1 に示す。ECM システムはメタデータ抽出処理を内部に持つ。ECM システムの導入においては，管理対象とする文書の種

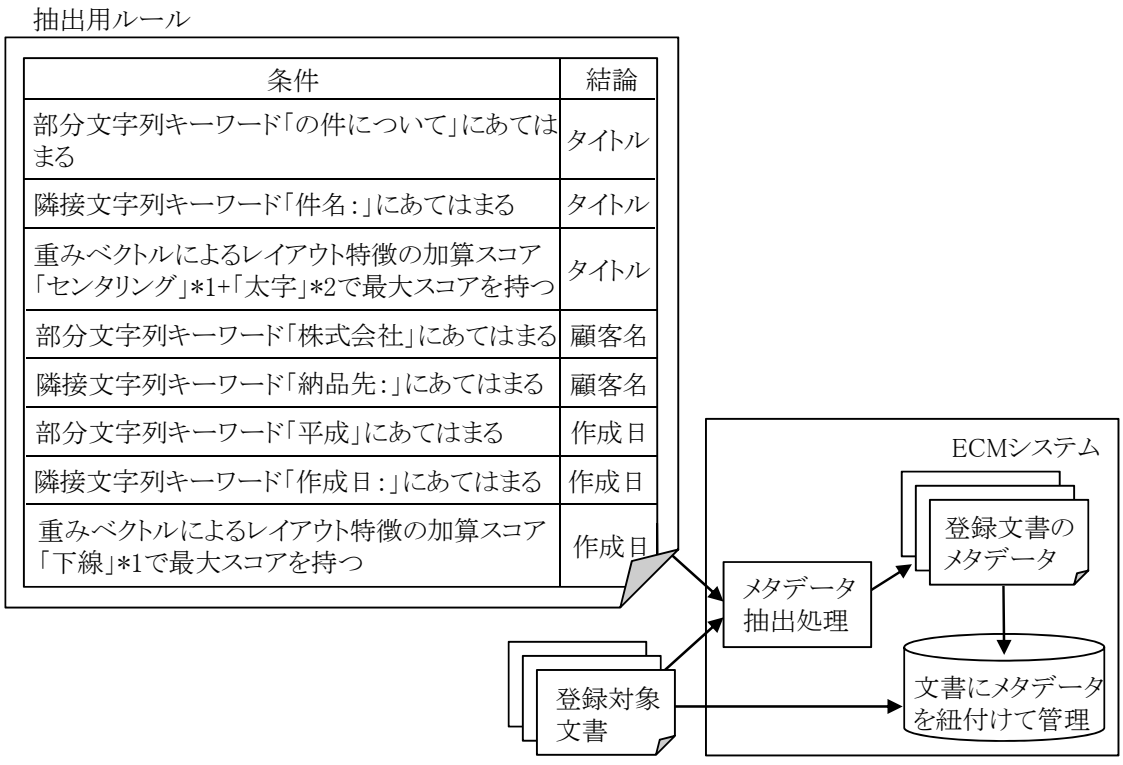



図 3.1　ECM システムおよびメタデータ抽出の概要

類，数量および，文書の管理に用いるメタデータの種類を検討し，メタデータ抽出用ルールを設定する必要がある。システムの稼働開始後は，メタデータ抽出処理により抽出用ルールに基づいて登録対象となる文書からメタデータが抽出され，文書とメタデータが紐付けられて管理される。本章では，メタデータ抽出技術を搭載した先駆的な ECM システムである，コンテンツ運用支援ソリューション MEANS[日立ソフト 2010]におけるメタデータ抽出技術を対象とする。

メタデータ抽出用ルールとしては，タイトルや顧客名などのメタデータの種類ごとに，メタデータに含まれる部分文字列（以下，部分文字列キーワード），メタデータに隣接して記載される文字列（以下，隣接文字列キーワード），どのようなレイアウト上の特徴（以下，レイアウト特徴）がメタデータにおいてどのような頻度で指定されるかを示す重みベクトルの三種類の情報が設定される。このうち，部分文字列キーワードおよび隣接文字列キーワードとしては，任意の文字列を任意の数だけ指定できる。また，レイアウト特徴は，ビジネス文書において頻繁に使用される，表 3.1 に例示するような約五十種類のレイアウトである。レイアウト特徴の重みベクトルは，各レイアウト特徴に 0 以上の整数を割り当てて構成される。

メタデータの抽出では，タイトルや作成日などの種類ごとに，部分文字列キーワードとして設定された文字列を含む文字列，隣接文字列キーワードとして設定された文字列の前後に記載された文字列，レイアウト特徴の有無を重み加算したスコアが文書中で最大となった文字列が抽出される。この中でレイアウト特徴のスコアによる文字列の抽出は，下記の手順で行う。まず，文書中に含まれる全ての文字列に対し，それぞれのレイアウト特徴を持つかどうかを調べ，レイアウト特徴を持つならば 1，持たないならば 0 を割り当てる。この値を重みベクトルを用いて加算し，文字列のスコアとする。

具体的なメタデータ抽出処理について，図 3.1 に示す抽出用ルールおよび図 3.2 に示すビジネス文書を例として説明する。まず，部分文字列キーワードに関し，作成日の抽出用に設定された「平成」が文書の右上に記載されている文字列「平成 22 年 2 月 16 日」に含まれる。そこで，この文字列が作成日であるとして抽出される。次に，隣接文字列キーワードに関し，顧客名の抽出用に設定された「納品先：」が文書下部の案件の詳細説明部分の一項目に記載されている。そこで，この文字列に隣接して記載されている「AB 法人アイウエオ」が顧客名であるとして抽出される。レイアウト特徴の重みベクトルに関しては，タイトル抽出用に設定されたセンタリング指定および太字指定の重みに従い，文書の左上に記載された「営業管理部担当者様」はセンタリング指定されていないが太字指定されているためスコアとして 2 を，文書の中央上部に記載された「案件報告書」はセンタリング

表 3.1　レイアウト特徴の例

| No. | レイアウト特徴 |
|---|---|
| 1 | 太字指定されている |
| 2 | センタリング指定されている |
| 3 | 下線を持つ |
| 4 | 囲み枠を持つ |
| 5 | 上下に隣接文字列がない |
| 6 | 上下に間隔が空いている |
| 7 | フォントが他と異なっている |
| 8 | 文字色が他と異なっている |
| 9 | 背景色が他と異なっている |
| 10 | 左下隅に近い |
| 11 | 右下隅に近い |
| 12 | 左上隅に近い |
| 13 | 右上隅に近い |
| 14 | フォントの高さが小さい |
| 15 | フォントの高さが大きい |
| 16 | フォントサイズが大きい |
| 17 | ページの上半分にある |
| 18 | フォントがページ内で最小サイズである |

と太字が共に指定されているためスコアとして 3 を，文書の中央に記載された「記」はセンタリング指定されているが太字指定されていないためスコアとして 1 を持つ。その他の文字列のスコアは 0 である。「案件報告書」が最大のスコアを持つため，タイトルとして抽出される。作成日については全ての文字列のスコアが 0 になるため，レイアウト特徴の重みベクトルに基づく抽出は行われない。

### 3.2.2 メタデータ抽出用ルールの設定における課題

上記のようなメタデータ抽出を実現するためには，抽出用ルールが適切に設定されている必要がある。ECM システムに登録された文書を探す際の検索洩れを防ぐため，メタデータ抽出では再現率が重視されるが，高い再現率を達成する抽出用ルールを設定するため

**営業管理部担当者 様**

案件管理番号:23-0401
平成22年2月16日
第一営業部

## 案件報告書

いつもお世話になっております。
下記の案件について注文書を受領しましたので添付してお送りします。
手続きいただきたくよろしくお願いいたします。

記

納品先:　AB法人アイウエオ
品名:　営業文書管理システムの開発および運用
担当部署: 第二開発部
見積番号: H23-1B-1846

**図 3.2　ビジネス文書の例**

には以下の課題がある。メタデータ抽出は，メタデータの種類間で競合することに注意する必要がある。すなわち，文書中に記載された単一の文字列は，たかだか一種類のメタデータとしてしか抽出されないため，複数の種類のメタデータの抽出用ルールにあてはまると競合が起きる。例えばタイトルに対して不必要な部分文字列キーワード・隣接文字列キーワードを定義すると，本来顧客名として抽出するべき文字列までタイトルだとみなして抽出するために，顧客名の再現率を悪化させる危険性が生じる。このため，部分文字列キーワード・隣接文字列キーワードは必要十分なものを設定しなくてはならない。さらに，再現率の高いメタデータ抽出のためには抽出用ルールは二種類の条件を満たすものを設定する必要がある。第一の条件は，キーワードの副作用が生じないものであることである。例えば「様」を隣接文字列キーワードとして設定することにより顧客名(文書の宛先)を抽出することができると期待されるが,図3.3に示すような文書はタイトルの4文字目に「様」という文字を含むため，誤ってタイトルの 1〜3 文字目である「要求仕」を顧客名として抽出してしまう危険性がある。従って，キーワードの生成に際しては，このような誤抽出を起こさないかを確認する必要がある。第二の条件は，レイアウト特徴の重みベクトルによる副作用が生じないルールを生成することである。例えば，タイトルは下線を持つことがあるため，直感的には,「下線を持つ」というレイアウト特徴に大きい重みを設定するこ

注文No.123-4568

日立ソフトウェアエンジニアリング株式会社 御中

2010年2月16日
〒111-1111
東京都千代田区大手町1-2-3
株式会社ABCD製作所

# 要求仕様

## 1. はじめに

本文書は、当社が導入を検討している新規業務システムについて基本となる考え方および仕様について記載したものである。

## 2. システム導入の方針

### 2．1 新規業務システムの導入目的

当社の業容拡大に伴い、企画・研究開発・設計・営業・マネージメントなどの非定型業務で扱われるデータ量が増大し、これらのデータから業務遂行上重要な箇所

**図 3.3 抽出用ルールが不適切である場合メタデータを抽出できないビジネス文書の例**

とがタイトルの抽出に有効なのではないかと期待される。しかし，実際には，図 3.3 に示すように注文書番号のような文書 ID(Identifier)や価格が下線を持つ確率の方が高い。このため，「下線を持つ」というレイアウト特徴に過度な重みを設定すると，文書 ID や価格をタイトルとして誤抽出することにつながってしまう。従って，重みベクトルの設定に際しては「抽出したい文字列で高頻度に現れており，かつ抽出したくない文字列では現れていないか」を確認する必要がある。

部分文字列キーワードおよび隣接文字列キーワードは，任意の文字列を設定でき，また，個数の制限もない。また，レイアウト特徴としては約五十種類のものがあるため，高い再現率を与える組み合わせおよび重みベクトルを設定する工数は膨大なものとなる。これらの理由により，抽出用ルールを人手で設定するには数時間以上を要する。さらに，抽出用ルールの設定を行うためには，メタデータ抽出技術の原理を理解している必要があるが，習得には数日以上を要する。

## 3.3 メタデータ抽出用ルール生成アルゴリズム

### 3.3.1 サンプル提示によるメタデータ抽出システム

上記で述べた課題を解決するため，本章ではサンプル文書からメタデータ抽出用ルールを生成する SCA 法を提案する。SCA 法によってサンプル文書からメタデータ抽出用ルールを生成し，そのルールを用いてメタデータを抽出する ECM システムの概要を図 3.4 に示す。新規の登録対象文書への ECM システムの導入時には，左上の破線で囲んだ部分に示すように，サンプル文書と正解メタデータの組を入力として SCA 法により抽出用ルールを生成する。ECM システム稼動後は，右下の破線で囲んだ部分に示すように，このルールを用いて登録対象文書からメタデータを抽出し，文書とメタデータを合わせて登録する。

このシステムを実現するためには，下記の三つの要求がある。

- 生成された抽出用ルールを用いて，人手で設定したルールと同等の再現率でメタデータを抽出できなくてはならない。
- 誤抽出が発生しないような抽出用ルールを生成する必要がある。
- 実用的な時間で計算を行うために，効率的に抽出用ルールを生成しなくてはならない。

なお，SCA 法へ入力するサンプル文書については，従来から企業での利用が広まっている帳票の認識技術[Fujio2001]と同様に，ECM システムの稼動後に登録される文書の種類

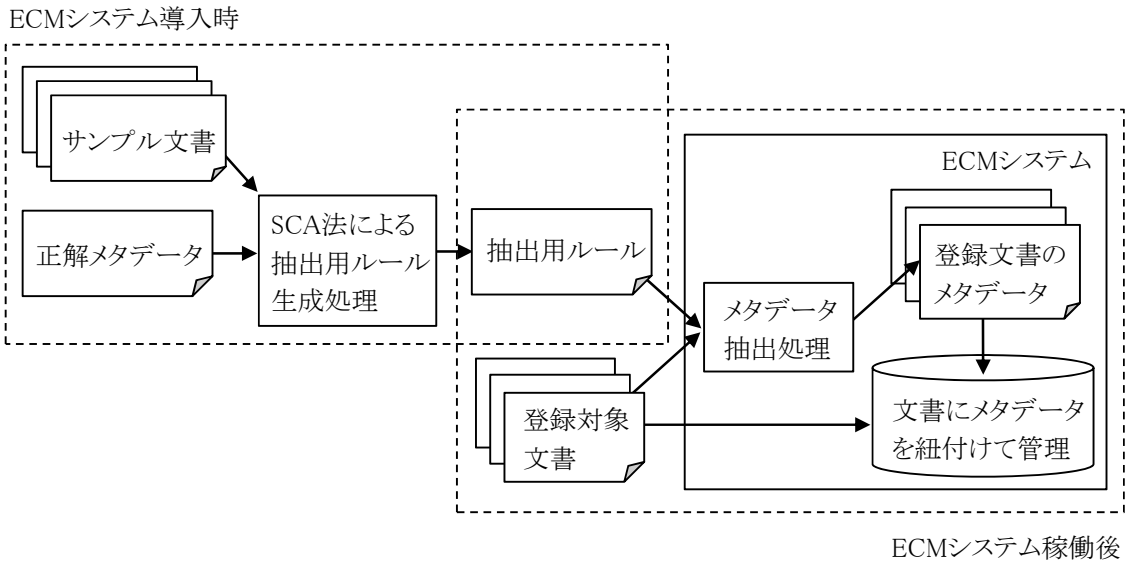



図 3.4　サンプル提示によりメタデータを抽出する ECM システム

あたり複数枚の文書を集める。例えば，対象が営業文書であれば何件かの取引について引合から入金までに発生した文書を，対象が設計文書であれば何件かの製品について基本設計からテストまでに発生した文書を揃える。

### 3.3.2 SCA 法の概要

ビジネス文書の記載上の三種類の傾向に基づき，SCA 法の処理方式を定める。

(1) 固定的文字列の利用

ビジネス文書では語彙が統制されており，例えば顧客名であれば「株式会社」や「(株)」のような文字列が多く用いられる。このため，サンプル文書のメタデータにおいて頻繁に用いられる文字列は，部分文字列キーワードとして有効である可能性が高い。

(2) ID 番号に対する名称の併記

図 3.2 の右上に記載された案件管理番号のように，ビジネス文書では重要な情報が読み取りやすいように，ID 番号の隣には名称が記載される場合が多い。そこで，サンプル文書のメタデータに隣接して記載される文字列は，隣接文字列キーワードとしての効果が期待できる。

(3) 文書の中で重要な文字列のレイアウト

タイトルのように文書の中で特に重要な文字列は，読み取りやすさの観点から，図 3.2 の例のように大きいフォントサイズ・太字・センタリングなど，他の文字列と比べて視認性の高いレイアウトが指定される場合が多い。したがって，サンプル文書のメタデータがどのようなレイアウト特徴を持つか調べることで，重みベクトルとしてどのレイアウト特徴を重視するべきか情報が得られると考えられる。

SCA 法の概要は図 3.5 に示すものであり，メタデータ抽出技術の処理内容を踏まえ，三つの特徴を持つ。

- メタデータの記載上の特徴を洩れなく集めるため，サンプル文書において指定された正解メタデータの出現から抽出用ルール生成のための情報を集めることである。正解メタデータおよびその周辺文字列から，部分文字列キーワードおよび隣接文字列キーワードの候補を集める。また，サンプル文書において正解メタデータで特異的に観察されたレイアウト特徴を候補として利用する。ビジネス文書における三種類の傾向により，サンプル文書における正解メタデータを調べることで，記載上の特徴を効率よく集めることができると考えられる。メタデータの記載上の特徴を洩れなく集めることで，メタデータ抽出用ルールの生成における第一の要求である，

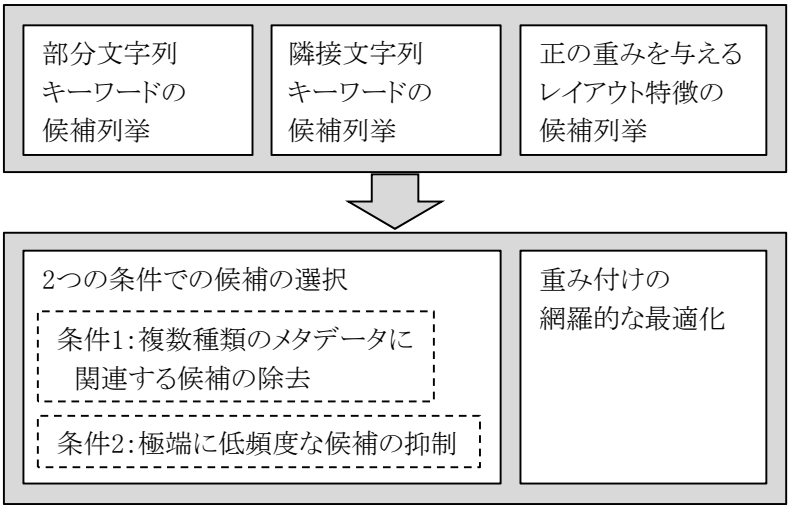



**図 3.5　SCA 法の概要**

高い再現率の達成を狙う。

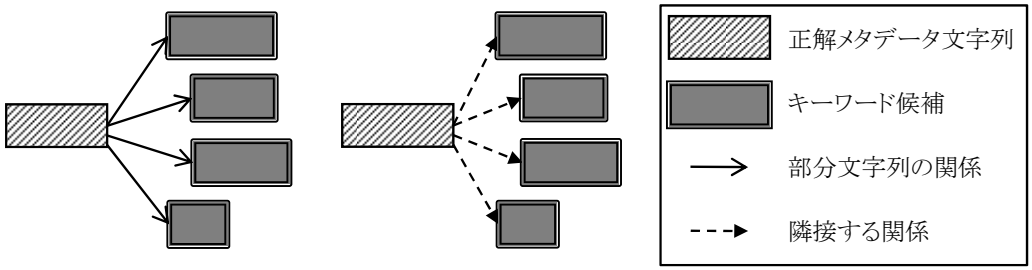
- メタデータの誤抽出の防止である。必要性の低い抽出用ルールにより誤った抽出が行われることを防ぐため，サンプル文書に含まれる正解メタデータ以外の文字列を用いて，部分文字列キーワード・隣接文字列キーワードおよびレイアウト特徴がメタデータに特有のものであるかを確認する。これにより，メタデータ抽出用ルールの生成における第二の要求である，誤抽出の防止を実現する。
- 第三の特徴は，まず候補を列挙してから，キーワード候補の選択およびレイアウト特徴の重みの最適化を行うという二段階方式の採用である。これにより，メタデータ抽出用ルールの生成における第三の要求である，効率的な生成を実現する。

### 3.3.3 部分文字列キーワードおよび隣接文字列キーワードの生成

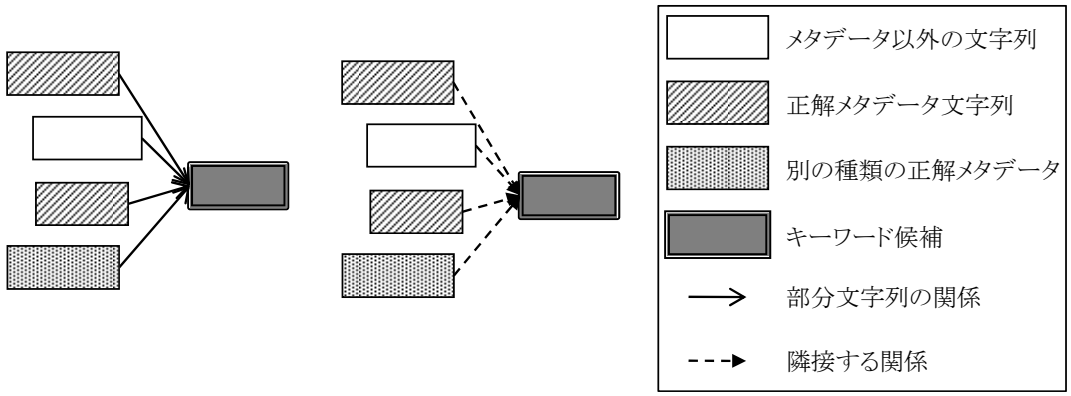
サンプル文書におけるメタデータの記載ごとに部分文字列キーワードの候補および隣接文字列キーワードの候補を列挙する処理では，図 3.6 中の(A)に示すように，単一の正解メタデータ文字列から複数の候補を列挙する。列挙の手順を図 3.7 中の(A)に示す。正解メタデータ全体または部分文字列を，部分文字列キーワードの候補として列挙する。また，正解メタデータに隣接する文字列を，隣接文字列キーワードの候補として列挙する。

