Signatureの局所的パターンマッチによる電子メールからの送信元住所録情報の抽出とそれを用いた住所録管理システム

浅野久子[†] 加藤恒昭[†,☆] 高木伸一郎[†]

電子メールには，送信者の姓名や電話番号等の住所録情報を持つsignatureとよばれる署名が付与される場合が多い．また，headerには送信者のメールアドレスが存在し，さらには姓名等の情報が存在する場合もある．これらの情報の有効利用を可能とするために，日本語で書かれた電子メールからheaderとsignatureを自動検出し，そこから送信元の住所録情報を抽出する方法を提案する．この抽出方法の主な特徴は，レイアウト情報に基づくsignature検出，および，signatureの局所構造を単位とした，属性の定型度順のパターンマッチによる住所録情報の抽出である．本方法を200通の電子メールを用いて評価した結果，signature検出の適合率96.3%，再現率86.1%，住所録情報抽出の全抽出対象属性に対する適合率93.4%，再現率88.7%が得られ，本検出・抽出方法の有効性が確認された．また，本抽出方法を利用した住所録管理システムについても述べる．本システムにより，ユーザは非定期に起こる住所録の変更の必要性やそのタイミングに留意している必要がなくなり，加えて，新規データ登録や更新も簡単なGUI操作のみで行える．さらに，更新した住所録情報を同僚等の住所録に通知することも可能である．これらにより，従来の住所録管理の負荷が大幅に軽減されることが期待できる．

Extraction of Sender Information from E-mails Based on Local Pattern Matching of Signatures and Its Application to Address Book Management

HISAKO ASANO,[†] TSUNEAKI KATO[†,☆] and SHINICHIRO TAKAGI[†]

E-mails often have signatures, which include sender information (e.g., name, telephone number, etc.). E-mail headers have the sender's e-mail address and generally sender name. For making good use of these data, we propose a method to find header and signature in Japanese e-mails and extract sender address book information. The main features of the method are signature separation using e-mail layout information, and sender information extraction based on pattern matching of the local structure of signatures. For 200 e-mails, the precision rate is 96.3%, and recall rate is 86.1% for signature separation; the precision and recall rates are 93.4%, 88.7% for sender information extraction respectively. These results show the proposed method is effective. As an application of the extraction method, we describe an address book management system. The users of this system need not care about the maintenance of their address book, and can use an easy GUI for data changes and new data registration. Furthermore, they can inform their colleagues about new data or changes. Therefore, the system is expected to simplify address book management.

1. はじめに

近年，電子メールを用いた情報伝達の機会が増大し，各種指示，連絡や挨拶等にも電子メールが利用されている．この電子メールには，末尾にsignatureとよばれる送信者の署名が付与される場合が多く，このsignatureには送信者の姓名や電話番号等の住所録情報が含まれていることが多い．また，電子メールの配送に必要な情報を持つheaderは，通常，送信者のメールアドレスを含み，さらに姓名等の情報を含む場合もある．受信する電子メールからこれらの送信元の住所録情報を自動的に抽出して住所録の維持更新に利用で

† NTT情報通信研究所
NTT Information and Communication Systems Laboratories
☆ 現在，NTTコミュニケーション科学研究所
Presently with NTT Communication Science Laboratories

きれば，非定期に起こる住所録の変更の必要性やそのタイミングをユーザが意識している必要がなくなり非常に有益である．しかしsignatureは必ず存在するわけではなく，その表現形式は多彩であり，含まれる情報も様々である．このため，電子メールからの住所録情報の抽出は単純なパターンマッチでは困難であり，具体的な手法の提案は行われていなかった．

本稿では，電子メールからの送信元住所録情報の抽出方法を提案する．最初にインターネット上の電子メールの特徴と，そこからの住所録情報抽出における課題を述べる（2章）．次に，日本語の電子メールを対象として，headerとsignatureを自動検出し，送信元の住所録情報を抽出する方法を提案する（3章）．主な特徴は，レイアウトに基づくsignature検出，および，signatureの局所構造を単位としたパターンマッチによる住所録情報の抽出である．そして，この抽出方法に対する評価と考察を述べる（4章）．さらに，提案した抽出方法を用いて到着した電子メールから住所録情報を抽出し，既存の住所録情報と差分がある場合，ユーザにそれを通知し，ユーザが簡単なGUI操作で新規登録や変更を行える住所録管理システムについて述べる（5章）．