隣接文字列キーワードおよび部分文字列キーワードの候補に対し，各サンプル文書での記載を確認して二種類の条件で選択を行い，矛盾した候補やメタデータ抽出に効果が見込まれない候補を除く。それぞれの条件の詳細を以下に説明する。

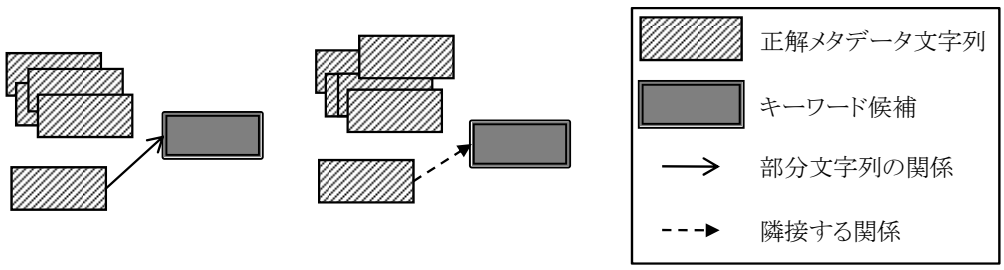
第一の条件は，キーワード候補を実際に抽出用ルールとして用いた場合，不必要な文字




(A)　正解メタデータ文字列に対し，全ての部分文字列および隣接して記載する文字列を部分文字列キーワードおよび隣接文字列キーワードの候補として列挙する。




(B-1)　不必要な文字列が抽出されたならば，その候補は削除される（第一の条件）




(B-2)　十分な数の正解メタデータがあるにも関わらず一度しか登録されなかった候補は，低頻度過ぎるとして除く（第二の条件）

図 3.6　キーワード候補の列挙および選択

```
for (メタデータの種類 km ) {
  for (kmについての正解メタデータ) {
    全ての部分文字列をkmの部分文字列キーワードの候補として挙げる
    全ての隣接する文字列をkmの隣接文字列キーワードの候補として挙げる
  }
}
```

(A)　キーワード候補の列挙

```
for (メタデータの種類 km ) {
  for (kmの部分文字列キーワードの候補) {
    if (候補がサンプル文書において種類km'の正解メタデータに含まれる AND km ≠ km') 候補から除く
    if (候補が種類kmの正解メタデータに含まれる回数
      ＞ 候補がメタデータ以外の文字列に含まれる回数) 候補から除く
  }
  for (kmの隣接文字列キーワードの候補) {
    if (候補がサンプル文書において種類km'の正解メタデータの隣に記載される
      AND km ≠ km') 候補から除く
    if (候補が種類kmの正解メタデータの隣に記載される回数
      ＞ 候補がメタデータ以外の文字列の隣に記載される回数) 候補から除く
  }
}
```

(B-1)　キーワード候補に対する選択（第一の条件）

```
閾値lを設定する
for (メタデータの種類 km ) {
  if (種類kmに対してl以上の正解メタデータが指定されている) {
    for (種類kmの部分文字列キーワードまたは隣接文字列キーワードの候補) {
      if (候補が種類kmの正解メタデータに1回だけ含まれるまたは隣接して記載される) 候補から除く
    }
  }
}
```

(B-2)　キーワード候補に対する選択（第二の条件）

図 3.7　キーワードの列挙および選択アルゴリズム

列を誤って抽出しないか確認するためのものである。図 3.6 中の(B-1)に示すように，単一の候補に対し，その候補を包含または隣接する複数の文字列を用いて確認を行う。確認の手順を図 3.7 中の(B-1)に示す。部分文字列キーワードの候補を包含する文字列について，SCA 法では二種類の確認を行う。

- 包含する文字列が意図しない種類の正解メタデータとして指定されたものではないかの確認を行う。すなわち，その候補を列挙する元になったものと異なる種類の正解メタデータに包含されている場合は，その種類のメタデータ抽出を阻害してしまうので，候補から除く。
- 意図した種類の正解メタデータとして指定された文字列に包含される方が指定されていない文字列に包含される場合よりも多いかどうかの確認を行う。指定されていない文字列に包含される場合の方が多いならば，無関係な文字列を誤って抽出してしまう副作用の方が大きいと考えられるためは，候補から除く。

隣接文字列キーワードの候補についても同様に二種類の確認を行う。

- 隣接して記載される文字列が意図しない種類の正解メタデータとして指定されたものではないかの確認を行う。
- 意図した種類の正解メタデータとして指定された文字列と隣接して記載される方が指定されていない文字列と隣接して記載される場合よりも多いかどうかの確認を行う。

上記の条件により隣接文字列キーワードの候補が除外される動作例を二つ述べる。

例 1　対象データ：顧客名の隣接文字列キーワード
　　理由：タイトルの一部を誤抽出してしまうため
　　キーワード候補から除外するもの：「様」
例 2　対象データ：請求書番号の隣接文字列キーワード
　　理由：注文番号を誤抽出してしまうため
　　キーワード候補から除外するもの：「No.」

まず，図 3.3 を用いて例 1 を説明する。これは，3.2.2 節で述べた「副作用による誤抽出

を発生させるキーワード」をこの条件で除外できる例である。図 3.3 に示す文書を含むサンプル文書から，文字列「様」が顧客名の隣接文字列キーワードとして列挙されたとする。図 3.3 では「様」に隣接する文字列「要求仕」はタイトルの一部であることから，意図した種類の正解メタデータとして指定された文字列ではない。これにより，「様」を顧客名の隣接文字列キーワードの候補から除外することができる。

次に図 3.3 および図 3.8 を用いて例 2 を説明する。これは，「意図とは別の種類のメタデータの隣接文字列キーワードとしても作用してしまうキーワード」をこの条件で除外できる例である。図 3.8 の左上に記載されている「AB1-234567890」が請求書番号である場合，「No.」という文字列が隣接文字列キーワードの候補として挙げられる。しかし，図 3.3 では「No.」に隣接する文字列は注文番号であり，意図した種類の正解メタデータとして指定された文字列ではない。これにより，「No.」を請求書番号の隣接文字列キーワードの候補から除外し，注文番号を請求書番号として誤抽出するのを防ぐことができる。

第二の条件は，極端に低頻度なキーワード候補の抑制である。図 3.6 中の(B-2)に示すように，候補を包含または隣接する文字列の数を用いて確認を行う。確認の手順は図 3.7 中の(B-2)に示す通り，正解メタデータが指定されているサンプル文書が $l$ 以上あるような種類のメタデータにおいては，一度しか登録されなかった候補は低頻度過ぎるとして除く。サンプル文書数が多い場合は，全ての種類のメタデータにおいて正解メタデータが指定さ

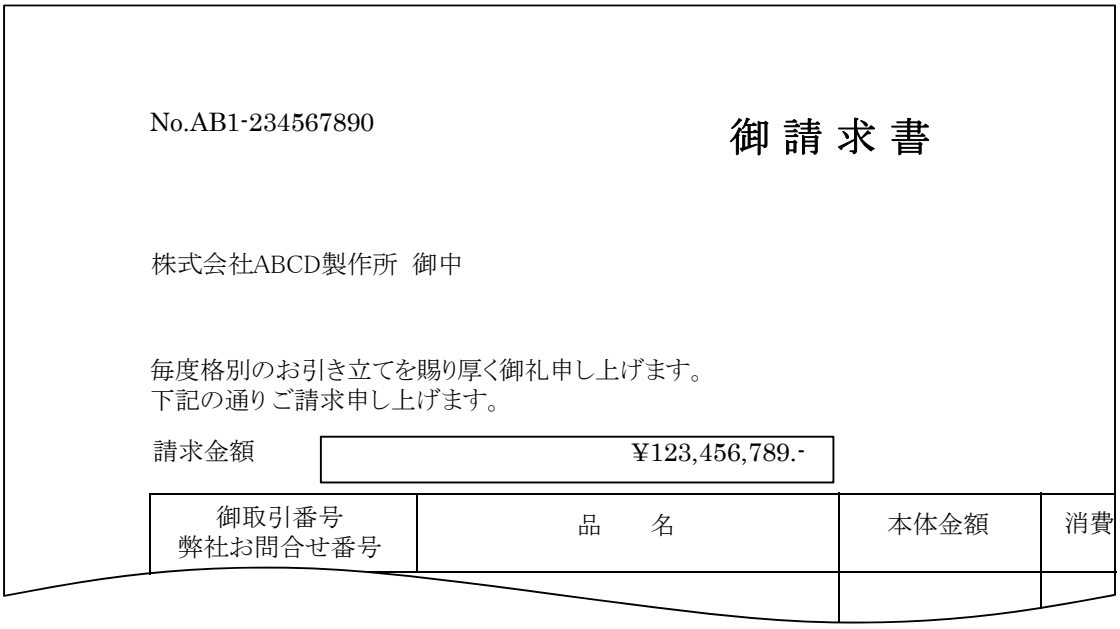



**図 3.8 抽出用ルールが不適切である場合メタデータを抽出できないビジネス文書の例**

れているサンプル文書の数が $l$ 以上となり，$l$ は抽出用ルールの生成結果に影響を与えなくなる。本章においては $l = 5$ とする。

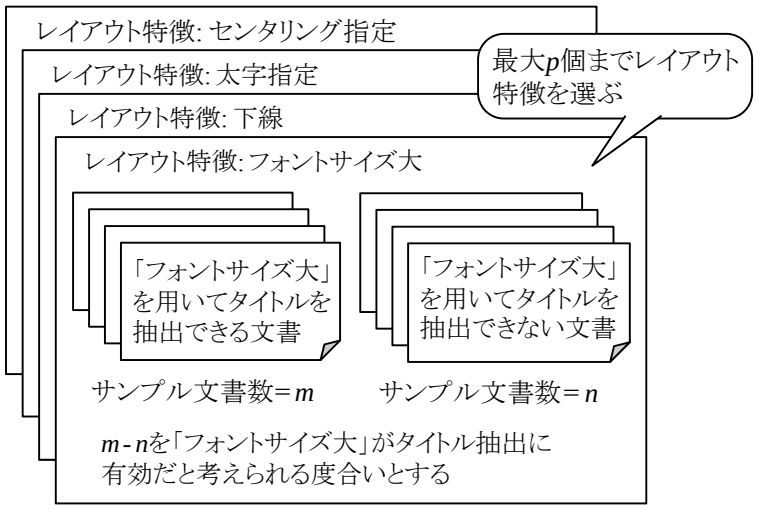
### 3.3.4 レイアウト特徴の重みベクトルの生成

レイアウト特徴の重みベクトルについては，メタデータの種類ごとに，表 3.1 に例示したレイアウト特徴のどれをどの重みで用いるかを設定する。有望と思われるレイアウト特徴についてのみ重点的に重みパラメータを設定するため，二段階の処理を行う。まず，それぞれのレイアウト特徴に着目して，以下に述べる「その種類のメタデータを抽出するために有効だと考えられる度合い」を計算し，この度合いが大きい順に最大 $p$ 個まで候補として選ぶ。次に，レイアウト特徴の候補に対して重みを網羅的に計算し，最も多くのサンプル文書でメタデータを正しく抽出できる重みの組み合わせを得る。

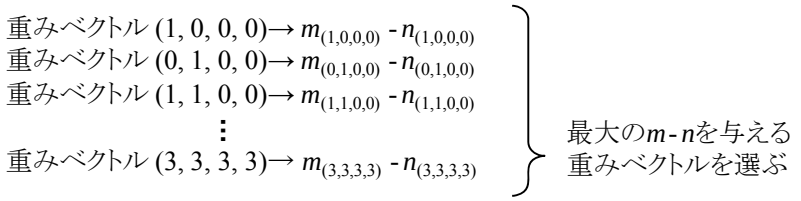
レイアウト特徴に対するメタデータを抽出するために有効だと考えられる度合いの評価について，図 3.9 中の(A)を用いて説明する。レイアウト特徴の利用にあたっては，3.2.2 節で述べた「正解メタデータ以外の文字列でも多く設定されるレイアウトに過度な重みを設定してしまうことによる誤抽出」の発生を考慮する必要がある。そこで，正解メタデータだけがレイアウト特徴を持ち，その他の文字列はそのレイアウト特徴を持たないサンプル文書の数を $m$ ，正解メタデータはレイアウト特徴を持たず，他の文字列の中にそのレイアウト特徴を持つものがあるサンプル文書の数を $n$ として，“$m-n$”をそのレイアウト特徴がメタデータ抽出のために有効だと考えられる度合いとする。例として，図 3.2 に示す文書がサンプル文書として与えられ，タイトルの正解メタデータとして「案件報告書」という文字列が指定されていたとする。表 3.1 の No.16 に示す「フォントサイズが大きい」というレイアウト特徴に着目すると，図 3.2 に示すサンプル文書においては正解メタデータだけがこのレイアウト特徴を持ち，他の文字列は持たない。このような条件を満たすサンプル文書の数が「フォントサイズが大きい」というレイアウト特徴に対する $m$ の値となる。$n$ についても同様に求める。

上記の手順で全てのレイアウト特徴に対してメタデータの抽出に関して有効だと考えられる度合“$m-n$”を計算する。この度合が大きい順に，$m>0$ である範囲で，最大 $p$ 個までレイアウト特徴を選んでメタデータの抽出に関しての候補とする。帳票の設計基準[JIS-Z8303]でタイトルの記載にあたり考慮すべきと指定されているレイアウト特徴の数が 4 であることから，$p=4$ とする。

次に，レイアウト特徴の候補について，重み付けの調整を行いメタデータ抽出に最も有効な組み合わせを求める処理を，図 3.9 中の(B)を用いて説明する。正解メタデータにおけ




(A)　各レイアウト特徴に対して，メタデータの抽出に関して
有効だと考えられる度合の計算




(B)　レイアウト特徴の候補について重み付けの調整を行い
メタデータ抽出に最も有効な組み合わせを求める処理

**図 3.9　レイアウト特徴の重みベクトルの列挙および最適化**

るレイアウト特徴の有無を重み加算したスコアがそのサンプル文書中の他の文字列におけるスコアより大きい文書数を $m$，正解メタデータ文字列におけるスコアよりもそのサンプル文書中の他の文字列の方がスコアが大きい文書数を $n$ として，“$m-n$”を目的関数とする。各レイアウト特徴の候補に対して 0 から 3 の範囲で網羅的に重みを変えながら，目的関数“$m-n$”を最大化する。最適な組み合わせを用いた場合に $m-n>0$ が成り立つならば，その組み合わせを採用する。そうでなければ，レイアウト特徴のスコアを用いて抽出することそのものが不適切であると判断する。

## 3.4 実験結果

3.3 節で述べた抽出用ルール生成アルゴリズムの有効性を評価するため，二種類のビジネス文書を用いて，アルゴリズムで自動生成したルールと，人手で設定したルールを比較する実験を行った。人手で設定したルールは，メタデータ抽出アルゴリズムの開発において用いられたものを使用した。実験には Core 2 Duo (2GHz) CPU および 3GB メモリの PC (Personal Computer)を利用した。実験の手順は以下の通りである。ビジネス文書を二組に分け，最初の一組を導入時作業で用いるサンプル文書として，もう一組を ECM システム稼動後に登録する文書として扱った。サンプル文書とそれに対応する正解メタデータから，各メタデータについての部分文字列キーワード・隣接文字列キーワード・レイアウト特徴の重みベクトルを抽出用ルールとして生成した。次に，登録文書からメタデータを抽出し，メタデータの種類ごとに再現率を求めた。この操作を，分け方を変えながら繰り返し，平均値および標準偏差を求めた。

一種類目のビジネス文書は，6 つの異なる顧客とのビジネス案件において，引合から入金までの業務の過程で作成された営業文書である。各案件の文書数は表 3.2 に示す通りである。同じ種類の文書であっても作成した顧客が異なれば形式は異なり，また，同じ内容でも顧客によって名称が異なっていることもある。さらに，顧客によって受発注プロセスが異なることから，一部の案件でのみ作成される文書も含まれる。

6 つの案件のうち $d_{teacher}$ 案件における営業文書を導入時作業におけるサンプル文書として扱った。$d_{teacher} = 2, 3, 4$ の場合について，SCA 法で生成した抽出用ルールおよび，人手で設定したルールを用いてメタデータ抽出を行った場合の再現率を図 3.10 に示す。作成日を除く全てのメタデータについて，$d_{teacher}$ が大きいほど再現率が高く，$d_{teacher} \geq 3$ の場合は

**表 3.2　評価に用いた営業文書**

| ビジネス案件番号 | 営業文書の数 |
|---|---|
| 1 | 12 |
| 2 | 27 |
| 3 | 13 |
| 4 | 15 |
| 5 | 35 |
| 6 | 29 |

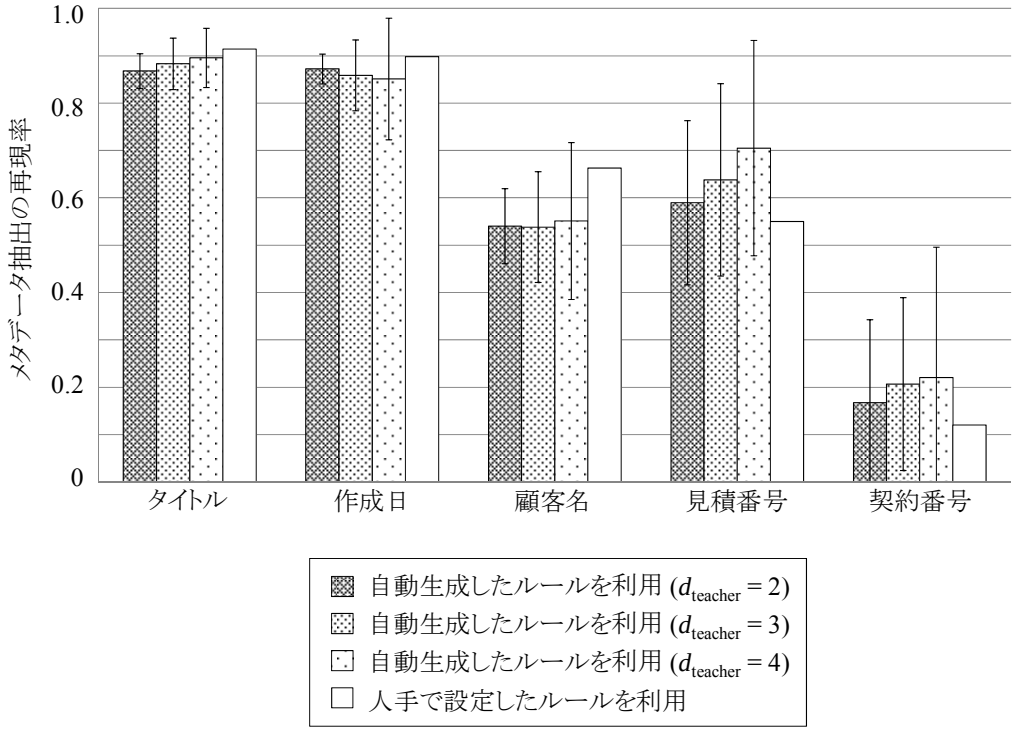



**図 3.10　自動生成または人手で設定したルールを用いた**
**営業文書からのメタデータ抽出の再現率**

人手で設定した抽出用ルールにおける再現率の値が平均±標準偏差の範囲に含まれていた。また $d_{teacher}$ = 2, 3, 4 すべての場合において，SCA 法で生成したルールを用いた抽出と人手で設定したルールを用いた抽出とにおける再現率は，有意水準 10%において有意差は見られなかった。また，抽出用ルールの生成に要した時間は平均 28 秒であった。なお，同じキーワードが文書によって異なる意味で使われるケース（例えば「No.」という隣接文字列キーワードは見積書において見積番号を注文書においては注文番号を示すことなど）があるため，本章で対象とするメタデータ抽出技術においては，人手で設定したルールを用いてもメタデータの抽出を完全に正しく行うことはできない。

二種類目のビジネス文書は，5 つの研究プロジェクトにおける週次作業報告書である。各プロジェクトの文書数は表 3.3 に示す通りである。報告書の利用範囲は部署内に限られるため形式の統一は図られていない。プロジェクトメンバーによって，報告内容が文章で記載されたりリスト形式で記載されたりする。また，同一の作業内容に対し異なる用語が用いられる場合もある。プロジェクトの数が少ないことから，プロジェクト間のメタデータの記載のされ方の多様性を含んだサンプル文書を用いるため $d_{teacher}$ = 3 の場合のみにつ

表 3.3　評価に用いた週次作業報告書

| プロジェクト番号 | 週次作業報告書の数 |
|---|---|
| 1 | 20 |
| 2 | 20 |
| 3 | 30 |
| 4 | 20 |
| 5 | 20 |

いて， SCA 法で生成した抽出用ルールおよび，人手で設定したルールを用いてメタデータ抽出を行った場合の再現率を図 3.11 に示す。タイトルおよび有給休暇行使については人手で設定した抽出用ルールにおける再現率の値が平均±標準偏差の範囲に含まれており有意水準 10%において有意差は見られなかったが，打合せについては有意差が見られた。また，抽出用ルールの生成に要した時間は平均 38 秒であった。

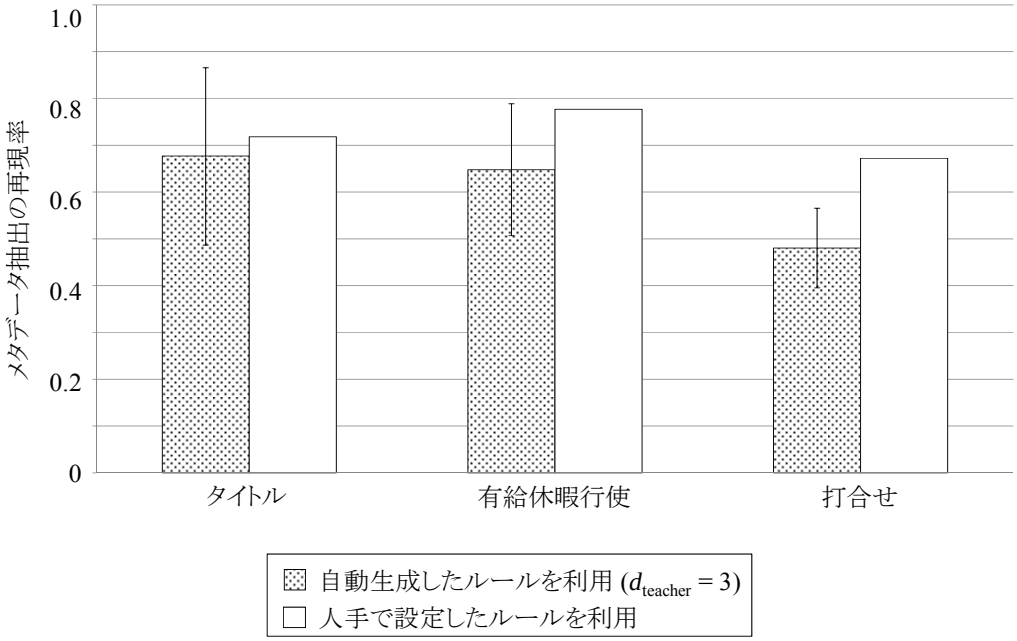



図 3.11　自動生成または人手で設定したルールを用いた
週次作業報告書からのメタデータ抽出の再現率

## 3.5 実験結果の安定性の評価

3.4 節で述べた通り，週次作業報告書における打合せの場合を除き，SCA 法で自動生成したルールを用いた場合と人手で設定したルールを用いた場合とで，メタデータ抽出の再現率には有意差は見られなかった。従って，3.3.1 節で述べた第一の要求「人手で設定した抽出用ルールと同等の精度でメタデータを抽出できなくてはならない」を多くの場合で達成できた。この結果が安定的に得られるのか評価した結果を下記に示す。

### 3.5.1 サンプル文書が再現率に与える影響

図 3.10 に示す通り，抽出用ルール生成に用いる案件の数 $d_{teacher}$ が大きいと多くの種類のメタデータで抽出の再現率が高いという結果が得られた。このことには二つの原因が考えられる。第一に，多くの文書をサンプルファイルとして用いれば，部分文字列キーワードや隣接文字列キーワードをより洩れなく挙げることができるようになる。一件のサンプル文書では一種類のキーワードが現れ，キーワードの頻度が Zipf の法則[Zipf1932]に従う，すなわち，i 番目に高頻度なキーワードの生起確率が $\frac{1/i}{\sum_j 1/j}$ であると仮定した場合に全てのキーワードをサンプル文書から収集できる確率を図 3.12 に示す。例えば，メタデータ抽出のためにキーワードを五つ設定する必要がある場合でも，サンプル文書数が 50 程度あればほとんどの場合(90%以上の確率)で洩れのないキーワード登録が可能となる。表 3.2 に示す営業文書では，$d_{teacher} \geq 3$ の場合はただ一つの組み合わせを除いてサンプル文書数が 50 以上になり，人手で設定した抽出用ルールを用いた場合とほぼ同等の再現率を達成できたことを裏付けている。

第二の原因として，より多くの案件で発生した文書を用いることにより，サンプル文書の多様性が増加する。例えば営業文書において $d_{teacher}$ の値が小さすぎる場合，案件による用語の偏りが顧客名の再現率を悪化させている例があった。顧客企業名を抽出するためには「御中」および「殿」を隣接文字列キーワードとして利用することが特に有効である。しかし案件 1，案件 2，案件 6 には，「殿」を記載している営業文書が含まれないため，サンプル文書がこれらの案件の文書のみから成る場合は「殿」を隣接文字列キーワードとして生成することができなかった。また，週次作業報告書において自動生成ルールによる打合せの抽出の再現率が低かったことも，用語の偏りが原因であった。2 番目のプロジェクトでは「レビュー」，4 番目のプロジェクトでは「定例」という用語で打合せについて報告

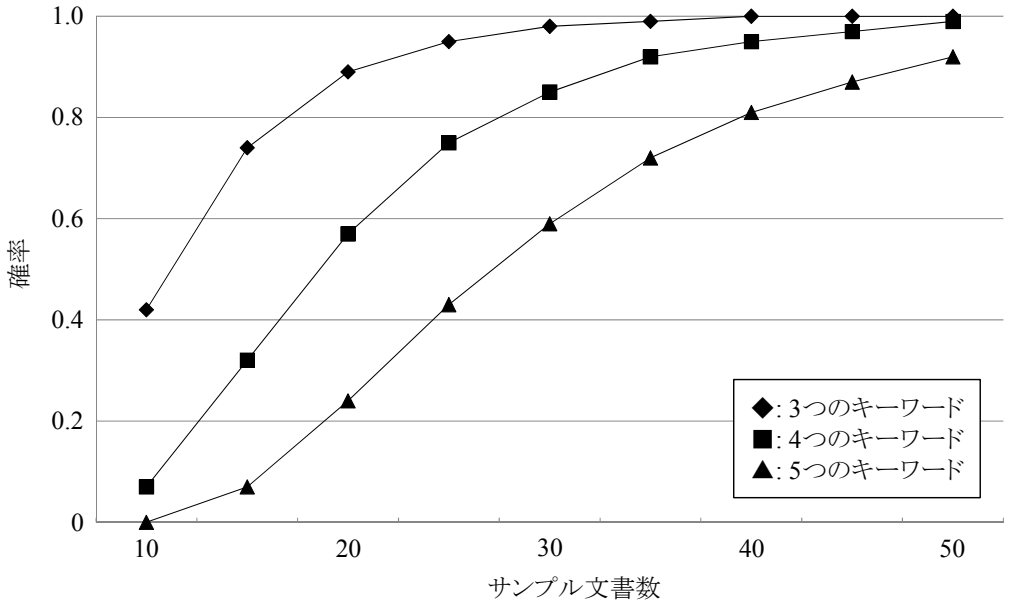



**図 3.12　サンプル文書から全てのキーワードを収集する確率**

する場合があり，他のプロジェクトの週次作業報告書からルールを生成するとこれらの用語を利用した抽出ができなかった。

### 3.5.2 抽出用ルールの調整による効果

正解メタデータ以外の文字列を用いた調整を行った場合と行わなかった場合の，営業文書からのメタデータ抽出の再現率の違いを図 3.13 に示す。作成日および顧客名においては文字列の選択と重み付けの調整を行うことにより，少しずつ再現率が改善した。また，タイトル，作成日，顧客名，および見積番号では，文字列の選択により再現率が大きく改善し，有意水準 5%で有意差が見られた。ただし，契約番号では文字列の選択を行わない方が再現率が高く，人手で設定した抽出用ルールにおける値を上回ることもあった。この原因は，契約番号は業務システムによって割当てられており，固定的文字列部分が長かったためであった。部署や年度によってこの文字列が変わる可能性もあるため，本来は「契約No.」などの隣接文字列キーワードから抽出する方が安定的な抽出を実現できると考えられる。

また，調整を行う前後の部分文字列キーワードおよび隣接文字列キーワードの数を表 3.4 および表 3.5 にそれぞれ示す。調整を行うことで，隣接文字列キーワードは多くのメタデータで数個程度まで絞り込まれた。人間が抽出用ルールを設定する場合も，同程度の数の文字列を挙げると考えられる。大幅に候補数が削減された原因は，3.3.3 節の第二の条

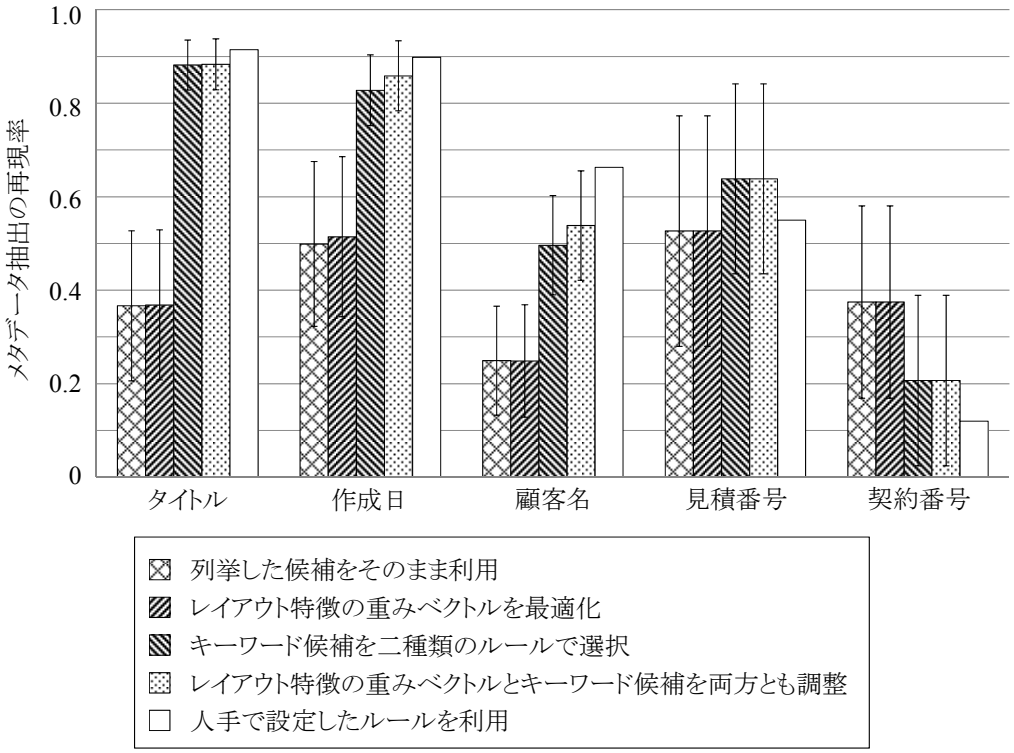



**図 3.13　抽出用ルールの調整を行わない場合と行う場合における営業文書からのメタデータ抽出の再現率**

**表 3.4　営業文書における調整前後の部分文字列キーワードの数**

| | タイトル | 作成日 | 顧客名 | 見積番号 | 契約番号 |
|---|---|---|---|---|---|
| 調整前 | 3,567.9 | 490.0 | 719.3 | 107.2 | 115.5 |
| 調整後 | 543.9 | 154.2 | 358.5 | 21.9 | 114.3 |

**表 3.5　営業文書における調整前後の隣接文字列キーワードの数**

| | タイトル | 作成日 | 顧客名 | 見積番号 | 契約番号 |
|---|---|---|---|---|---|
| 調整前 | 185.1 | 28.8 | 249.7 | 33.5 | 29.7 |
| 調整後 | 21.0 | 4.5 | 7.4 | 6.3 | 5.9 |

件により，サンプル文書において偶然メタデータに隣接して記載されていた多数の文字列が，一度しか登録されなかったキーワード候補として除かれたためである。図 3.13 においてキーワードの調整によりメタデータ抽出の再現率が低下しなかったことにも表れている

ように，除かれたキーワード候補は低頻度過ぎるため他の文書からメタデータを抽出するために有効に働くことはなかった。

部分文字列キーワードも調整により大きく絞り込まれたが，依然として数十以上の文字列が抽出用ルールとして用いられた。多数の文字列が残った原因は，サンプル文書における正解メタデータの数および文字列長に応じて候補数が増えることおよび，正解メタデータ以外の文字列にも現れる候補文字列は長さの短いものに限られるためであった。ただし，過剰な部分文字列が抽出用ルールに挙がっていても利用されないだけであり，メタデータ抽出の再現率には影響しない。特に，ECM システムの管理対象文書として紙文書のスキャン画像を OCR(Optical Character Recognition)で文字認識したものを含める場合は，文字誤認識の混入に備えて様々な長さの部分文字列キーワードを抽出用ルールとして挙げておくことが有効である。

部分文字列キーワードおよび隣接文字列キーワードの調整における，正解メタデータが指定されているサンプル文書数の閾値 $l$ を変えた場合の，メタデータ抽出の再現率を図 3.14 に示す。$l$ の値を変えても，メタデータ抽出にはほとんど影響を与えなかった。また，

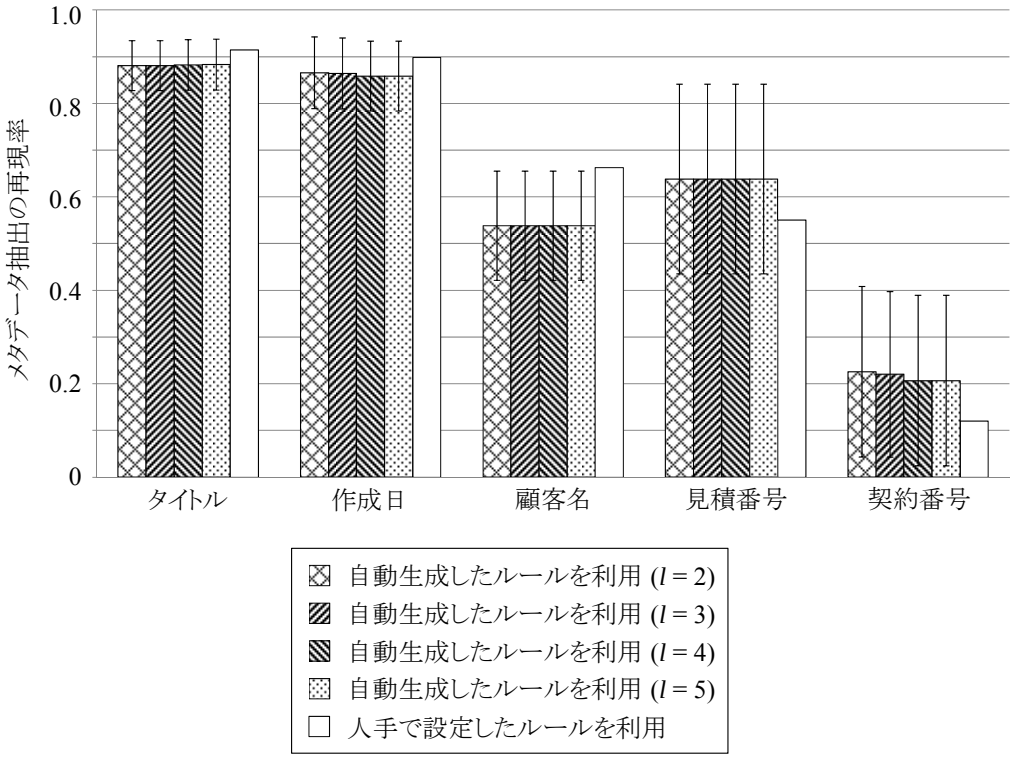



**図 3.14　正解メタデータが指定されたサンプル文書数の閾値を変えた場合の営業文書からのメタデータ抽出の再現率**

調整後の隣接文字列キーワードの数を表 3.6 に示す。契約番号以外のメタデータでは，$l$ の値は隣接キーワード選択に影響を与えなかった。また，部分文字列キーワードの数は，全ての種類のメタデータにおいて $l$ の値によらず一定であった。$d_{teacher} = 3$ の場合でも正解メタデータが指定されたサンプル文書数は十分多かったため，キーワードの調整結果やメタデータ抽出の再現率が $l$ の値の影響をほとんど受けなかったと考えられる。正解メタデータが指定されたサンプル文書数が $l$ と同程度以上に多ければ，3.3.3 節で述べたように，$l$ の値が閾値として意味を持つことはなくなる。このため，文書の種類やメタデータの記載のされ方に依らず，本章での実験と同様にメタデータ抽出の再現率への $l$ の値の影響はないと考えられる。

一方，レイアウト特徴の重みベクトルの調整は，営業文書におけるタイトル，見積番号，および契約番号では二つの原因から精度向上の効果は見られなかった。第一の原因は，上記三種類のメタデータにおいては部分文字列または隣接文字列として固定的な表現が用いられる場合が多く，キーワードが抽出精度に支配的な影響を及ぼしているためであった。第二の原因は，メタデータ抽出に有効なレイアウト特徴が限定的だったことである。見積番号および契約番号は営業文書中の様々な場所に記載されるため，レイアウト特徴が抽出に有効でなかった。タイトルについても，下線のような特異性の低いレイアウト特徴は正解メタデータよりもそれ以外の文字列で設定されていることが多いため有効ではないと判断され，候補として挙げられなかった。また，フォントの高さが大きいこととセンタリング指定されていることが抽出に効果を持つものの，均等な重みで効果を発揮するため，重みベクトルの調整を行っても再現率は改善しなかった。

また，レイアウト特徴の候補数の閾値 $p$ を変えた場合のメタデータ抽出の再現率を図 3.15 に示す。$p = 1$ の場合にタイトルの抽出の再現率がやや低かったものの，$p \geq 2$の場合は全ての種類のメタデータにおいてほとんど抽出の再現率は変わらなかった。この結果は，本章で実験に使用した文書におけるメタデータの記載のされ方を反映したものであり，メ

**表 3.6　営業文書における調整後の隣接文字列キーワードの数**

| $l$ の値 | タイトル | 作成日 | 顧客名 | 見積番号 | 契約番号 |
|---|---|---|---|---|---|
| 2 | 21.0 | 4.5 | 7.4 | 6.3 | 8.1 |
| 3 | 21.0 | 4.5 | 7.4 | 6.3 | 7.4 |
| 4 | 21.0 | 4.5 | 7.4 | 6.3 | 5.9 |
| 5 | 21.0 | 4.5 | 7.4 | 6.3 | 5.9 |

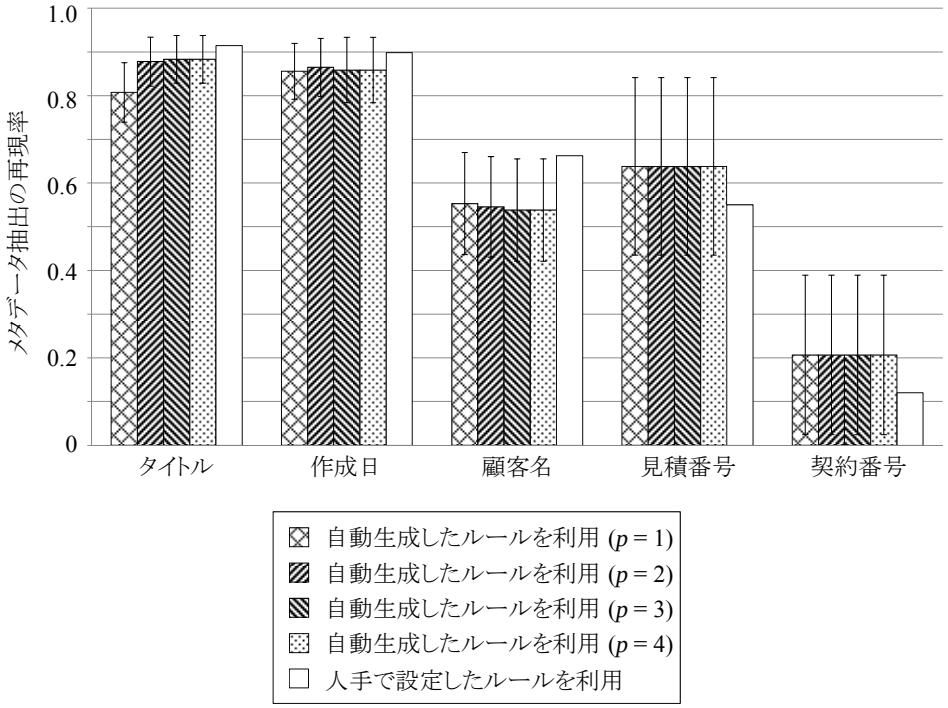



**図 3.15　レイアウト特徴の候補数の閾値を変えた場合の**
**営業文書からのメタデータ抽出の再現率**

タデータに多くのレイアウト特徴が指定される種類の文書では，$p$ が大きいほどメタデータ抽出の再現率が改善される可能性がある。ただし，そのような場合でも，3.3.4 節で述べたように $p = 4$ であれば十分な大きさであると考えられる。