2. 電子メールの特徴と住所録情報抽出における課題

2.1 電子メールのフォーマット

インターネット上の電子メール（以後，メールと略記する）はRFC☆822において，

- headerとbodyとよばれる2つの部分から構成され（bodyはなくてもよい），その境界は空行である．
- headerは複数のfieldから構成され，各fieldはfield-nameとfield-bodyから構成される．

とそのフォーマットが定められている．

headerのfieldは配送に必要な種々の情報であり，bodyは送信者が伝える内容そのものである．fieldのうちFrom fieldは送信者情報を表し，送信者のメールアドレスと，多くの場合，姓名の情報を持つ．日本語で書かれたメールにおいては，RFC822，RFC2047の規定を満たす形で，一般的に**表1**に示すFrom field-bodyのパターンが存在する．ここで姓名の部分は，アルファベット表記（例：Taro Yamada）されている場合と日本語表記（例：山田太郎）がMIMEエンコードされている場合とがある．

図1にインターネット上の電子メールの例を示す．

☆ インターネット標準のTCP/IP規格集

表1 From field-bodyの表現パターン
Table 1 From field-body pattern.

パターン	例（MIMEデコード後）
1）メールアドレス	yam@xx.jp
2）メールアドレス（姓名）	yam@xx.jp (Taro Yamada)
3）姓名〈メールアドレス〉	山田太郎〈yam@xx.jp〉

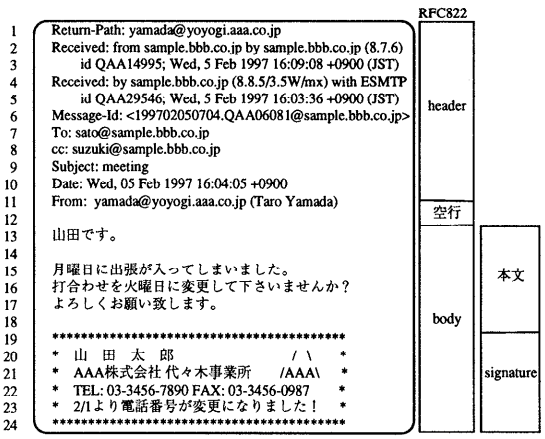


図1 電子メール例
Fig. 1 Example of e-mail.

2.2 signature

メールには，慣習的にbody末尾にsignatureとよばれる送信者の署名が含まれる場合が多い．以下では，bodyのsignature以外の部分を本文とよぶことにする（図1参照）．signatureには通常，送信者の氏名や電話番号，メールアドレス等の住所録情報が存在する．このsignatureの特徴をまとめると以下の5点となる．

(1) つねに存在するとは限らない．

(2) 文字本来の意味を持たない飾り専用の文字（飾り文字）が多用される．この飾り文字は記号の場合が多い（図1の外枠の“*”）が，漢字やアルファベットでさえ飾りのために使われる場合がある（図1第21行の会社ロゴの一部である“/AAA\”）．

(3) 含まれる情報の数やその位置・順序が定まっていない．また，変更通知（図1の“2/1より電話番号が変更になりました!”），格言，各種案内といった住所録情報以外の情報を含む場合もある．

(4) 全角文字と半角文字，正式名と略記名などの表記のゆれが存在する（“AAA株式会社”，“AAA（株）”）．

(5) レイアウトを考慮して，文字単位に分かち書きされる場合がある（図1の“山田太郎”）．

2.3 住所録情報抽出の課題

住所録情報の抽出は，テキストから特定の内容を抽出する情報抽出の一種である．しかし従来の情報抽

出では，対象テキストは主に新聞記事などに限定されることが多く[1),2)]，2.2節で示したような特徴を持つsignatureにそのまま適用するのは難しい．

また，会告記事用電子ニュースグループから会議のサマリーを抽出するシステム[3)]では，個条書きや行スタイル（センタリング行，左寄せ行等）といった行単位の表示上のスタイルを利用してタイトルや開催期日等の抽出を行うが，同一行内に姓名と会社名というように複数の情報が含まれる場合が多いsignatureには，行単位のスタイル情報のみでは不十分である．

住所録情報の抽出システムとしては，名刺読み取りシステム[4),5)]が実用化されているが，メールからの住所録情報抽出における特有の課題，

- signatureの存在有無を判定し，存在する場合にはその範囲を検出する必要があること
- signatureに含まれる情報とその配置や表現にバラエティが多く，それに適切に対処する必要があること

には対応していない．

3. 電子メールからの住所録情報抽出

3.1 header・signature検出

前章で述べたようにメールからの住所録情報の抽出では，抽出の対象領域となるheaderやsignatureの存在有無の判定や，その範囲の検出が必要である．headerはRFC822の規定に基づき，最初に現れる空行までを検出すればよい．しかし，signatureの表記に関する規定はない．そこで，signatureに関する以下の3つの特徴を利用したsignature検出方法を提案する．もちろんこれらの特徴を満たさないsignatureも存在する．その割合等については，次章の評価で論じる．

(1) signatureはbody末尾に存在し，“*”等の飾り文字からなる行または空行により本文と区切られる．
(2) signatureはメール1通につき0個または1個存在する．
(3) signatureは長くても十数行である．

signatureのこれらの特徴を利用するために，bodyの各行を**表2**に示す4種類に分類し，この行分類を並びをそのレイアウトの表現と考える．そして，それが一定のレイアウトのパターンとマッチするかでsignature検出を行う．より具体的には，body各行の行分類をアルファベットとし，それを末尾行から先頭方向に並べたものを文字列と見なし，これを**$body**と呼ぶ．ここで，**$body**はPerlで表現すると☆1以下のようなパターンとマッチすることになる．**$sig**がsignature，

表2 signature検出のための行分類
Table 2 Line classification for signature separation.

空行（N）	改行文字のみの行
飾り記号行（M）	空白文字☆2，記号☆3，長音のみからなる行
文末行（S）	行末が文末表現☆4となる行
デフォルト行（D）	上記以外の行

$textが本文末尾，**N***がメール末尾に含まれることのある無意味な空行に相当する．

```
$body =~ /^N*$sig$text/
```

これは上記の(1)と(2)の特徴を反映している．そして，signature検出のパターンとしては**$sig**と**$text**の対を与え，**$body**とのマッチングが成功した場合，そのマッチングの中で**$sig**と**$text**にマッチした文字数（行数）が一番少ないものを解とする．**$sig**に空列がマッチした場合はsignatureなしと判定される．一方，**$sig**にマッチした文字列の長さが一定値（signature最大行数として任意設定可能，実装評価時には15と設定）以上の場合は(3)の特徴から，これもsignatureなしと判定される．検出パターンは9種類用意した．以下2つのパターンについて，3つの電子メール例を用いて具体的に説明する．

signatureなしパターン1

```
$sig =~ /^$/, $text =~ /^S/
```

$sig**は空列を表現し，**$textは文末行を表現しているので，このパターンは末尾行が文末行のメールとマッチする．たとえば，**図2**の(1)では，末尾行の行末が“記号以外の文字+句点”であり文末行に分類されるので，このパターンを満たし，**$sig**は空列，つまり，signatureなしと判定される．

signatureありパターン1

```
$sig =~/^([DM].*M|[DM][DS]*N)?$/,
$text =~ /^[^M]{$TMAX}/
```

これは，飾り記号行（[DM].*M）から，存在しなければ末尾に最も近い空行（[DM][DS]*N）から，どちらも存在しなければメール末尾から，**$TMAX**行（仮先頭行最大行数）以上飾り記号行が存在しない場合に，飾り記号行，存在しなければ空行までをsignatureと判定，どちらも存在しなければsignatureなしと判定するパターンである．

このパターンにおいて，**$sig**の末尾（本文との境界

☆1 本稿で用いるパターン表記は，Perlの正規表現の表記法に若干の拡張を加えたものである．詳しくは付録A.1にまとめた．
☆2 スペースまたはタブ
☆3 漢字，カタカナ，ひらがな，アルファベット，数字，長音，空白文字以外の文字
☆4 句点で終わる等，付録の表9にその表記法を示した正規表現によって定義されている．

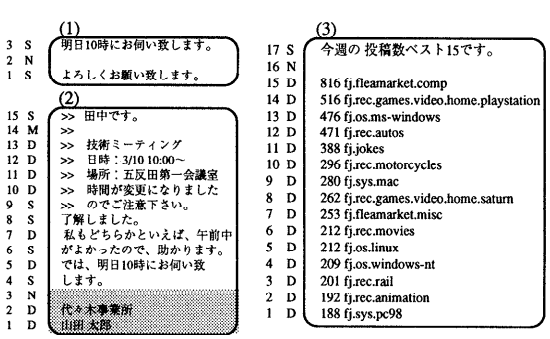


図2 行分類パターンによるsignature検出例
Fig. 2 Example of signature separation based on line classification patterns.