### 3.5.3 自動生成した抽出用ルールによる適合率

SCA 法で生成した抽出用ルールおよび人手で設定した抽出用ルールを用いて営業文書からメタデータ抽出を行った場合の適合率を図 3.16 および図 3.17 に示す。見積番号の抽出において，人手で設定した場合との差が特に大きい。自動生成した抽出用ルールでは，見積番号が記載されていない文書からも何らかの文字列を抽出してしまうことが影響していた。これはメタデータでないものを誤ってメタデータであると抽出してしまう誤りである。$d_{teacher}$ が増加するとタイトルの再現率は改善するが，適合率は低下する。ただし図 3.10 に示した通り，この適合率の低下は他の種類のメタデータ抽出を妨げ再現率を低下させるものではなかった。固定的な表現が用いられる場合が多い見積番号や契約番号では，単調な低下傾向は見られなかった。また，図 3.17 に示す通り，レイアウト特徴の重みベクトルおよびキーワードの調整により，図 3.13 に示す再現率の場合と同様の改善が見られた。ま

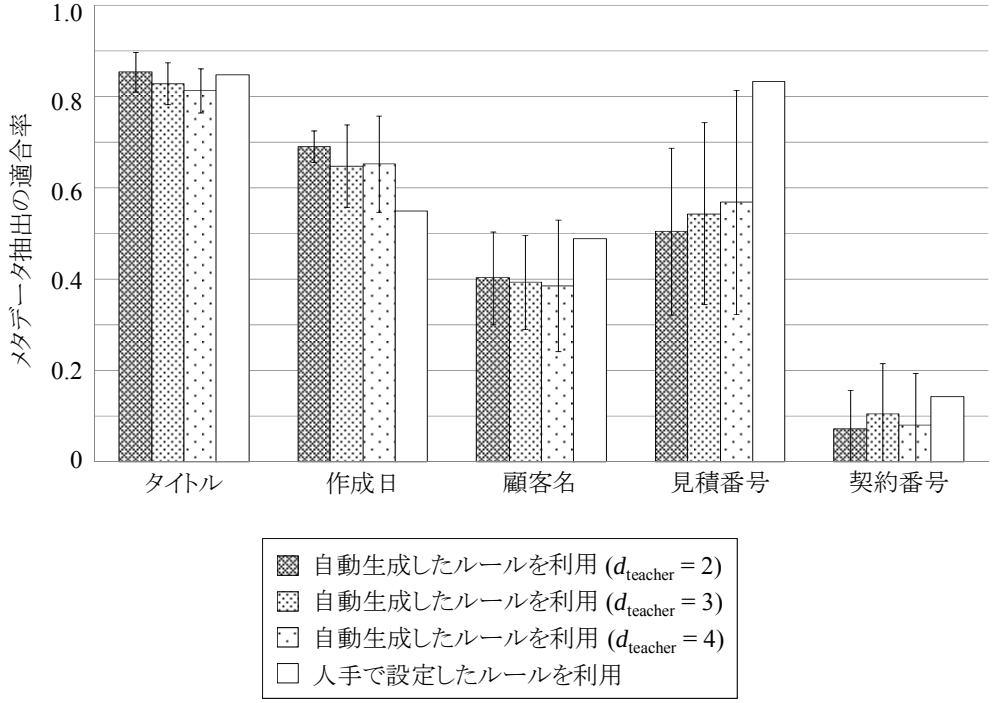



**図 3.16　自動生成または人手で設定したルールを用いた**
**営業文書からのメタデータ抽出の適合率**

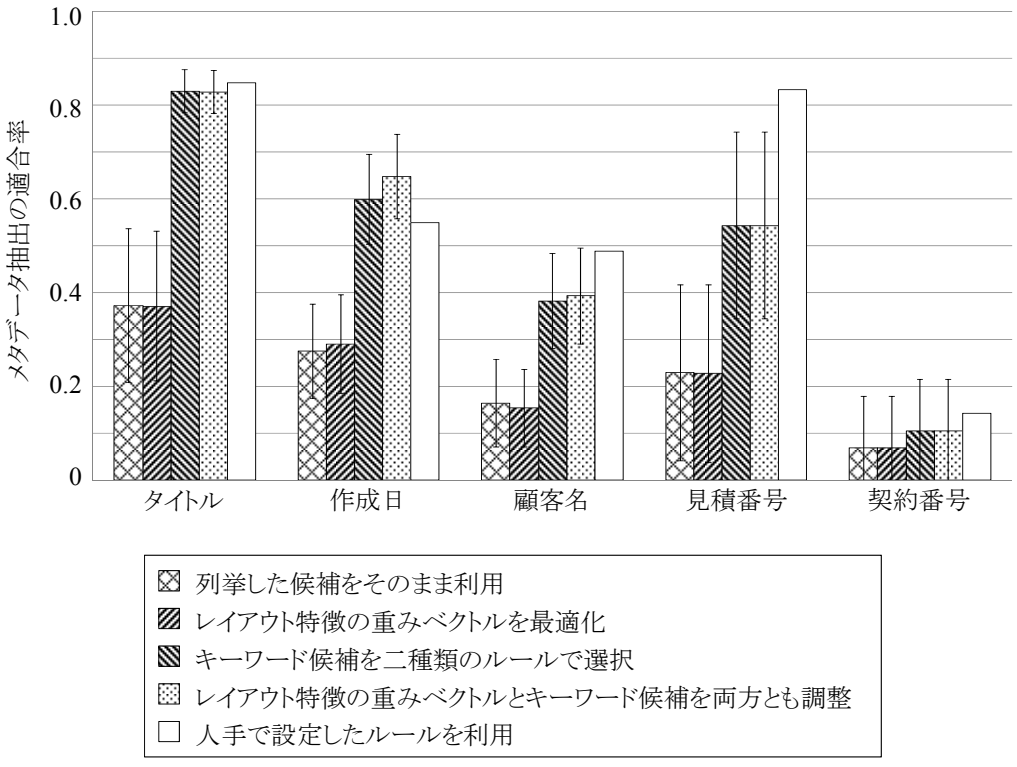



**図 3.17　抽出用ルールの調整を行わない場合と行う場合における**
**営業文書からのメタデータ抽出の適合率**

た週次作業報告書における適合率では，表 3.7 に示す通り有給休暇行使において，人手で設定した場合との差が大きい。これは，文章中から該当する文字列を抽出しようとして，前後の文字列も一緒に抽出してしまう場合が影響していた。これらの適合率の低下は，他の種類のメタデータ抽出を妨げ再現率を低下させるものではない。従って，3.3.1 節で述べた第二の要求「誤抽出が発生しないような抽出用ルールを生成する」を概ね満たしている。

### 3.5.4 抽出用ルールの自動生成の適用範囲

SCA 法では営業文書に対しては平均 28 秒，週次作業報告書に対しては 38 秒での生成が可能であり，3.2.2 節で述べた人手でのルール設定と比較して大幅な効率化を実現するとともに，3.3.1 節で述べた第三の要求「実用的な時間で計算を行えるために，効率的に抽出用ルールを生成しなくてはならない」を達成できた。同じ営業文書であっても業種や取引形態が異なれば作成する種類や記載項目は大きく異なることが多い。また，顧客名などの取引情報やプロジェクトの作業内容は通常は社外秘として指定される。これらのことから ECM システムの導入案件ごとにサンプル文書および正解メタデータを準備する必要があると考えられるが，正解メタデータの指定はファイルあたり 1 分程度で行うことができるため，3.5.1 節で述べたように 50 件のサンプル文書を用意する場合，全体の作業を 1 時間で終えることができる。3.2.2 節で述べた通り従来は数日以上を要していたことから，SCA 法により人手での設定と比べて大幅な省力化を実現できる。

また，SCA 法に基づくメタデータ抽出の適用性について，営業文書と週次作業報告書を二つの観点から比較する。第一に，週次作業報告書では営業文書の場合と異なり，メタデータの名称とその内容が並べて記載されることはほとんどない。このため，週次作業報告書においては隣接文字列キーワードは用いられないが，SCA 法は営業文書の場合と同様に有効であることが示された。第二に，週次作業報告書では「レビュー(1/16)」や「12/07 打合せ」のように作業内容と実施日が並んで記載されることが多い。SCA 法が対象とするメタデータ抽出方式では作業内容と実施日を別々に抽出することができないため，いったん

**表 3.7　自動生成または人手で設定したルールを用いた週次作業報告書におけるメタデータ抽出の適合率**

| | タイトル | 有給休暇行使 | 打合せ |
|---|---|---|---|
| 自動生成したルール | 0.99 | 0.64 | 0.15 |
| 人手で設定したルール | 1.00 | 0.84 | 0.65 |

まとめて抽出してから正規表現などで日付部分を別途切り分ける必要がある。

SCA 法は 3.3.2 節で述べたビジネス文書に共通的な特徴を用いてメタデータ抽出用ルールを生成することで，メタデータの記載上の特徴を効率よく利用することを可能としている。このようなアプローチは，機械学習分野における強化学習[木村 1999]，過学習の防止[銅谷 2005]および，属性選択[Yang1997]とは異なるものである。

## 3.6 結言

メタデータ抽出のためのルールを自動的に生成する SCA 法を提案した。SCA 法では，メタデータ検索を行う際の洩れ防止を目的として，再現率を重視した抽出用ルール生成を行う。サンプル文書における正解メタデータの記載に基づいて候補を列挙し，その後に詳細な調整を行うという二段階の方式により，人手による抽出用ルールと同等の再現率の達成・誤抽出の抑制・計算時間の抑制による効率的な生成という三つの課題をいずれも解決することができた。これらの特徴から，SCA 法は 3.5.4 節で述べた通り，従来は数日以上を要していたメタデータ抽出用ルールの設定を 1 時間程度で行うことができ，大幅な導入工数削減が実現できた。SCA 法は（株）日立ソリューションズのコンテンツ運用支援ソリューション MEANS 紙文書電子化ソリューションに搭載されており，導入作業の工数削減を通じて ECM システムによる営業文書の登録効率化を実現している[日立ソフト 2010]。

SCA 法によるメタデータ抽出用ルールの生成は，サンプル文書数およびメタデータの種類数に対しおおむね線形の計算時間で実行できるためスケーラビリティが高い。また，SCA 法を利用することにより，企業に ECM システムを導入する際の工数を大幅に削減し，ECM システムによるメタデータ抽出技術の実用性を改善した。これにより，多くの企業で大量のビジネス文書を効率よく管理することが可能となる。

なお，SCA 法では，キーワードとレイアウト特徴の重みベクトルを独立に生成しているが，抽出用ルール全体を相関ルールとして扱って同時に最適化させることによる改良も考えられる。また，SCA 法は英語文書からのメタデータ抽出においても有効であると期待できるが，単語の活用形やハイフンの利用など英語独自の表現への対応を行ったルール生成を行えば，より再現率の高いメタデータ抽出の実現が期待される。さらなる拡張としては，サンプルファイルから正解メタデータを指定する処理の自動化[Yoshinaga2006]による工数削減，シソーラスを用いて部分文字列キーワードや隣接文字列キーワードの候補の同義語をこれらのキーワード候補に加えることによる必要なサンプル文書数の削減および，メタデータ種類に優先度を割り当て重要なメタデータの抽出精度を優先的に向上させるため

の抽出用ルール生成方式の拡張が考えられる。

# 第 4 章
# 業務情報周知のための業務遂行状況に応じた提示要否の判別技術

## 4.1 緒言

通知，規則，連絡事項に代表される企業内で用いられる文書データの周知を実現するためには，業務状況に応じて着目するべき文書を抽出し，提示する技術が有効であると期待される。このため本章では，業務上利用するアプリケーションの表示文字列の例とそれぞれの状況における参照要否を入力として，提示要否の判別条件を構成・修正する方式について提案する。

企業内の文書のデジタル化の進展に伴い，データ量が急速に増大している[Lyman2003] [富士通 2009]。その中でも，多くのデータが関係データベースに蓄積されているのではなく単なるファイルとして存在していることが指摘されている[梅原 2008] [Shilakes1998]。業務を行う上で必要なファイルの迅速な取得，社内に蓄積された成果物の再利用推進による業務効率向上，および法令や社内規則の周知徹底を行うことが益々重要となり，企業内の業務情報活用への効率化ニーズが高まっている。

業務情報を登録し活用を図るナレッジマネジメントシステムや，様々な業務情報をまとめて参照の利便を図るポータルシステムはいろいろと検討されてきた[前田 2006] [Fan2002] [野中 2004] [清水 2004] [中山 1997] [野中 2008] [田村 2005] [野々口 2008] [田中 2004]。これらの手法ではいずれも，実務部門ユーザは，新たな業務情報がナレッジマネジメントシステムやポータルシステムなどに登録されたことを通知された折に，できるだけその内容を確認し，記憶に留める必要がある。実際にその業務情報を利用するのがいつになるのか事前には分からないにも関わらず,「その時」が訪れたらその業務情報を即座に思い出し，適切に業務を遂行することが求められている。この状況のように，将来行うことに関する記憶は「展望的記憶」と呼ばれ[Einstein1990]，疲労や多忙により記憶に失敗してしまう危険性が高まることが報告されている[森田 2000]。実際，ナレッジマネジメ

ントシステムが活用されるかどうかは，業務情報入力のための工数を確保できるかだけでなく，実務部門ユーザが参照するための興味および工数を確保できるかに依存することが指摘されている[白石 2007] [紺野 1998]。もし，必要な業務情報を思い出すことができなかったり，思い出しても処理の煩雑さを嫌って従わなかったりすると，コンプライアンス違反や業務効率の低下などのリスクが生じてしまう。

これに対し，実務部門各ユーザの業務遂行状況を監視し，登録済みの業務情報の利用が必要になった時点でポップアップ形式による注意喚起を行うことができるならば，実務部門ユーザへ負担なく業務情報の周知を実現できる。この実現にむけて，本章では，業務上利用するアプリケーションの表示文字列の特徴を利用して，実務部門ユーザの業務遂行状況に対して業務情報の提示要否を判別する Conditioning Business Information Pop-up（以下，CBIP 法）と呼ぶ方法を提案する。具体的には，メール，ブラウザ，ワープロソフト，グループウェア等のアプリケーションを用いて文書作成，閲覧に関わる業務を遂行しているユーザに対し必要な業務情報を洩れなく提示し，不要な業務情報が提示された場合に「不要である」旨を表明するだけで確実にその後の提示を抑制することを実現する。また, 業務に用いるアプリケーションの表示文字列には過度に高頻度な表現が現れること，および業務内容に特徴的な表現では頻度の上昇が見られることに着目し，提示要否の判別を行うための文字列の特徴を効率よく調べることを目指す。

本章の構成は，以下の通りである。4.2 節では業務遂行状況に応じた業務情報の提示要否判別の要件について述べる。4.3 節では CBIP 法の詳細について述べる。4.4 節では CBIP 法の有効性を評価するための実験について述べ，4.5 節でその結果について評価を行う。4.6 節ではまとめを述べる。

## 4.2 業務遂行状況に応じた提示要否判別の要件

CBIP 法を用いた業務遂行手順の変化について，図 4.1 に示す。従来の業務遂行手順では，管理部門ユーザが登録した通知，規定，適用事例，チェックリストなどの業務情報それぞれについて，業務部門ユーザはあらかじめ閲覧・理解しておき，必要に応じて思い出さなくてはならない。CBIP 法を用いる場合は, 管理部門ユーザが各業務情報の登録時に，その業務情報の参照が必要/不要な業務遂行状況の例を指定する。管理部門ユーザは実務の全容を定義できないために詳細な条件を手動で定義することが困難であることから，業務遂行状況の例から判別基準を生成し登録しておく。ユーザの業務遂行状況として，業務上利用するアプリケーションの表示文字列（以下，業務上表示文字列）を用いる。業務上表

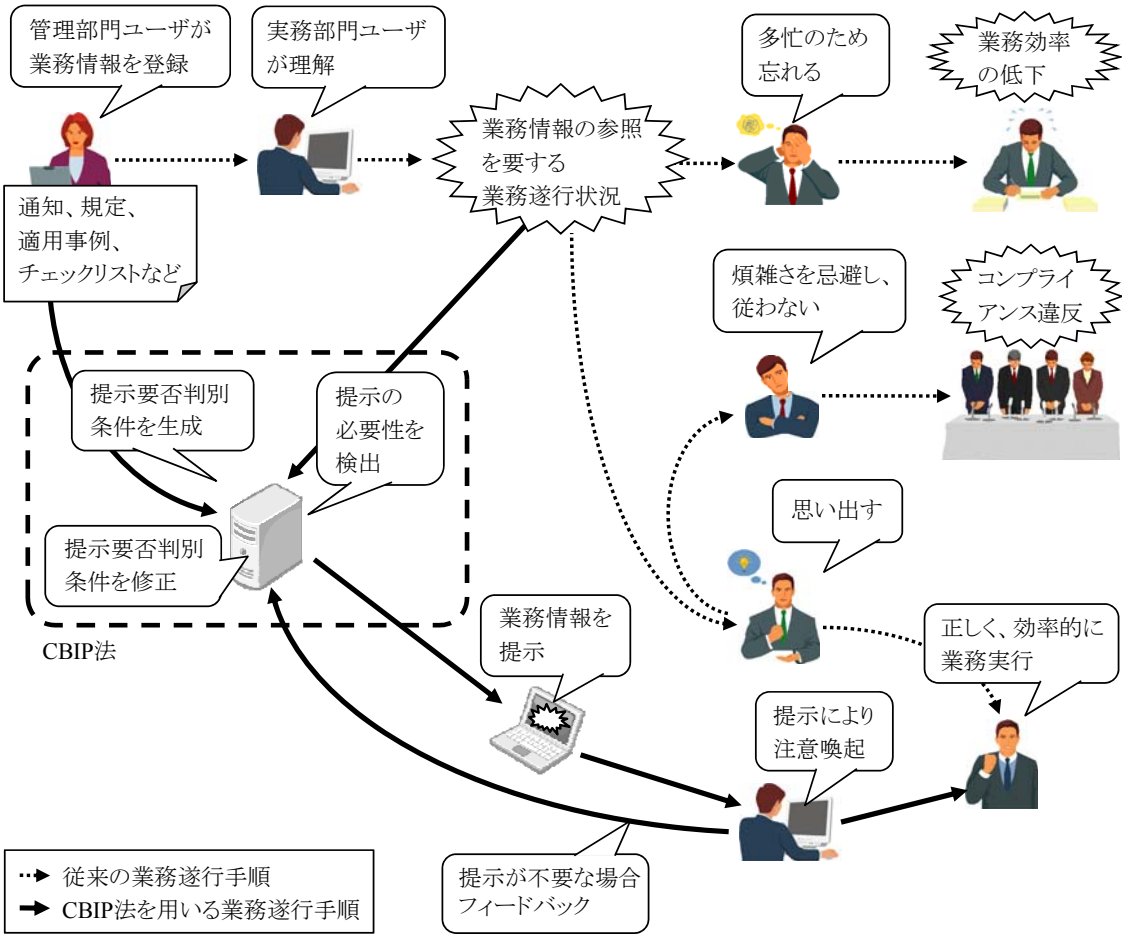



**図 4.1　業務遂行手順の変化**

示文字列に対する提示条件との照合により業務情報の参照の必要性が検出され，業務情報がポップアップ形式で提示されることにより実務部門ユーザの注意を喚起する。また，提示された業務情報に対し実務部門ユーザから不要である旨のフィードバックが与えられた場合は，提示条件を修正する。これにより，業務情報の周知徹底および活用を推進する。

CBIP 法により業務情報の提示が行われる例を図 4.2 に示す。この例では，実務部門ユーザは業務上利用するアプリケーションとしてメール作成ソフトを利用している。アプリケーションには社外の人物に対して打合せのため来社するよう述べている文字列が表示されており，「本社ビル」および「打合せ」を含む。このため提示を要するかどうかの判別条件に合致することから，来館者受け入れに関する規定が提示される。

CBIP 法を実用的に運用するためには，実務部門ユーザが「提示が不要である」旨のフィードバックを行うことができ，かつそのフィードバック提示要否の判別基準に確実に反映されることが必要である。提示される業務情報に実務部門ユーザの注意を向けさせて周