に相当する）が飾り記号行もしくは空行になっていることが前述の特徴(1)に対応している．また，この例は**$text**部分のマッチングの必要性も説明している．つまり，signature内には複数の飾り記号行，空行が存在する場合があるため，飾り記号行または空行が存在してもすぐにそれをsignatureの境界とはせず，飾り記号行または空行を検出した後，それ以降**$TMAX**行内に，境界として再設定すべき行がない場合にそれを境界とすることをこの**$text**によって表現しているのである．

このパターンによってsignatureが検出される例として，図2の(2)では，**$TMAX**=10の場合，末尾から第13行目までを処理した段階で，**$sig = 'DDN'**，**$text = 'SDSDSSDDDD'**となり，上記のパターンを満たす．そこで第1行から第3行までをsignatureとして検出する．

図2の(3)では，末尾から第1行〜第15行まですべてデフォルト行であり，末尾から10行を処理した段階で，**$sig = ''**（空列），**$text = 'DDDDDDDDDD'**となり上記パターンを満たす．そこでsignatureなしと判定される．

ここでは，2つのパターンについて記述した．他の7つのパターンは，引用やsignature内の文末行の存在等を扱うためのものとなっている．

3.2 住所録情報抽出

本節では，姓名，会社名，所属名（会社名より下位の組織名），役職，メールアドレス，郵便番号，住所，電話番号，FAX番号の9つの住所録属性を抽出する方法を提案する．ここで，住所録情報の抽出対象メールはsignatureを持つメールのみとする．

前章で述べたように，住所録情報の抽出においては，そこに含まれる情報とその配置にバラエティが多いというsignature独自の問題に対処しなければならない．

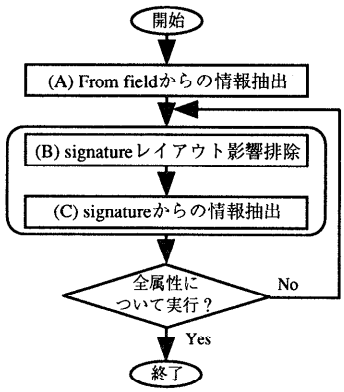


図3 抽出処理フロー
Fig. 3 Flow of extraction process.

表3 抽出プロセス
Table 3 Extraction process.

No.	プロセス名	役割
1	姓名・姓名読み候補抽出	(A)
2	メールアドレス抽出(1)	(A)
3	スコープ設定	(B)
4	メールアドレス抽出(2)	(C)
5	連続飾り文字の削除	(B)
6	スコープ再設定	(B)
7	電話・FAX番号抽出	(C)
8	郵便番号・住所抽出	(C)
9	スコープ再設定	(B)
10	姓名抽出	(C)
11	スコープ再設定	(B)
12	会社名・所属・役職抽出	(C)
13	スコープ再設定	(B)
14	住所抽出(2)	(C)
15	備考データ作成	(C)

提案方式ではこの問題に次のような方針で対処した．

- 定型度が高い等，高い信頼度で抽出できる属性（メールアドレスや電話番号）を先に抽出し，それを抽出が難しい属性（姓名や会社名）の多義の解消，縮小に利用して全体の抽出精度を高める．
- signatureのレイアウトの影響を排除するために，その全体を情報抽出の対象とするのではなく，いくつかの抽出単位に分割して，それぞれに対して局所的なパターンマッチを行うことで情報抽出する．この抽出単位をスコープとよぶ．

この方針に基づき，3つの役割を持った処理から構成される**図3**のフローにより抽出を行う．まず，フォーマットが規定されており（表1），最も信頼性が高いFrom fieldから住所録情報を抽出する(A)．次にsignatureを対象に，定型度が高い属性から順に，抽出対象領域と抽出単位，つまりスコープを設定し(B)，その属性に関する情報を抽出する(C)という処理を繰り返す．**表3**に具体的な処理順序を示す．ここに示した