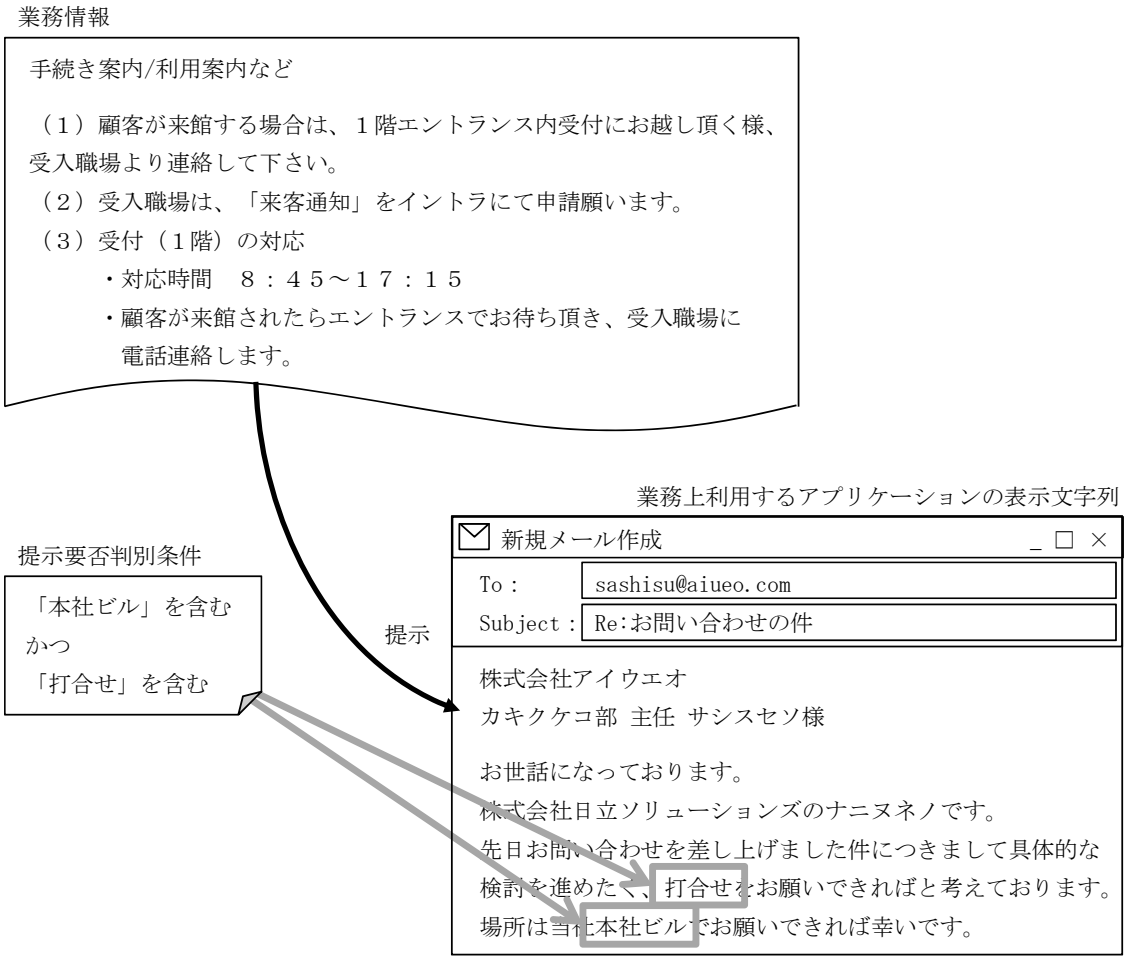



**図 4.2　業務情報の提示の例**

知徹底および活用を実現するためには,「必要としない業務情報がフィードバック後も再度提示されること」を確実に防ぐ手段が必須である。さらに，将来の不要な提示を未然に防ぐと共に必要な提示を維持することも必要である。一方で，実務部門ユーザから「提示されたものは適切である」または「必要な提示が行われなかった」旨のフィードバックが行われることは期待できない。負例に関してのみ直接的なフィードバックが与えられることは業務情報の提示における特徴的な前提である。

## 4.3 業務情報の提示要否の判別方式

### 4.3.1 CBIP 法のアプローチ

CBIP 法では，次節に述べる業務上表示文字列の特徴に基づき，文書集合に特徴的な単語列（以下，言い回し表現）に着目して処理を行う。業務情報の登録時に指定した，その

業務情報の参照が必要/不要である業務上表示文字列の例から言い回し表現を抽出し，判別器を構成する。また，言い回し表現から抑止キーワードを選択することで実務部門ユーザのフィードバックを反映させる。「抑止キーワードを含む業務上表示文字列に対しては提示を行わない」とすることで，実務部門ユーザは「提示が不要である」ことを表明するだけでフィードバックを確実に反映させることができる。

CBIP 法の概要について，図 4.3 に示す。最初に想定される提示条件の生成に当たり，業務情報を提示すべき状況における業務上表示文字列および提示する必要のない状況における業務上表示文字列をそれぞれ正例および負例として用いる。実務部門ユーザからフィードバックが与えられた際の業務上表示文字列と区別するため，それぞれ「初期負例」および「追加負例」と呼ぶ。CBIP 法は三つの主な処理から成る。まず，正例および初期負例から典型文字列の除去を行い，高頻度な形態素列を抽出して文書集合に特徴的な言い回

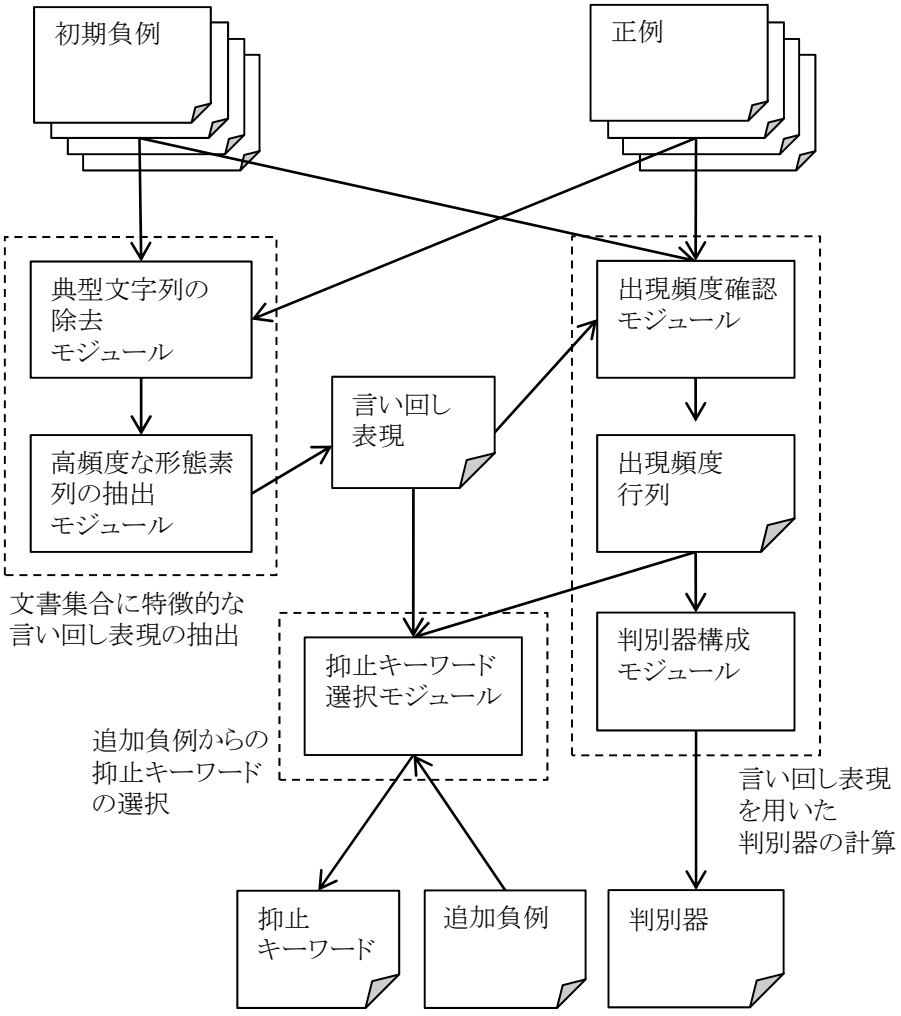



**図 4.3　CBIP 法の概要**

し表現を得る。次に，正例および初期負例における言い回し表現の出現頻度を確認して出現頻度行列を求め，これを元に判別器を構成する。ユーザからフィードバックとして追加負例が与えられた場合に言い回し表現を参照して抑止キーワードを選択する。図 4.3 において破線で囲んだ，これらの処理についてそれぞれ以下に述べる。

### 4.3.2 言い回し表現の抽出

文書分類の分野では，形態素の頻度に基づいた bag of words モデルによる特徴把握が広く行われている[Yang1997] [高村 2001] [高村 2003]。そこで，業務上表示文字列における頻度および形態素の出現の独立性について二種類の予備調査を行った。この結果に基づき，図 4.4 に示すように，典型文字列を避けて言い回し表現を抽出することにより業務上表示文字列の特徴を把握する。

第一の予備調査は，業務上表示文字列に過度に高頻度に現れる表現である。業務上表示

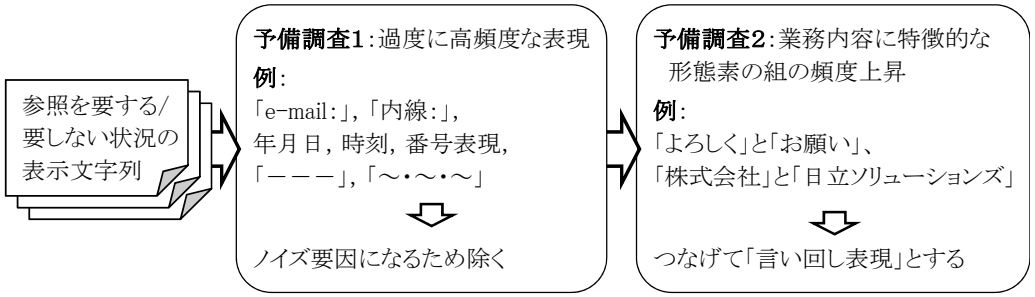



**図 4.4　二種類の予備調査に基づく言い回し表現の抽出**

**表 4.1　業務上利用するアプリケーションの表示文字列における過度に高頻度な文字列**

| 種類 | 例 |
|---|---|
| 電子メールアドレスの項目名 | “e-mail”や“MAIL” |
| 電話番号の項目名 | “TEL” や “内線” |
| 日付 | “2010 年 9 月 21 日(火)” や “2010/09/21” |
| 時刻 | “10:30” や “1 時 30 分～50 分” |
| 項目番号 | “(1)”や “[2]”や“3)” |
| 電子メールのヘッダ | “Subject”や“In-Reply-To” |
| 罫線 | “－－－” や “～・～・～” |

文字列では，内容に関わらず表 4.1 に示すような文字列が高頻度に現れる。これらの文字列の有無は，業務情報を提示するべきかどうかとは関係ないため，ノイズ源になると考えられる。従って，あらかじめ除去することが適切と考えられる。第二の予備調査は，文字列頻度分布である。業務上表示文字列においても，一般文書の場合と同様に，形態素の表層表現の頻度分布は Zipf の法則[Zipf 1932]に従う。すなわち，頻度とその順位の逆数が比例している。しかし業務上表示文字列においては，連続する二つの形態素の頻度分布では「よろしく」と「お願い」の組や「株式会社」と「日立ソリューションズ」の組など，業務内容に特徴的な表現で頻度の上昇が見られる。単純な形態素ではなくこのような特異的な言い回しを利用することにより業務上表示文字列の特徴を捉えやすくなると考えられる。これらの予備調査の結果を踏まえ，図 4.5 に示す通り，下記の手順で言い回し表現の抽出

業務上利用するアプリケーションの文字列(*str*と表す)から
典型文字列を除く　　} 手順 (1)

*str*を*numw*の形態素に分割する(*w*[*numw*]と表す)
要素番号*idx*を1にセットする
While (*idx* ≤ *numw*である間) {
　形態素列*strp*に*w*[*idx*]を設定する
　While (*w*[*idx*]と*w*[*idx*+1]が同じ文に属し、
　　*w*[*idx*]と*w*[*idx*+1]が同じ種類の文字からなり、かつ
　　(*w*[*idx*+1]をつなげた形態素列の頻度)が
　　(*w*[*idx*]までの形態素列の頻度)*0.5以上である間) {
　　形態素列*strp*に*w*[*idx*+1]をつなげる
　　*idx*に1を追加する
　}
　高頻度な形態素列のリスト*strps*に*strp*を追加する
　*idx*に1を追加する
}　　} 手順 (2)

英数字のみから成る形態素列を*strps*から除く
助詞・代名詞のみから成る形態素列を*strps*から除く　　} 手順 (3)

正例または負例で頻度が0.1以上である形態素列を選択する　　} 手順 (4)

**図 4.5　頻出言い回しの抽出方法**

を行う。

（1） 正例および初期負例の文字列から第一の予備調査で挙げた典型文字列を除く。

（2） 形態素に区切り，下記の条件を満たす限り形態素をつなげて言い回し表現とする。

- 文の区切りをまたがない
- 仮名漢字・英数字・記号などの文字種が同じである
- つなぎ合わせても頻度が半分以下にならない

（3） 英数字のみから成るもの・助詞・代名詞などを除く。

（4） 正例または初期負例のいずれかにおいて一割以上の出現率を持つもののみを抽出する。

上記の手順(2)の第一の条件については，文の区切りをまたいでも頻度が下がらない形態素列は，挨拶文の後に現れる「さて」のように，業務内容に特徴的であるとは言えない例が多かったため設定した。第二の条件については，表 4.1 に示した日付，時刻および，項目番号に類する例を除くため設定した。また，第三の条件については，ある形態素に対し最も頻度が高くなる形態素をつなぎ合わせるため，頻度が半分以下にならないことを条件とした。また手順(4)においては，第二の予備調査の結果を元に閾値を設定した。

### 4.3.3 言い回し表現を用いた判別器構成

正例および初期負例の業務上表示文字列を $s_1$, …, $s_p$ および $s_{p+1}$, …, $s_{p+n}$ で表し，4.3.2 節で抽出した言い回し表現を $f_1$, …, $f_m$ で表す。言い回し表現が正例および初期負例の文字列に現れているかどうかを

$\mathrm{Exist}(s_i, f_j) = 1$ ($s_i$ が $f_j$ を含む場合)

$= 0$ ($s_i$ が $f_j$ を含まない場合)

で表し，$\mathrm{Exist}(s_1, f_1)$, …, $\mathrm{Exist}(s_{p+n}, f_m)$を要素として持つ属性行列を作成する。例えば，図 4.2 に示すメールの文字列が正例 $s_1$，言い回し表現 $f_1$ および $f_2$ が「株式会社日立ソリューションズ」および「よろしくお願い」である場合，$s_1$ は $f_1$ を含むが $f_2$ は含まない。このため，属性行列の 1 行 1 列の要素は 1，1 行 2 列の要素は 0 となる。その後，属性行列から決定木を構成する[Hall2009]。

### 4.3.4 抑止キーワードの選択

追加負例として与えられた業務上表示文字列 $t$ に対し，提示が不要な業務上表示文字列を正しく判別できる確率（以下，不要文書非提示率）の改善と再現率の維持のバランスを

図るための抑止キーワードを選択する。将来現れる追加負例に含まれる確率が高い抑止キーワードを選択できれば追加負例を未然に抑止することができ，不要文書非提示率を改善できる。また，再現率の維持のためには，本来提示が必要な文書に合致してしまう可能性の低い抑止キーワードを選ぶことが必要である。追加負例における出現頻度を初期負例における頻度で，将来現れる提示が必要な文書における出現頻度を正例における頻度で近似して考える。すなわち，言い回し表現 $f_i$ に対し正例における出現度数

$$freqP = \sum_{j=1}^{p} \mathrm{Exist}(s_j, f_i)$$

および初期負例における出現度数

$$freqN = \sum_{j=p+1}^{p+n} \mathrm{Exist}(s_j, f_i)$$

をもとに，$freqP = 0$ かつ $\mathrm{Exist}(t, f_i) = 1$ である $f_i$ のうち最大の $freqN$ を持つものを抑止キーワードとする。

## 4.4 実験結果

### 4.4.1 実験に用いた業務情報

表 4.2 に示す三種類の業務情報に関する業務上表示文字列，および表 4.3 に示す一般の業務上表示文字列を用い，CBIP 法の評価を行った。第一の業務情報は，他社製品につい

表 4.2　評価に用いた業務情報

| No. | 種類 | 提示要の業務上表示文字列 | 提示不要の業務上表示文字列 |
|---|---|---|---|
| 1 | ソフトウエア製品に対する窓口部署設置の通知 | ソフトウェア製品「秘文」に対する問合せメール 43 件 | 「秘文」に関する問合せ以外のメール 54 件 |
| 2 | 外部の参加者があるセミナーを社内の会議室で開催する際の事務手続き | 外部の参加者があるセミナーの準備および告知のメール 60 件 | 社外会場または社内参加者のセミナーの告知のメール 60 件 |
| 3 | PC 棚卸の担当者変更の通知 | 棚卸結果の提出メール 69 件 | PC 以外の棚卸に関するメール 71 件 |

表 4.3　一般の業務上表示文字列

| 種類 | 件数 |
|---|---|
| 通達 | 83 |
| 事務処理手続の説明 | 52 |
| 新規ビジネス企画 | 138 |
| 提案書 | 52 |
| 仕様書 | 68 |
| 調査集計資料 | 74 |

て担当部署を設置し，問合せ窓口を一本化したことを通達する文書である。従来は，実務部門ユーザは通達を読んで「窓口部署が設置された」ということを記憶に留めておき，以降の問合せは窓口部署宛に行うことを求められていた。しかし実際には，窓口部署の設置を失念して直接開発元企業に対して問合せを行い，対応の遅れを招いたり有利な契約条件を利用し損ねたりするリスクがある。そこで，ソフトウェア製品「秘文」を例とし，図 4.6 に示すような問合せのメール 43 件と，インストール指示など問合せ以外のメール 54 件および一般の業務上表示文字列を用いて，判別を正しく行えるかを評価した。

第二の業務情報は，外部の参加者があるセミナーを社内の会議室で開催する際の事務手続きについて定めた規定である。従来はセミナーを開催する実務部門ユーザは規定を事前に確認し，適切な手続きを行うことが求められていた。しかし，実際には多くの実務部門にとって外部の参加者があるセミナーを開催する機会はまれであり，事務手続きの規定の

担当者様

仕様に関しての質問です。
秘文AEを利用してますが、遠隔地にある支社へのネットワークは
秘文とは切り離されており、秘文サーバへ接続できません。
WANを介した環境においても秘文を導入することは可能ですか？

以上、よろしくお願いいたします。

図 4.6　ソフトウェア製品「秘文」に対する問合せ

存在そのものを知らないリスクがある。そこで，図 4.7 に示すような外部の参加者があるセミナーを社内の会議室で開催するための準備および告知のメール 60 件と，社外会場で行われるか参加者が社内に限られるセミナーの告知のメール 60 件および一般の業務上表示文字列とを用いて，判別を正しく行えるかを評価した。

第三の業務情報は管理業務の担当者を変更することを通達する文書である。従来は，実務部門ユーザは通達を読んで「担当者が変更された」ということを記憶に留めておき，以降のその管理業務の遂行においては新しい担当者に結果を提出することを求められていた。しかし実際には，前回のメールを再利用して報告を行うことが頻繁に行われ，旧来の担当者宛に提出してしまうリスクがある。また，旧来の担当者は新しい業務を担当して社内の各部署とメール送受信を行う必要があるため，単純に新しい担当者へメールの転送を行うことはできない。そこで，社内で利用している PC の棚卸業務を例とし，図 4.8 に示すような棚卸結果を取りまとめ部門担当者に提出するメール 69 件と，PC 以外の棚卸に関する

平成22年 3月 2日
***コンソーシアム会員各位

***コンソーシアム
＊＊委員会
＊＊＊＊＊＊＊グループ

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊開催のご案内(最終版)

拝啓　時下ますますご清祥のこととお慶び申しあげます。
平素は当会の活動に格別のご協力、ご支援を賜り心より御礼申しあげます。

本メールにていよいよ最終のご案内になります。全日程がほぼ確定しておりますのでご確認ください。年度末のお忙しい時期かと思いますが、ぜひお時間をとってご参加いただければと考えております。

なお、今年度は3月10，11日が＊＊＊＊＊＊＊の＊＊＊＊＊様、16～18日が＊＊の＊＊＊＊・＊＊・＊＊様が会場となっております。

皆様のふるってのご参加をお待ち申し上げております。

==========
テーマ
==========

**図 4.7　外部の参加者があるセミナーの告知**

お世話になっております。
研究部のPC管理担当をしております、松本です。

2009年度PC棚卸について、部内PCの確認を行いましたので
結果を添付して提出いたします。

お手数をおかけしますが、よろしくお願いいたします。

**図 4.8　棚卸し結果の提出**

メール 71 件および一般の業務上表示文字列を用いて,判別を正しく行えるかを評価した。

表 4.3 に挙げた一般の業務上表示文字列の例を図 4.9 から図 4.14 に示す。通達は図 4.9 に例を示す通り，新しい制度や施策および，業務に関連したニュースなどについて広く周知を図るための文書である。事務処理手続の説明は，図 4.10 に例を示す通り，事務処理を行う際の規則と手順について説明したものである。新規ビジネス企画に関わる文書には，図 4.11 に例示するような商品のポジショニングを検討するためのマップやビジネスモデルの検討資料，収支計画などがある。提案書は，図 4.12 に示すようなカタログのほか，プレゼンテーション資料やキャンペーンチラシなどがある。仕様書は図 4.13 に例示するように,製品の機能や前提となる動作環境などを記載した文書である。調査集計資料は,図 4.14 に例示するような顧客ニーズに関する調査状況をまとめた表のほか，競合製品との機能比

全社掲示板　　2009.6.1
総 務 部

2009年度『クールビズ』実施について

題記の件、当社は国民運動『チーム・マイナス６％』のチーム員として地球温暖化対策を推進しております。本年もビル内の冷房時の室温を２８℃にします。

尚、当社におきましては、既に「ノーネクタイ・ノー上着」を通年実施しておりますが、夏場は特に『クールビズ』を意識して実施するよう、ご協力をお願い致します。

記

**図 4.9　通達の例**

較や製品ポートフォリオにおける売上動向などがある。

手続き案内/利用案内など

（1）顧客が来館する場合は、１階エントランス内受付にお越し頂く様、受入職場より連絡して下さい。
（2）受入職場は、「来客通知」をイントラにて申請願います。
（3）受付（１階）の対応
・対応時間　８：４５～１７：１５
・顧客が来館されたらエントランスでお待ち頂き、受入職場に電話連絡します。

**図 4.10　事務処理手続きの例**

## （付録）競合比較

＜競合との差別化要素＞
・＊＊＊＊＊への＊＊＊＊＊＊＊の＊＊を＊＊＊にできる
・＊＊＊＊＊＊＊＊を＊＊＊として＊＊できるため、＊＊を＊＊＊＊＊がない
（＊＊＊の新たな＊＊が発生しない。＊＊＊＊＊＊などの＊＊・＊＊＊＊も＊＊化。）
・他社＊＊＊製品よりも低価格

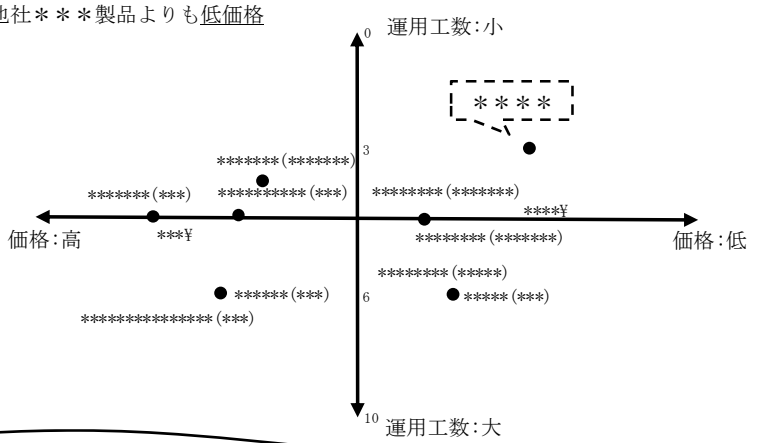



**図 4.11　新規ビジネスの企画の例**


図 4.12　提案書の例

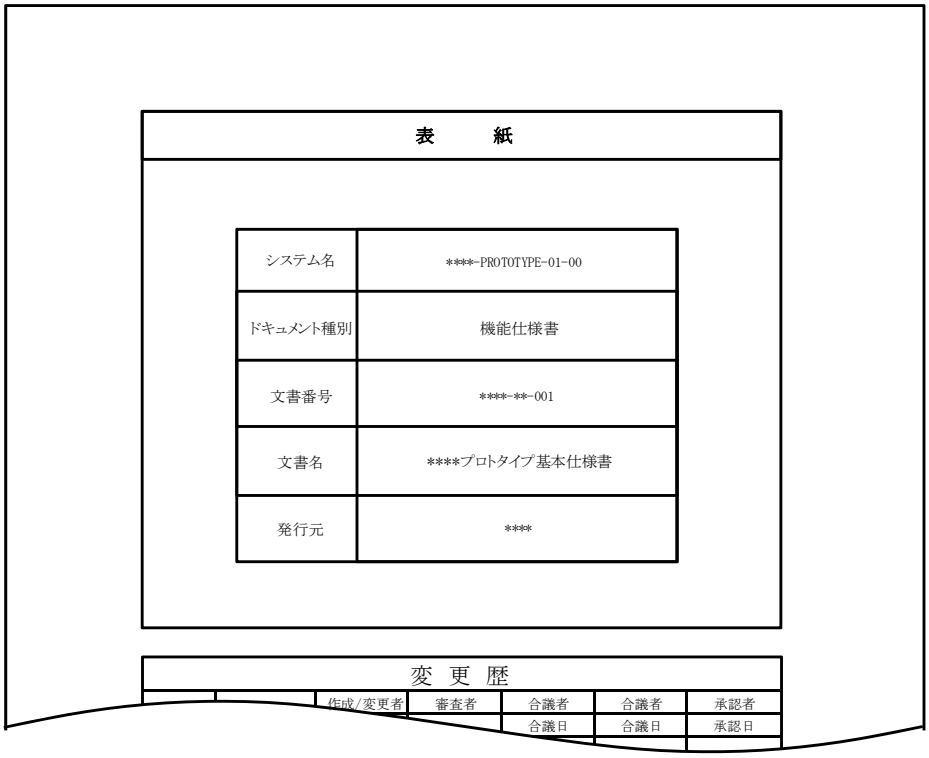
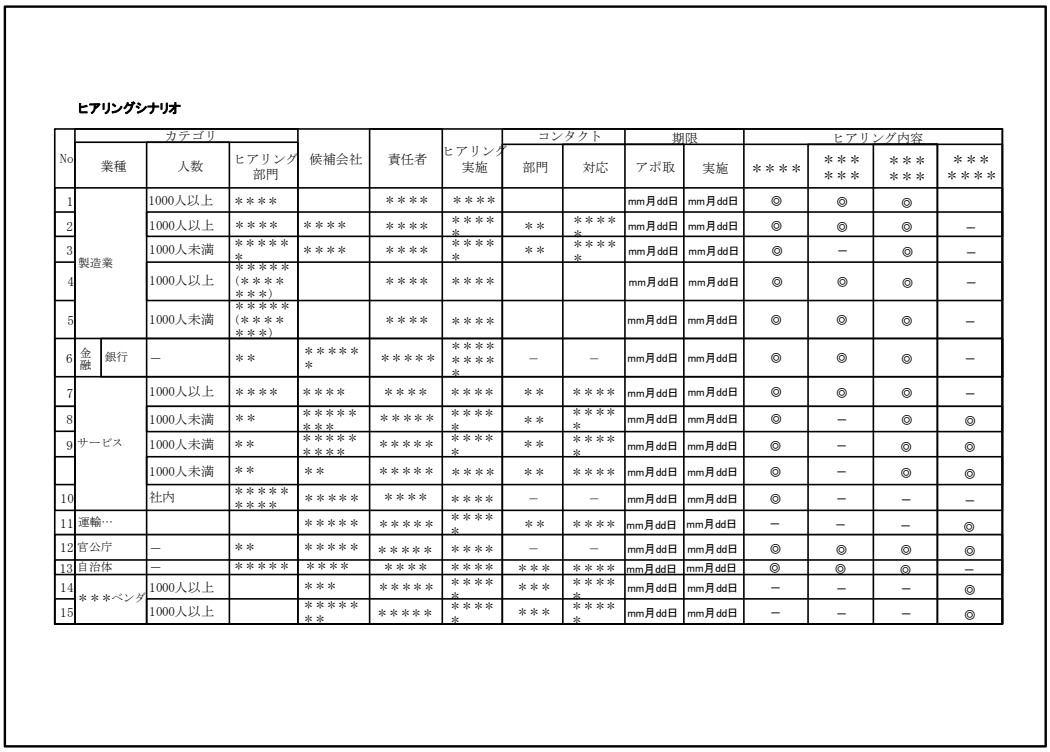



図 4.13　仕様書の例



ヒアリングシナリオ

| No | カテゴリ | | | 候補会社 | 責任者 | ヒアリング実施 | コンタクト | | 期限 | | ヒアリング内容 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 業種 | 人数 | ヒアリング部門 | | | | 部門 | 対応 | アポ取 | 実施 | **** | ***<br>*** | ***<br>*** | ***<br>**** |
| 1 | 製造業 | 1000人以上 | **** | | **** | **** | | | mm月dd日 | mm月dd日 | ◎ | ◎ | ◎ | |
| 2 | | 1000人以上 | **** | **** | **** | ****<br>* | ** | ****<br>* | mm月dd日 | mm月dd日 | ◎ | ◎ | ◎ | – |
| 3 | | 1000人未満 | *****<br>* | **** | **** | ****<br>* | ** | ****<br>* | mm月dd日 | mm月dd日 | ◎ | – | ◎ | – |
| 4 | | 1000人以上 | *****<br>(****<br>***) | | **** | **** | | | mm月dd日 | mm月dd日 | ◎ | ◎ | ◎ | – |
| 5 | | 1000人未満 | *****<br>(****<br>***) | | **** | **** | | | mm月dd日 | mm月dd日 | ◎ | ◎ | ◎ | – |
| 6 | 金融 銀行 | – | ** | *****<br>* | ***** | ****<br>****<br>* | – | – | mm月dd日 | mm月dd日 | ◎ | ◎ | ◎ | – |
| 7 | サービス | 1000人以上 | **** | **** | **** | **** | ** | **** | mm月dd日 | mm月dd日 | ◎ | ◎ | ◎ | – |
| 8 | | 1000人未満 | ** | *****<br>*** | ***** | ****<br>* | ** | ****<br>* | mm月dd日 | mm月dd日 | ◎ | – | ◎ | ◎ |
| 9 | | 1000人未満 | ** | *****<br>**** | ***** | ****<br>* | ** | ****<br>* | mm月dd日 | mm月dd日 | ◎ | – | ◎ | ◎ |
| | | 1000人未満 | ** | ** | ***** | **** | ** | **** | mm月dd日 | mm月dd日 | ◎ | – | ◎ | ◎ |
| 10 | | 社内 | *****<br>**** | ***** | **** | **** | – | – | mm月dd日 | mm月dd日 | ◎ | – | – | – |
| 11 | 運輸… | | | ***** | ***** | ****<br>* | ** | **** | mm月dd日 | mm月dd日 | – | – | – | ◎ |
| 12 | 官公庁 | – | ** | ***** | ***** | **** | – | – | mm月dd日 | mm月dd日 | ◎ | ◎ | ◎ | ◎ |
| 13 | 自治体 | – | ***** | **** | **** | **** | *** | **** | mm月dd日 | mm月dd日 | ◎ | ◎ | ◎ | – |
| 14 | ***ベンダ | 1000人以上 | | *** | ***** | ****<br>* | *** | ****<br>* | mm月dd日 | mm月dd日 | – | – | – | ◎ |
| 15 | | 1000人以上 | | *****<br>** | ***** | ****<br>* | *** | ****<br>* | mm月dd日 | mm月dd日 | – | – | – | ◎ |

**図 4.14　調査報告資料の例**

### 4.4.2 実験方法

三種類の業務情報それぞれについて，抑止キーワード登録による不要文書非提示率の改善および再現率の維持を評価した。不要文書非提示率は，提示を必要としない状況のうち，提示を要しないと正しく判別された割合を計算する。不要文書非提示率および提示するべき状況の発生確率から，提示された業務情報のうち実際に参照を要する確率，すなわち適合率を求めることができるが，本章では提示するべき状況の発生確率から独立して議論をすすめるため，不要文書非提示率での評価を用いる。実験には Core i5 M520 (2.4GHz) CPU および 3GB メモリの PC を利用した。

評価の手順には，管理部門ユーザによる業務情報の登録および実務部門ユーザの業務上表示文字列に対する提示処理の両方を含めた。図 4.15 を参照して手順の詳細を説明する。まず，業務情報の登録時を想定して，教師データから 4.3 節で述べた手順で判別器を構成する。正例は，表 4.2 に示す対象とする種類の業務情報における提示要の業務上表示文字列から 1 件残したものを用いる。また，初期負例は以下の三種類の業務上表示文字列の和集合を用いる : 1) 対象とする種類の業務情報における提示不要の業務上表示文字列, 2) 対象としない二種類の業務情報における提示要および不要の業務上表示文字列, 3) 表 4.3 に

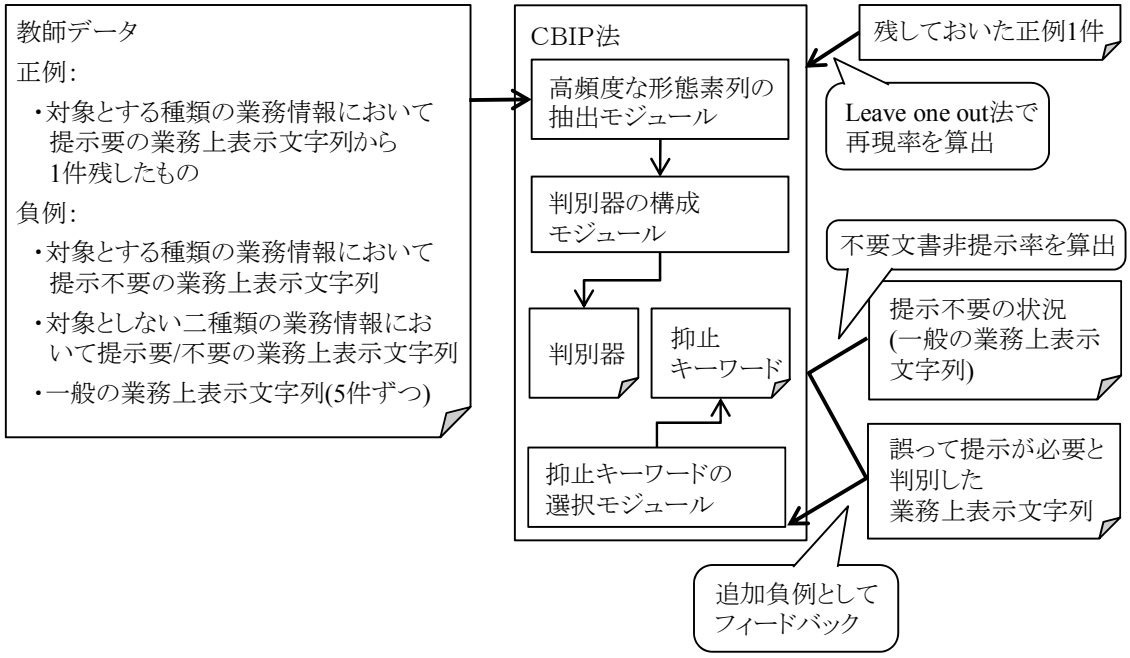



**図 4.15　評価の手順**

示す一般の業務上表示文字列を 5 件ずつ選んだもの。その後，実運用開始後の業務上表示文字列を想定して表 4.3 に示す一般の業務上表示文字列のうち初期負例に用いなかったものに対して順番に参照必要性の判別を試行する。誤って提示が必要と判別するたびにその業務上表示文字列が追加負例としてフィードバックされたと考え，抑止キーワードを選択する。$a$ 件の一般の業務上表示文字列に対して判別を行うまでに誤って提示が必要と判別した回数 $b$ に対し，$(a-b)/a$ を不要文書非提示率として求める。また，$a$ 件の一般の業務上表示文字列に対して判別を試行しフィードバックに応じて抑止キーワードを選択した後の状態で，残しておいた正例 1 件に対して業務情報の参照を要すると判別されるか調べる。対象の正例 1 件を変えながら繰り返すこと，すなわち Leave one out 法により再現率を求めた。なお，表 4.3 に示す一般の業務上表示文字列から初期負例として選ぶ件数は，管理部門ユーザが業務情報を登録する際の工数を抑制する観点から設定した。

また，抑止キーワード選択方法の比較のため，*freqP* の値によらず $\mathrm{Exist}(t, f_i)=1$ である $f_i$ のうち最大の *freqN* を持つものを抑止キーワードとする単純な方式も評価した。この方式では *freqP* を確認しないため，$\mathrm{Exist}(t, f_i)=1$ である $f_i$ のうち最大の *freqN* を持つものにおいて $freqP > 0$ である場合は提案方式よりも大きい *freqN* を選ぶことになり，再現率は無視して不要文書非提示率のみを優先することになる。

### 4.4.3 実験結果

三種類の業務情報について，抑止キーワードの登録による不要文書非提示率の改善および再現率の維持について評価した結果を図 4.16 から図 4.18 に示す。横軸は業務情報の参照を要しない一般の業務上表示文字列に対して提示要否の判別を試行した回数 $a$，縦軸は不要文書非提示率および再現率である。不要文書非提示率について，提案手法および単純な方式は同等に改善することができ，三種類の業務情報全てについてほぼ 100%に達した。単純な方式では一般の業務上表示文字列への適用を進めて抑止キーワードの登録が増加するに従い再現率が急激に低下してしまったが，提案手法では再現率の低下は軽度に留まっており，不要文書非提示率の改善と再現率の維持の両立を実現できた。また，言い回し表現の抽出および判別器構成に要した時間の合計は，ソフトウェア製品に対する窓口部署設置の通知では平均 14 秒，外部参加者があるセミナーを社内の会議室で開催する際の事務手続きおよび PC 棚卸の担当者変更の通知では約 16 秒であった。抑止キーワードの選択に要する時間は，三種類の業務情報全てにおいて 1 秒未満であった。

## 4.5 考察

提案手法では「業務情報の提示が不要である」旨のフィードバックに対して抑止キーワ

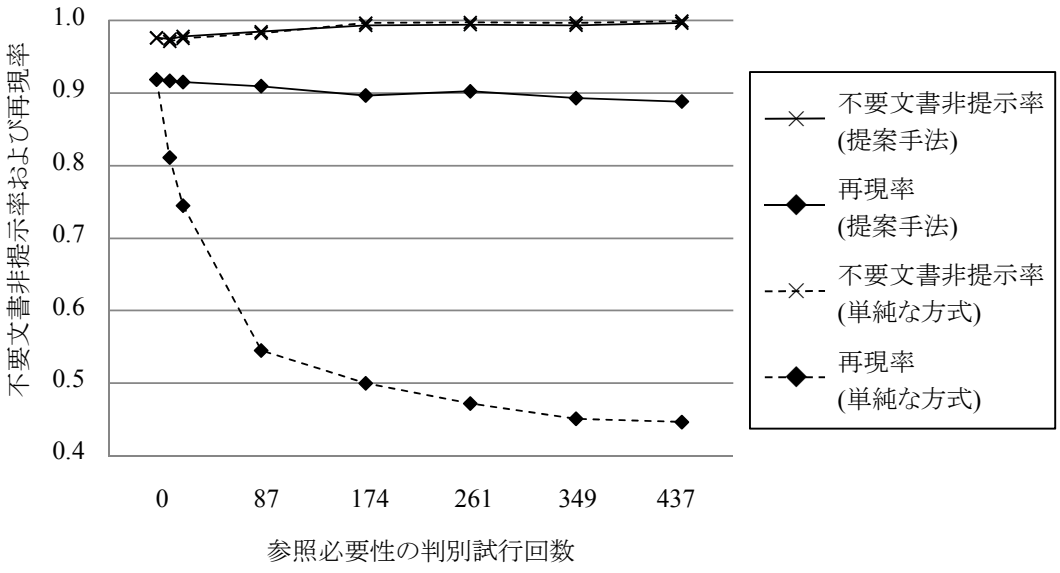



**図 4.16　ソフトウエア製品に対する窓口部署設置の通知に対する抑止キーワードによる不要文書非提示率および再現率**

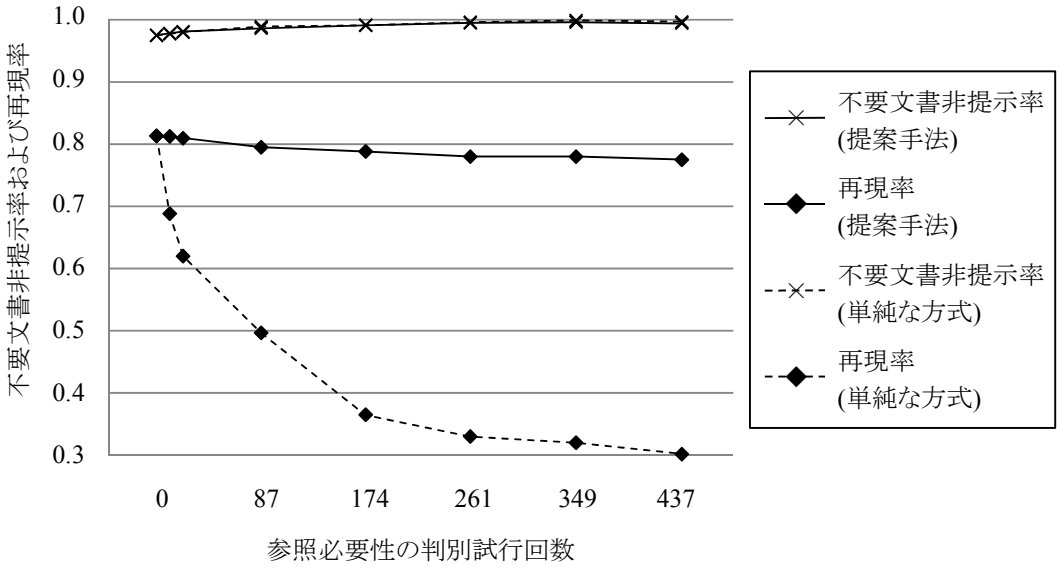



**図 4.17　外部参加者があるセミナーを社内の会議室で開催する際の事務手続きに対する抑止キーワードによる不要文書非提示率および再現率**

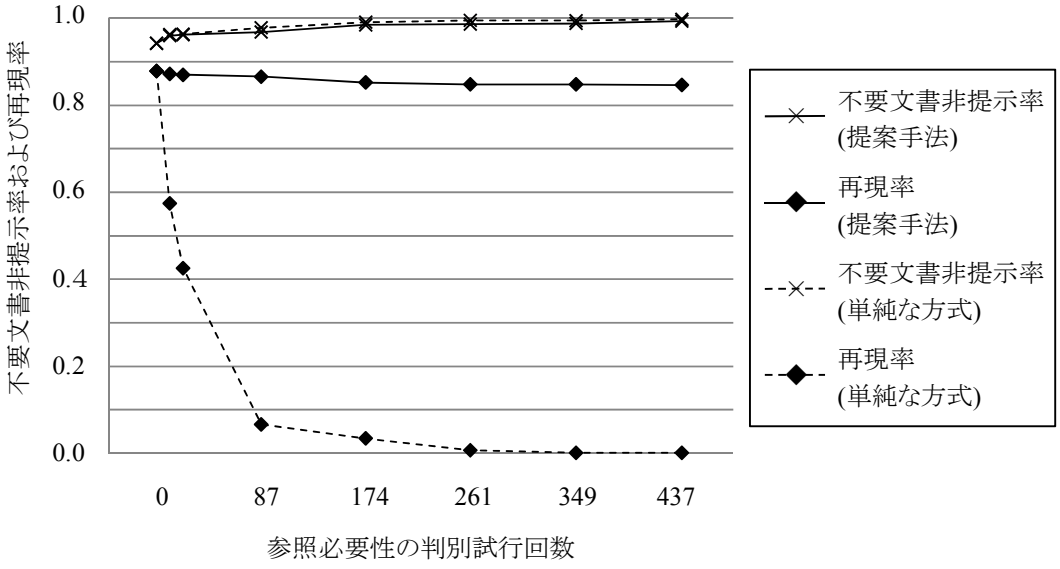



**図 4.18　PC 棚卸の担当者変更の通知に対する抑止キーワードによる不要文書非提示率および再現率**

ードを選ぶため，フィードバックを受けた業務情報が同じ業務上表示文字列に対して再度提示されることを確実に防ぐことができる。また，判別試行回数が増えると不要文書非提

示率が改善してほぼ 100%に達したことから，抑止キーワードにより将来の不要な提示を未然に防ぐことができた。さらに，図 4.16 から図 4.18 に示す通り，参照必要性の判別試行回数を増やして抑止キーワードを選んでも再現率はほとんど低下せず，不要文書非提示率の改善と再現率の維持の両立を実現できている。従って，4.2 節で述べた要件を満たすことができた。