情報の抽出順序が定型度を反映している．以降では，各役割別にその内容を説明する．

3.2.1 From field からの住所録情報抽出

プロセス1の姓名・姓名読み候補抽出では，姓名または姓名読み候補を抽出する．表1の2), 3)において「姓名」が漢字かな表記の場合には，それを「姓名」として抽出，アルファベット表記の場合には，ローマ字カナ変換を行い，変換後のカナを姓名読み候補とする（例：YAMADA → ヤマダ）．姓名読み候補は，プロセス10の姓名抽出で利用する．

プロセス2のメールアドレス抽出(1)では，表1の「メールアドレス」部分を抽出する．

3.2.2 signature のレイアウト影響排除

signature においてそれを整形・装飾するレイアウト表現としてよく用いられるのは，2.2節の signature の特徴における(2)飾り文字と(5)文字単位の分かち書きである．住所録情報を抽出する場合，(2)飾り文字はそれらを誤抽出する可能性があり，(5)文字単位の分かち書きにより抽出パターンにマッチしない可能性が出てくるため，それぞれ対処を行う必要がある．ここで飾り文字は，罫線のようにつねに連続して現れるわけではないことが問題を難しくしている．以下，本提案方式で行った連続する飾り文字への対処，不連続な飾り文字への対処，文字単位の分かち書きへの対処について述べる．

連続する飾り文字への対処 連続する飾り文字を削除するために，signature の縦または横に3文字以上連続する情報記号☆1以外の記号，または5文字以上連続する空白文字以外の同一文字を連続飾り文字として処理済文字☆2とする（プロセス5）．

不連続な飾り文字への対処 不連続な飾り文字を削除することは，それで区切られていない連続した文字列を得ることである．この文字列をスコープとよぶ．1つのスコープには1つ以上の情報（住所録情報や変更通知などのその他の情報）が含まれており，一般にはある住所録情報が複数のスコープにまたがっていることはないので☆3，この単位で局所的なパターンマッチを行うことで情報抽出が可能である．以下，スコープの区切りとなる，通常，住所録情報に含まれない文字を「スコープ区切り文字」と呼ぶ．ここで，住所録の各属性ごとに含まれる文字が異なるため，属性別にスコープ区切り文字を動的に変更する．たとえば，姓名が記号や数字を含むことはありえないが，住所や電話番号は，数字や“-”等を含む．そこで，姓名の抽出の際には記号や数字をスコープ区切り文字とするが，住所の抽出では，それらをスコープ区切り文字としない．スコープ設定は，プロセス3, 6, 9, 11, 13の5プロセスで行われる．これらの各プロセスにおけるスコープ区切り文字を**表4**に示す．

表4 スコープ区切り文字
Table 4 Scope separation characters.

No	スコープ区切り文字
3	改行文字
6	基本スコープ区切り文字☆4
9	基本スコープ区切り文字，数字，アルファベット，情報記号
11	基本スコープ区切り文字，情報記号☆5
13	基本スコープ区切り文字

文字単位の分かち書きへの対処 空白文字（の連続）で区切られた，漢字，ひらがな，カタカナ，アルファベット1文字からなるスコープが2つ以上連続して存在した場合には，その連続するスコープを統合して1スコープとする．このスコープ統合により，文字単位に分かち書きされている文字列を，1単位として扱えるようになる．

3.2.3 signature からの住所録情報抽出

住所録情報のうち，“姓名”は姓名辞書☆6，単漢字辞書☆7を用いた辞書マッチにより，その他の属性はパターンマッチにより抽出を行う．これは，

- 電子メールでは姓名の読み情報が得られる場合が多い（プロセス1：姓名読み候補）ので，読みをキーとした辞書検索が非常に有効である．
- 会社名，所属名等は膨大な種類があり，かつ増え続けており，signature では表記のゆれも存在するため，データベースの維持管理が困難である．

という2つの理由からである．

次に，パターンマッチおよび辞書マッチによる抽出についてそれぞれ説明する．

パターンマッチによる抽出 パターンマッチに用いる表現をキーワードとよぶ．キーワードは，任意文字指定などが記述できる任意設定可能な正規表現であり，表記のゆれに対処するために，1バイト（半角）文字と2バイト（全角）文字，大文字と小文字を区別しな