単純な方式と提案手法の再現率の低下の度合いの違いは，本来提示が必要な文書に合致してしまう可能性の低い抑止キーワードを選ぶことができたかどうかが影響していた。例えば第二の業務情報において，提案手法では「通知」・「提出」・「申請」などのセミナー開催業務と関連性の低いものが抑止キーワードとして選択されたのに対し，単純な方式では「講演」・「テーマ」・「本社」などのセミナーの内容や開催場所に関連するものが抑止キーワードとして選択された。このため，セミナー開催告知についての業務上表示文字列にも抑止キーワードが含まれてしまい，本来必要な提示が阻害されてしまった。

CBIP 法による抑止キーワード登録と，追加負例から抑止キーワードを登録するのではなく決定木を構成し直す手法を判別精度と保持するデータ量の観点から比較する。追加負例を初期負例に加えて決定木を構成し直す場合について不要文書非提示率および再現率を求めた結果を表 4.4 に示す。第一および第三の業務情報においては，学習用データの負例が増加すると正例でも負例であると推測する確率が高くなり，不要文書非提示率は改善するが再現率は CBIP 法の場合よりも低下した。また第二の業務情報のように再現率が維持される場合であっても，決定木を再構成する手法においては CBIP 法のように抑止キーワードを登録する手法と異なり，実務部門ユーザから与えられた「業務情報の提示が不要である」旨のフィードバックを確実に反映できるとは限らない。

運用時に保持するデータ量からの比較結果は以下の通りである。追加負例を貯めておき一定期間ごとに決定木を構成し直すためには，正例・初期負例および追加負例すべてについて全文データを保持しておかなくてはならない。実務部門ユーザの数を $num_{user}$，一人の実務部門ユーザが一日に作成・閲覧する業務上表示文字列の数を平均 $num_{str}$，登録済の業務情報の数を $num_{info}$，業務情報を登録してからの運用日数を平均 $num_{day}$，不要文書非提示率を平均 $p_{trn}$ とおく。業務上表示文字列の数 $num_{str}$ については，例えば部下が作成した作業計画書に対して管理職向けの労務管理規則を参照して指導を行う場合や，上司から与えられた概要レベルの業務指示に対し手順書を確認して作業を行う場合などを想定し，作成だけではなく閲覧も対象に含める。この場合，$num_{user} * num_{str} * (1 - p_{trn}) * num_{day} * num_{info}$ 件の追加負例の全文テキストを保持し続け，正例，初期負例，および $num_{user} * num_{str} * (1 - p_{trn}) * num_{day}$ 件の追加負例を用いた決定木の再構成を $num_{info}$ 回行う必要がある。例えば $num_{user}$ =

表 4.4　決定木を再構成する場合における判別精度

(A)　ソフトウエア製品に対する窓口部署設置の通知

| | 参照必要性の判別試行回数 | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 87 | 174 | 261 | 349 | 437 |
| 不要文書非提示率（提案手法） | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 |
| 不要文書非提示率（決定木再構成） | --- | 0.97 | 0.99 | 0.99 | 0.99 | 0.99 |
| 再現率（提案手法） | 0.91 | 0.90 | 0.89 | 0.90 | 0.89 | 0.88 |
| 再現率（決定木再構成） | --- | 0.91 | 0.86 | 0.83 | 0.83 | 0.83 |

(B)　外部の参加者があるセミナーを社内の会議室で開催する際の事務手続き

| | 参照必要性の判別試行回数 | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 87 | 174 | 261 | 349 | 437 |
| 不要文書非提示率（提案手法） | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 |
| 不要文書非提示率（決定木再構成） | --- | 0.97 | 0.99 | 0.99 | 0.99 | 0.99 |
| 再現率（提案手法） | 0.81 | 0.79 | 0.78 | 0.78 | 0.78 | 0.77 |
| 再現率（決定木再構成） | --- | 0.81 | 0.83 | 0.83 | 0.80 | 0.80 |

(C)　PC 棚卸の担当者変更の通知

| | 参照必要性の判別試行回数 | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 87 | 174 | 261 | 349 | 437 |
| 不要文書非提示率（提案手法） | 0.94 | 0.96 | 0.98 | 0.98 | 0.98 | 0.99 |
| 不要文書非提示率（決定木再構成） | --- | 0.94 | 0.98 | 0.99 | 0.99 | 0.99 |
| 再現率（提案手法） | 0.87 | 0.86 | 0.85 | 0.84 | 0.84 | 0.84 |
| 再現率（決定木再構成） | --- | 0.87 | 0.85 | 0.82 | 0.80 | 0.79 |

3000，$num_{str}=50$，$num_{info}=500$，$num_{day}=100$，$p_{trn}=0.9999$ の場合，75 万件の追加負例の全文テキストを保持し続け，1500 件の追加負例を用いた決定木の再構成を一定期間ごとに 500 回行う必要があることになる。$num_{user}$ は企業規模の分類に一般に使用される従業員数の基準から，$num_{info}$ は社内規則・業務手順・通知類がそれぞれ数百件規模で蓄積されているとして設定した。言い回し表現の抽出および判別器の構成に要する時間が正例および初

期負例の数に比例すると仮定すると，計算には約 20 時間を要する。このため，夜間や週末などの業務停止中に計算を行わなくてはならない時間制約の点から運用は困難である。これに対して提案手法では，決定木構成に用いた言い回し表現および各追加負例について選択した抑止キーワードのみ保持しておけばよく，全文テキストを保持し続ける必要はない。従って両方の観点において，CBIP 法は単純な判別器再構成と比較して運用上優位と考えられる。

本章では，第一の業務情報に対する正例として問合せのメールを，第三の業務情報に対する正例として棚卸結果提出のメールを用いた。メールのヘッダに現れる「To:」などの文字列は業務上表示文字列においては過度に高頻度であるため, 4.3.2 節で述べた手順により除かれている。このため，Web の入力フォームやグループウェアなどを介して行う文書作成の業務状況に対しても同様の効果が期待できる。また，第二の業務情報に対する正例としては外部の参加者があるセミナーを社内の会議室で開催するためのメールを用いたが，同様の理由から，ワープロソフトや HTML(HyperText Markup Language)エディタなどを用いる文書作成の業務状況においても同様の効果が期待できる。さらに，文書作成だけでなく閲覧に関わる業務状況においても同様である。

## 4.6 結言

本章では，実務部門ユーザの業務遂行状況に応じて業務情報の提示要否を判別する CBIP 法を提案した。ユーザのフィードバックを確実に反映し，不要文書非提示率と再現率の両立を実現するため，業務上利用するアプリケーション表示文字列で過度に高頻度に現れる典型文字列の除去および言い回し表現を用いることによる各文字列の特徴の把握を行った。言い回し表現を用いた抑止キーワード登録により，不要文書非提示率のみを優先した抑止キーワード選択と比較して再現率を改善することができた。また単純な判別器再構成と比較して判別精度およびデータ量の観点から利点がある。これにより，業務遂行状況に応じて業務情報を提示し，管理部門および実務部門双方のユーザにとって負担なく業務情報の周知徹底が図られると考える。登録されている業務情報の数が多いほど，実務部門ユーザの記憶にのみ依存した参照は困難になり，提案手法の効果が大きくなる。また，提案手法は登録されている業務情報の数に対して線形の計算時間であり，スケーラビリティが高い。

CBIP 法は，業務上表示文字列において過度に高頻度に現れる文字列があること，および業務内容に特徴的な表現で頻度の上昇が見られることに着目し，高頻度な形態素列を用

いた判別器の構成および抑止キーワードの選択を行うことで業務上表示文字列における特徴を効率よく利用することを可能としている。このようなアプローチは，機械学習分野における属性選択[Yang1997]の考え方を企業内で用いられる定型性の低い文書データの特徴に合わせて適用したものだと考えることができる。

CBIP 法のさらなる拡張として，提示した業務情報が利用されたことを検知できれば，参照要否判別の再現率を改善できると期待できる。さらに，業務情報をあらかじめ人事・経理・設備管理・輸出管理などのカテゴリに分類しておき，カテゴリ間で抑止キーワードを共有することによる，不要文書非提示率の改善に要するフィードバック回数削減が考えられる。また，抑止キーワードには，顧客情報や関係者外秘の文字列が選ばれる可能性もあるため，セキュリティの検討が必要である。

# 第 5 章

# 結論

## 5.1 本研究のまとめ

本論文では，企業情報システムにおけるデータからの業務上着目するべき箇所の抽出の効率化についての研究成果を，以下の 4 章に分けて述べた。

第 1 章では，企業情報システムにおける業務効率化の課題について，業務の内容や遂行状況に応じて様々な内容が様々な書式で記載される定型性の低いデータでは，データから業務上重要な箇所を抽出する際に高度なスキルを用いた解釈や判断が必要となる旨を述べた。さらに，データに内在する法則性に基づいて抽出を行うことを方針とし，数値データおよび文書データそれぞれを対象として，本研究で取り上げる課題について延べ，関連研究を概観すると共に，それぞれの解決方針を示した。

第 2 章では，数値データからユーザの業務において着目するべき箇所の専門家の知見に基づいた抽出に関し，DNA 実験データを例として取り上げ，真のデータの識別技術を提案した。DNA 配列やサンプルによってノイズデータの現れ方は異なるため，まず，入力データからノイズデータの解釈が容易なサンプルを抽出してノイズデータの傾向を定量化し，その後，ノイズデータの傾向に基づいて観測データの識別を行う，二段階の処理方式を提案した。174 個の DNA 実験データを用いて提案方式の識別精度を評価し，ヒトゲノム全体で平均 94%の精度で真のデータを識別できることを示した。また，解釈が容易なデータを用いて,内在的法則性について蓄積された専門家の知見を定量化する手法について，企業内の他の数値データへの適用性を述べた。

第 3 章では，文書データからのメタデータ抽出のためのルール生成方式を提案した。メタデータの記載上の特徴を洩れなく集めるため，サンプル文書における正解メタデータの記載からルールの候補を列挙する。さらに，必要性の低いルールにより誤った抽出が行われるのを防ぐため，サンプル文書の中で目的外の文字列にもルール候補があてはまらないか調べることでルール候補の絞り込みを行う。6 つの異なる顧客とのビジネス案件において作成された営業文書および 5 つの研究プロジェクトにおいて作成された週次作業報告書

を用いて，人手で設定したルールと自動的に生成したルールとの比較を行った。メタデータ抽出の再現率がほぼ同等であり，ルールの設定に要する時間を大幅に削減できたことから，提案手法の有効性を示した。

第 4 章では，業務情報の周知のための業務遂行状況に応じた提示要否の判別方式を提案した。提案手法では，業務上利用するアプリケーションの表示文字列の例とそれぞれの状況における参照要否を入力として，提示要否の判別条件を構成すると共に，提示した業務情報に対してユーザから提示不要である旨の指摘があった場合はフィードバックとして判別条件の修正を行う。業務上利用されるアプリケーションの表示文字列における特徴に基づき，高頻度な形態素列を「言い回し表現」として抽出し，決定木の構成および提示を抑止するキーワードの選択に利用する方式を提案した。3 種類の業務情報に対して提案方式を適用し，提示要否の正確な判別と，フィードバックによる不要な業務情報の提示の防止が両立できていることを確かめ，提案手法の有効性を示した。

## 5.2 今後の課題

以下に今後の課題について述べる。

### （1）数値データと文書データからの抽出技術の連携

本論文では，数値データおよび文書データそれぞれに対して重要な部分を抽出する手法を提案した。さらなる取組として，これらの手法を連携させより高度なデータ抽出を行うことが考えられる。営業文書における注文数量と金額，スケジュール管理文書におけるプロジェクト進捗状況と完了予定日，決算報告文書における業績数値と次年度目標値など，文書中の記載位置としての近傍関係だけでなく数値データとしての近傍関係にも法則性が内在しているデータにおいて，不利な取引条件，注視するべき案件など重要なデータを抽出することは従来専門家の知見に基づいて人手で行われている。これらのデータについてもビジネスにおいては可読性を向上させる記載方法が行われていると期待できるため，本論文で提案した手法における記載上の特徴を調べるための技術を組み合わせることが必要になると考えられる。本論文で提案した手法の連携によりこれらの業務の効率化支援を実現することは今後の課題である。

### （2）データ抽出に基づく業務遂行支援

本論文で検討した，ユーザが業務を遂行するために企業情報システムで用いられるデー

タから重要な部分を抽出することの次の段階として，抽出したデータを用いてのユーザの業務判断についても支援を行うことが考えられる。本論文ではデータ抽出における場所に着目して特徴を調べることによる支援について検討を行ったが，業務判断の支援ではデータ抽出における理由に着目して特徴を調べることが必要になる。

ユーザの業務判断に対する情報システムによる支援が実現されれば，さらなる業務効率向上につながるとともに，業務遂行結果から個人の業務スキルに依存する部分を削減し，一定のレベルを保証することができる。このことにより，企業はより競争力を創出する業務に人的資源を集中することが可能となる。

# 謝辞

本研究の全過程を通じ，懇切なるご指導とご鞭撻を賜りました大阪大学大学院情報科学研究科マルチメディア工学専攻 薦田憲久教授に心から感謝申し上げます。

本研究をまとめるにあたって貴重なお時間を割いて頂き，丁寧なるご教示を賜りました大阪大学大学院情報科学研究科マルチメディア工学専攻 細田耕教授，原隆浩准教授，秋吉政徳准教授に謹んで深謝致します。

大学院博士後期課程において，マルチメディア工学全般に関して，親切なるご指導とご助言を賜りました大阪大学大学院情報科学研究科マルチメディア工学専攻 西尾章治郎教授，藤原融教授に厚く御礼申し上げます。

本研究は，日立ソフトウェアエンジニアリング（株）（現（株）日立ソリューションズ）において，社内外の多くの方々のご指導とご協力を得て行ったものが元となっています。本研究の機会を与えていただくとともに，業務の傍らで本論文を纏めるにあたり，暖かいご配慮を賜った，常務執行役員 前澤裕行氏，元執行役員（現日立 INS ソフトウェア株式会社取締役社長）露木陽介氏，技術開発本部 本部長 正村勉氏，元研究部長 小柳和子博士（現情報セキュリティ大学院大学教授），研究部長 小野山隆博士，主任研究員（現知的財産権センタ センタ長）中重亮氏に心から御礼申し上げます。

第 2 章の研究を進めるにあたり，懇切なるご指導とご鞭撻を賜りました，国立遺伝学研究所生命情報・DDBJ 研究センター 五條堀孝教授，東海大学医学部基礎医学系分子生命科学 猪子英俊教授，田宮元助教授（現山形大学医学部教授），生田智樹博士，矢倉勝博士（現国立遺伝学研究所細胞遺伝研究系微生物遺伝研究部門），牧野悟士博士，新屋みのり博士（現国立遺伝学研究所系統生物研究センター助手）に心から感謝申し上げます。

第 3 章の研究に関して，メタデータ抽出に関して技術のご提供および様々なご助言を頂きました（株）日立製作所 主任研究員 故 丸川勝美氏，主任技師 池田尚司氏，主任技師 永崎健氏，主任研究員 藤尾正和博士に心から御礼申し上げます。また，ビジネスニーズについて様々なご討論ご助言を頂くとともに，多大なるご支援を頂きました日立ソフトウェアエンジニアリング（株）ミドルソフト第 4 設計部（現（株）日立ソリューションズ スマートオフィスシステム部）担当部長 盛井恒男氏に心から御礼申し上げます。

また，研究を進めるにあたり日々様々なご支援とご配慮を頂きました日立ソフトウェア

エンジニアリング（株）（現（株）日立ソリューションズ） 研究部の先輩，同僚，後輩の方々に心から御礼申し上げます。

最後に，いつも暖かく励ましてくれた，家族に感謝します。

# 参考文献

[Ashish1997] N. Ashish and C. A. Knoblock: Wrapper Generation for Semi-Structured Internet Sources, *SIGMOD Record*, Vol.26, No.4, pp.8-15, 1997.

[Cesarini1998] F. Cesarini, M. Gori, S. Marinai, and G. Soda: INFORMys: A flexible invoice-like form reader system, *IEEE Trans. of PAMI*, Vol.20, No.7, pp.730-745, 1998.

[Clark1988] J. M. Clark: Novel Non-templated Nucleotide Addition Reactions Catalyzed by Procaryotic and Eucaryotic DNA Polymerases, *Nucleic Acids Research*, Vol.16, pp.9677-9686, 1988.

[Crow1988] J. E. Crow: Eighty years ago: the beginnings of population genetics, *Genetics*, Vol.119, pp.473-476, 1988.

[Dib1996] C. Dib, S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Kazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Weissenbach: A Comprehensive Genetic Map of the Human Genome based on 5,264 Microsatellites, *Nature*, Vol.380, pp.152-154, 1996.

[Dietrich1992] W. Dietrich, H. Katz, S. E. Lincoln, H. S. Shin, J. Friedman, N. C. Dracopoli, and E. S. Lander: A Genetic Map of the Mouse Suitable for Typing Intraspecific Crosses, *Genetics*, Vol.131, pp.423-447, 1992.

[Einstein1990] G. O. Einstein and M. A. McDaniel: Normal Aging and Prospective Memory, *Journal of Experimental Psychology: Learning, Memory and Cognition*, Vol.16, No.4, pp.717-726, 1990.

[Esposito2004] F. Esposito, D. Malerba, G. Semeraro, S. Ferilli, O. Altamura, T. M. A. Basile, M. Berardi, M. Ceci, and N. D. Mauro: Machine Learning Methods for Automatically Processing Historical Documents: from Paper Acquisition to XML Transformation, in *Proc. of 1st International Workshop on Document Image Analysis for Libraries*, pp.328-335, 2004.

[Fan2002] I. S. Fan, G. Li, and M. Lagos-Hernandez: A Rule Level Knowledge Management System for Knowledge Based Engineering Applications, in *Proc. of DETC2002*, pp.813-821, 2002.

[Freitag2000] D. Freitag and A. McCallum: Information Extraction with HMM Structures Learned by Stochastic Optimization, in *Proc. of American Association for Artificial Intelligence*, pp.584-589, 2000.

[Fujio2001] M. Fujio, N. Furukawa, S. Watanabe, and H. Sako: Automatic Generation of the Keywords Dictionary for Efficient Document Form Identification, *Technical report of IEICE. PRMU*, Vol.101, No.421, pp.93-98, 2001.

[Gyapay1994] G. Gyapay, J. Morissette, A. Vignal, C. Dib, C. Fizames, P. Millasseau, S. Marc, G. Bernardi, M. Lathrop, and J. Weissenbach: The 1993-94 Genethon Human Genetic Linkage Map, *Nature Genetics*, Vol.7, pp.246-339, 1994.

[Hall2009] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten: The WEKA Data Mining Software: An Update, *ACM SIGKDD Explorations Newsletter*, Vol.11, No.1, pp.10-18, 2009.

[Handley2005] J. C. Handley, A. M. Namboodiri, and R. Zanibbi: Document Understanding System Using Stochastic Context-Free Grammars, in *Proc. of 8th International Conference on Document Analysis and Recognition*, pp.511-515, 2005.

[Hauge1993] X. Y. Hauge and M. Litt: A Study of the Origin of 'Shadow Bands' seen when Typing Dinucleotide Repeat Polymorphisms by the PCR, *Human Molecular Genetics*, Vol.2, pp.411-415, 1993.

[Hu1993] G. Hu: DNA Polymerase-catalyzed Addition of Nontemplated Extra Nucleotides to the 3' End of a DNA fragment, *DNA and Cell Biology*, Vol.12, pp.763-770, 1993.

[Ishitani1999] Y. Ishitani: Logical Structure Analysis of Document Images Based on Emergent Computation, in *Proc. of 5th International Conference on Document Analysis and Recognition*, pp.189-192, 1999.

[JIS-Z8303] JIS-Z8303：帳票の設計基準，1953.

[Kerdprasop2011] K. Kerdprasop and N. Kerdprasop: Data Preparation Techniques for Improving Rare Class Prediction, in *Proc. of the 13th WSEAS International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems*, pp.204-209, 2011.

[Knapik1998] E. W. Knapik, A. Goodman, M. Ekker, M. Chevrette, J. Delgado, S. Neuhauss, N. Shimoda, W. Driever, M. C. Fishman, and H. J. Jacob: A Microsatellite Genetic Linkage Map for Zebrafish (Danio rerio), *Nature Genetics*, Vol.18, pp.338-343,

1998.

[Kramer2007] M. Kramer, H. Kaprykowsky, D. Keysers, and T. Breuel: Bibliographic Meta-Data Extraction using Probabilistic Finite State Transducers, in *Proc. of 9th International Conference on Document Analysis and Recognition*, pp.609-613, 2007.

[Leibelt2003] C. S. Leibelt, C. E. Boland, C. L. Brown, T. L. Hatch, Y. Daoudi, Y. Lou, and R. K. Roby: Verification of GeneMapper Software for STR Analysis, in *Proc. of the Annual Meeting of American Academy of Forensic Sciences*, Vol.9, pp.50-51, 2003.

[LeDuc1995] C. LeDuc, P. Miller, J. Lichter, and P. Parry: Batched Analysis of Genotypes, PCR Methods Application, Vol.4, pp.331-336, 1995.

[Lipkin1998] E. Lipkin, M. O. Mosig, A. Darvasi, E. Ezra, A. Shalom, A. Friedmann, and M. Soller: Quantitative Trait Locus Mapping in Dairy Cattle by Means of Selective Milk DNA Pooling using Dinucleotide Microsatellite Markers: Analysis of Milk Protein Percentage, *Genetics*, Vol.149, pp.1557-1567, 1998.

[Lyman2003] P. Lyman, H. Varian, K. Swearingen, P. Charles, N. Good, L. L. Jordan, and J. Pal: How Much Information? 2003, Regents of the University of California, http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/, 2003.

[Magnuson1996] V. L. Magnuson, D. S. Ally, S. J. Nylund, Z. E. Karanjawala, J. B. Rayman, J. I. Knapp, A. L. Lowe, S. Ghosh, and F. S. Collins: Substrate Nucleotide-determined Non-templated Addition of Adenine by Taq DNA Polymerase: Implications for PCR-based Genotyping and Cloning, *Biotechniques*, Vol.21, pp.700-709, 1996.

[Matsumoto2004] T. Matsumoto, W. Yukawa, Y. Nozaki, R. Nakashige, M. Shinya, S. Makino, M. Yagura, T. Ikuta, T. Imanishi, H. Inoko, G. Tamiya, and T. Gojobori: Novel Algorithm for Automated Genotyping of Microsatellites, *Nucleic Acids Research*, Vol.32, pp.6069-6077, 2004.

[Matsumoto2005] T. Matsumoto and R. Nakashige: Evaluating Robustness of Algorithm for Microsatellite Marker Genotyping, in *Proc. of 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp.158-164, 2005.

[Matsumoto2006] T. Matsumoto, R. Nakashige, T. Watanabe, and Y. Sugimoto: Applying Genotyping Algorithm for Microsatellite Markers to Bovine Data, *in Proc. of the 20th IUBMB International Congress of Biochemistry and Molecular Biology*,

2P-B-036, 2006.

[Matsumoto2010] T. Matsumoto, M. Oba, and T. Onoyama: Sample-based Collection and Adjustment Algorithm for Metadata Extraction Parameter of Flexible Format Document, in *Proc. of The 10th International Conference on Artificial Intelligence and Soft Computing*, pp.566-573, 2010.

[Memmel1997] M. Memmel and A. Dengel: Sharing Contextualized Attention Metadata to Support Personalized Information Retrieval, in *Proc. of ACM/IEEE Workshop on Contextualized Attention Metadata: Personalized Access to Digital Resources*, pp.19-26, 1997.

[Minagawa2006] A. Minagawa, Y. Fujii, H. Takebe, and K. Fujimoto: A Method of Logical Structure Analysis for Form Images with Various Layouts by Belief Propagation, *IEIC Technical Report*, Vol.106, pp.17-22, 2006.

[Murphy1986] J. J. Murphy: Technical Analysis of The Futures Markets: A Comprehensive Guide to Trading Methods and Applications, New York Institute of Finance, Prentice-Hall, 1986.

[Murray1993] V. Murray, C. Monchawin, and P. R. England: The Determination of the Sequences Present in the Shadow Bands of a Dinucleotide Repeat PCR, *Nucleic Acids Research*, Vol.21, pp.2395-2398, 1993.

[Palsson1999] B. Palsson, F. Palsson, M. Perlin, H. Gudbjartsson, K. Stefansson, and J. Gulcher: Using Quality Measures to Facilitate Allele Calling in High-throughput Genotyping, *Genome Research*, Vol.9, pp.1002-1012, 1999.

[Parmentier1997] F. Parmentier and A. Belaid: Logical Structure Recognition of Scientific Bibliographic References, in *Proc. of 4th International Conference on Document Analysis and Recognition*, pp.1072-1076, 1997.

[Perlin1994] M. W. Perlin, M. B. Burks, R. C. Hoop, and E. P. Hoffman, Toward Fully Automated Genotyping: Allele Assignment, Pedigree Construction, Phase Determination, and Recombination Detection in Duchenne Muscular Dystrophy, *American Journal of Human Genetics*, Vol.55, pp.777-787, 1994.

[Riloff1993] E. Riloff: Automatically Constructing a Dictionary for Information Extraction Tasks, in *Proc. of the Eleventh National Conference on Artificial Intelligence*, pp.811-816, 1993.

[SAP2008] SAP ジャパン：ビジネスインテリジェンス～中小企業のための完全ガイド～,

SAP White Paper, 49009213J, 2008.

[Shilakes1998] C. C. Shilakes and J. Tylman: Enterprise Information Portals, *Merrill Lynch In-depth Report*, 1998.

[Shimoda1999] N. Shimoda, E. W. Knapik, J. Ziniti, C. Sim, E. Yamada, S. Kaplan, D. Jackson, F. de Sauvage, H. Jacob, and M. C. Fishman: Zebrafish Genetic Map with 2000 Microsatellite Markers, *Genomics*, Vol.58, pp.219-232, 1999.

[Smith1995] J. R. Smith, J. D. Carpten, M. J. Brownstein, S. Ghosh, V. L. Magnuson, D. A. Gilbert, J. M. Trent, and F. S. Collins: Approach to Genotyping Errors Caused by Nontemplated Nucleotide Addition by Taq DNA Polymerase, *Genome Research*, Vol.5, pp.312-317, 1995.

[Stoughton1997] R. Stoughton, R. Bumgarner, W. J. Frederick III, and R. A. McIndoe: Data-adaptive Algorithms for Calling Alleles in Repeat Polymorphisms, *Electrophoresis*, Vol.18, pp.1-5, 1997.

[Tamiya2005] G. Tamiya, M. Shinya, T. Imanishi, T. Ikuta, S. Makino, K. Okamoto, K. Furugaki, T. Matsumoto, S. Mano, S. Ando, Y. Nozaki, W. Yukawa, R. Nakashige, D. Yamaguchi, H. Ishibashi, M. Yonekura, Y. Nakami, S. Takayama, T. Endo, T. Saruwatari, M. Yagura, Y. Yoshikawa, K. Fujimoto, A. Oka, S. Chiku, S. E. V. Linsen, M. J. Giphart, J. K. Kulski, T. Fukazawa, H. Hashimoto, M. Kimura, Y. Hoshina, Y. Suzuki, T. Hotta, J. Mochida, T. Minezaki, K. Komai, S. Shiozawa, A. Taniguchi, H. Yamanaka, N. Kamatani, T. Gojobori, S. Bahram, and H. Inoko: Whole Genome Association Study of Rheumatoid Arthritis using 27,039 Microsatellites, *Human Molecular Genetics*, Vol. 14, No. 16, pp.2305-2321, 2005.

[Tan2005] S. Tan: Neighbor-weighted K-nearest Neighbor for Unbalanced Text Corpus, *Expert Systems with Applications*, Vol.28, No.4, pp.667-671, 2005.

[Tautz1989] D. Tautz: Hypervariability of Simple Sequences as a General Source for Polymorphic DNA markers, *Nucleic Acids Research*, Vol.17, pp.6463-6471, 1989.

[Taylor1992] S. L. Taylor, R. Fritzson, and J. A. Pastor: Extraction of Data from Preprinted Forms, *Machine Vision and Applications*, Vol.5, pp.211-222, 1992.

[Tereba1999] A. Tereba: Tools for Analysis of Population Statistics, *Profiles in DNA*, Vol.2, pp.14-16, 1999.

[Weissenbach1992] J. Weissenbach, G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vaysseix, and M. Lathrop: A Second-generation Linkage Map of the

Human Genome, *Nature*, Vol.359, pp.794-801, 1992.

[Wnek2002] J. Wnek: Machine Learning of Generalized Document Templates for Data Extraction, in *Proc. of 2nd International Workshop on Document Analysis Systems*, pp.457-468, 2002.

[Yang1997] Y. Yang and J. O. Pedersen:, A Comparative Study on Feature Selection in Text Categorization, in *Proc. of International Conference on Machine Learning*, pp.412-420, 1997.

[Yoshinaga2006] N. Yoshinaga and K. Torisawa: Finding Specification Pages from the Web, *Transactions of the Japanese Society for Artificial Intelligence*, Vol.21, No.6, pp.493-501, 2006.

[Zipf1932] G. K. Zipf: Selected Studies of the Principle of Relative Frequency in Language, Cambridge, 1932.

[浅川 1992] 浅川悟志，川下靖司，坂田淳，畠山敦：フルテキストサーチシステム Bibliotheca/TS の開発(1) －システムの概要－，情報処理学会第 45 回全国大会予稿集，3-239-3-240, pp.243-244, 1992.

[医薬産業政策研究所 2007] 医薬産業政策研究所：製薬産業の将来像～2015 年に向けた産業の使命と課題～，産業レポート，No.4, 2007.

[梅原 2008] 梅原寿夫：ECM の最新動向，JIIMA 文書情報マネジメントセミナー，2008.

[川上 2010] 川上潤司：コスト削減や法令順守を視野に基幹システムからの帳票印刷を集中管理，IT Leaders, 2010 年 03 月号，p.52, 2010.

[川口 1999] 川口真一：勤務票システムの業務用件とシステム要件，*UNISYS TECHNOLOGY REVIEW*, Vol.63, pp.82-94, 1999

[川波 1998] 川波隆：イントラネットを核とした情報系コラボレーション，ユニシス・ニュース，Vol.444, p.3, 1998.

[喜連川 1997] 喜連川優：データマイニングにおける相関ルール抽出技法，人工知能学会誌，Vol.12, No.4, pp.513-520, 1997.

[キーマンズネット 2009] キーマンズネット：知識の横断検索！即解「企業内検索ツール」，http://www.keyman.or.jp/3w/prd/16/30002916/, 2009.

[木村 1999] 木村元，宮崎和光，小林重信：強化学習システムの設計指針，計測と制御，Vol.38, No.10, pp.1-6, 1999.

[栗原 2009] 栗原雅：情報の精査と分析・予測で"泥沼"を脱せよ，IT Leaders, Vol.2009-12, pp.22-23, 2009.

[小谷 1997] 小谷重徳：生産管理システム，オペレーションズ・リサーチ，Vol.42, No.2, pp.66-71, 1997.
[後藤 2010] 後藤和之，平博司，宮部泰成：企業の情報と知識の利活用を促進する対話型文書分類システム，東芝レビュー，Vol.65, No.2, pp.60-63, 2010.
[紺野 1998] 紺野登：知識資産の経営：企業を変える第５の資源，日本経済新聞社，1998.
[清水 2004] 清水勇喜，野中紀彦，西垣一朗，石川高司：開発設計のためのナレッジ活用型業務誘導システムの開発，日本機械学会年次大会講演論文集，pp.239-240, 2004.
[白石 2007] 白石弘幸:知識の共有と共用-応用地質の事例, 金沢大学経済学部論集, Vol.27, No.2, pp.129-148, 2007.
[鈴木 2006] 鈴木剛，前田薫，金崎克己，ハラルドホルツ，オーレグロスタニン，アンドレアスデンゲル：TaskNavigator：柔軟なワークフロー管理と適時情報配信，*Ricoh Technical Report*, No.32, pp.89-96, 2006.
[高野 2000] 高野明彦，西岡真吾，今一修，岩山真，丹羽芳樹，久光徹，藤尾正和，徳永健伸，奥村学，望月源，野本忠司：汎用連想検索エンジンの開発と大規模文書分析への応用，第 19 回 IPA 技術発表会，pp.383-384, 2000.
[高橋 2002] 高橋久尚，山下智志：大規模データによるデフォルト確率の推定－中小企業信用リスク情報データベースを用いて－，統計数理，Vol.50, No.2, pp.241-258, 2002.
[高村 2001] 高村大也，松本裕治：独立成分分析を用いた文書分類: SVM のための素性空間再構成，自然言語処理，Vol.143, No.3, pp.17-24, 2001.
[高村 2003] 高村大也，松本裕治：SVM を用いた文書分類と構成的帰納学習法，情報処理学会論文誌，Vol.44, SIG_3, (TOD_17), pp.1-10, 2003.
[田中 2004] 田中秀樹:使われるポータルサイト: ポータルソフトウェア開発を通じて，情報の科学と技術，Vol.54, No.8, pp.413-420, 2004.
[田村 2005] 田村泰彦，飯塚悦功，松川勇樹：工程設計のための不具合に関する知識の運用: 工程不具合の因果連鎖に関する知識構造の構築，品質，Vol.35, No.2, pp.95-113, 2005.
[特許庁 2007] 特許庁：平成 18 年度特許出願技術動向調査報告書ポストゲノム関連技術，2009.
[銅谷 2005] 銅谷賢治：教師あり学習，数理科学，No.507, pp.1-8, 2005.
[中山 1997] 中山康子，真鍋俊彦，竹林洋一：知識情報共有システム(Advice/Help on Demand)の開発と実践: 知識ベースとノウハウベースの構築，インタラクション'97, pp.1186-1194, 1997.
[日経 BP 企画 2005] 日経 BP 企画：実践 帳票マイグレーション―短期間&低リスク「帳

票」から始めるシステムオープン化，日経 BP 企画，2005.
[日経 BP 企画 2010] IT マーケットデータ年鑑 2010，日経 BP 社，2010.
[西村 2003] 西村実：ポストゲノムシーケンス時代のバイオ産業，日経産業新聞，2003.
[西村 2008] 西村健: ミドル・オフィスの生産性と情報・知識, 情報知識学会誌, Vol.18, No.5, pp.472-478, 2008.
[野中 2004] 野中紀彦，用田敏彦，針谷昌幸，横張孝志：プロセスとナレッジを融合した設計支援システムの開発，設計工学・システム部門講演会講演論文集，pp.292-293, 2004.
[野中 2008] 野中紀彦，清水勇喜，西垣一朗：設計プロセスを革新するナレッジベーストエンジニアリングの取り組み，日立評論，Vol.90, No.11, pp.906-909, 2008.
[野々口 2008] 野々口修次，松野二郎，荻野剛正：イントラネットポータルによる生産管理情報の共有，システム制御情報学会誌，Vol.52, No.4, pp.130-135, 2008.
[橋本 2003] 橋本文彦：テクニカル分析およびファンダメンタル要因分析による長期国債金利の動向予測，経済学雑誌，Vol.103, No.4, pp.1-11, 2003.
[日立ソフト 2008] 日立ソフトウェアエンジニアリング株式会社：ドキュメント統制・活用ソリューション『活文』，*Hitachi Soft REVIEW*, Vol.9, pp.6-7, 2008.
[日立ソフト 2010] 日立ソフトウェアエンジニアリング株式会社：ワークスタイル改革の実績を製品化『MEANS 紙文書電子化支援ソリューション』と『MEANS ファイルサーバスリム化ソリューション』，日立ソフトニュースレター，Vol.5, pp.1-3, 2010.
[富士通 2009] 富士通株式会社：ストレージ市場動向 第 3 回 データ量の増加への対応，http://storage-system.fujitsu.com/jp/news/sp/storage-market/vol03/, 2009.
[古川 2007] 古川直広，池田尚司，小西康介：デジタルペンを用いた研究ノートの開発，情報処理学会シンポジウム論文集, Vol.2007, No.4, pp.59-60, 2007.
[平本 2001] 平本純也 : 知っておきたいバーコード・二次元コードの知識，日本工業出版，第 5 版，2001.
[前田 2006] 前田慶太 : ビジネスコンテンツの戦略的活用に関する考察，*exa review*, No.7, pp.11-22, 2006.
[増山 2002] 増山繁，山本和英 : テキスト自動要約における新たな展開と展望，情報処理，Vol.43, No.12, pp.1310-1316, 2002.
[松野 2002] 松野成悟：企業間電子商取引と EDI の現状と課題 － アンケート調査による分析 －，宇部工業高等専門学校研究報告，Vol.48, pp.87-105, 2002.
[松本 2010A] 松本俊子，大峡光晴，小野山隆，薦田憲久：営業文書からのメタデータ自動抽出のための文書モデル自動生成技術，電気学会 情報システム研究会，IS-10-046,

pp.129-134, 2010.
[松本 2010B] 松本俊子，大峡光晴，小野山隆，薦田憲久：メタデータ抽出用パラメータの自動生成による導入容易な業務文書管理活用支援システム，平成 22 年度 電気学会 C 部門大会，TC15-3, pp.546-551, 2010.
[松本 2011A] 松本俊子，小野山隆，秋吉政徳：業務情報周知および活用を実現するビジネスレコメンデーション技術，電気学会 情報システム研究会，IS-11-041, pp.47-52, 2011.
[松本 2011B] 松本俊子，大峡光晴，小野山隆，秋吉政徳：ビジネス文書からのメタデータ抽出のためのルール自動生成技術，電気学会 C 部門論文誌，Vol.131, No.8, pp.1502-1511, 2011.
[松本 2011C] 松本俊子，小野山隆，秋吉政徳：業務情報周知のための業務遂行状況に応じた提示要否の判別方式，電気学会 C 部門論文誌，Vol.131, No.10, pp.1819-1827, 2011.
[水野 1993] 水野浩孝，青木由紀子，辻洋：データ可視化技法を用いた情報検索方式の提案，第 9 回ヒューマンインタフェース・シンポジウム，計測自動制御学会，pp.79-82, 1993.
[皆川 2009] 皆川明洋：帳票認識技術の応用と展開，信学技報，PRMU2008-219, pp.69-74, 2009.
[峰岸 2009] 峰岸達也，伊勢昌幸，新美礼彦，小西修：ロジスティック分析でのステップワイズ法と決定木による属性選択法の実データをもちいた比較，第 25 回ファジィシステムシンポジウム論文集，1A2-02, 2009 (in CD-ROM).
[宮川 2004] 宮川公男：経営情報システム，中央経済社，第 3 版，2004.
[村田 2004] 村田博士，小野田崇，由本勝久，中野幸夫，近藤修平：建物の外から電気機器の使用実態を把握するモニタリングシステム-実家庭への適用実験-，電気学会 C 部門論文誌，Vol.124, No.9, pp.1874-1880, 2004.
[森田 2000] 森田泰介：展望的記憶課題成績の規定因に関する調査的研究，日本教育心理学会総会論文集，No.42, p.596, 2000.
[薬師 2009] 薬師貴之，太田学，高須淳宏：CRF を用いた学術論文 OCR テキストからの自動書誌要素抽出，情報処理学会論文誌，データベース，Vol.2, No.2, pp.126-136, 2009.
[吉川 1990] 吉川武男：日英両国における原価計算システムの実態調査について，横浜経営研究，Vol.10, No.4, pp.407-429, 1990.