☆1 住所録情報の一部となりうる記号．任意設定可能．実装評価時は“・, -,',(,), 〒”の6文字を設定．
☆2 ある抽出処理でマッチし，以降では処理対象外とされる文字．
☆3 ただし，特定の属性においては，抽出する属性単位より小さい単位で1スコープを形成することがありうる．たとえば，姓名の姓と名，所属の部と課が隣接する2スコープからなることがある．
☆4 情報記号以外の記号または空白文字または処理済文字
☆5 会社名，所属名，役職キーワードに用いられている文字は除く．
☆6 表記，読み，姓名の区別を持つ．(例) 山田，ヤマダ，姓
☆7 漢字1文字ごとの表記，読みを持つ．(例) 部：ブ，ベ，へ

い．キーワードの表記法，および具体的な表現は付録の表9，表10に示す．キーワードは14種類存在し，属性値の（部分）文字列にマッチする属性値キーワードと，属性名の（部分）文字列にマッチする属性名キーワードに分けられる．必要な情報は，属性値キーワードにマッチする部分（たとえば“yamada@yoyogi.aaa.co.jp”）であるが，それだけでなく属性名キーワード（たとえば“E-?MAIL”）によるマッチを併用しているのは，次の2つの理由による．

- 他の住所録情報，あるいは住所録以外の情報ともマッチする可能性がある属性値キーワードは，対応する属性名キーワード，あるいは関連する属性値キーワードとの共起を利用して，誤抽出を防止する．たとえば“番号パターン”（属性値）は，“電話”（属性名）と“FAX”（属性名）のうち共起した方の属性（どちらとも共起しない場合は電話番号）として抽出する．また内線番号は，電話番号が抽出された場合に限り，その同一行後方で“内線”（属性名）と“内線パターン”（属性値）が共起した場合のみ抽出する．
- 属性値に加えて属性名も処理済文字とし，以後の抽出処理対象から除くことで，後続の抽出処理における処理対象スコープの削減，誤抽出防止となる．

各属性ごとの具体的な抽出条件等は付録A.2に示す．

辞書マッチによる抽出 プロセス10の姓名抽出では，プロセス1で姓名が抽出されていない場合に，全未処理スコープを対象としてまず姓名辞書，次に単漢字辞書とのマッチングを行う．さらに辞書より得られた読みと姓名読み候補のマッチングを行い，姓名の抽出を行う☆．姓名読み候補は，複数の姓名が抽出された場合の絞り込み，会社名等の誤抽出を防ぐのに有効である．

備考データの作成 姓名読み候補，signature未処理スコープを「備考」データとする．

3.2.4 抽出処理例

図1の電子メールを例に抽出処理の具体的な流れを**図4**に示す．プロセス1により，From fieldのアルファベット表記の姓名から“ヤマダ”と“タロ”という姓名読み候補が得られる．プロセス2により，From fieldのメールアドレス“yamada@yoyogi.aaa.co.jp”が抽出される．次に処理がsignatureに移る．プロセス5により外枠を表す“*”が削除される．プロセス6によって，図に示す10のスコープが設定される．ここで[山田太郎]は1文字からなる4つの連続するスコープが1つに統合されて得られるスコープである．プロセス7では，設定されたスコープのうち，網掛けのスコープが属性値である“番号パターン”，斜線のスコープが属性名である“電話”および“FAX”キーワードにマッチするため，図に示す電話，FAX番号が抽出される．プロセス9でスコープの再設定が行われ，プロセス10の姓名辞書のマッチングにおいて“山田”，“代々木”が姓，“太郎”が名にマッチする．ここで読み候補とマッチし，姓と名がペアで抽出されているため，“山田太郎”が姓名として抽出される．プロセス11で再びスコープ設定が行われ，プロセス12で斜線の文字列が“会社名”，網掛けの文字列が“組織名”にマッチするので，図に示す会社名と所属名が抽出される．

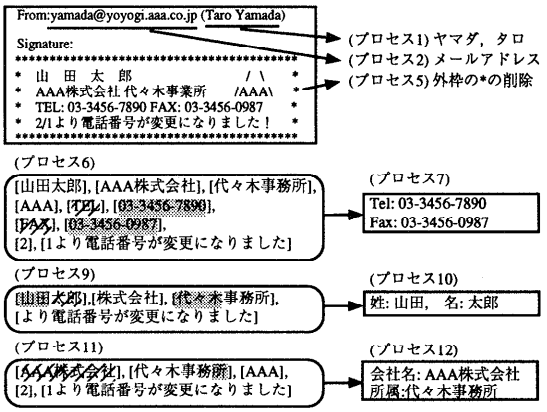


図4 抽出処理例
Fig. 4 Extraction example.

4. signature検出・住所録情報抽出の評価と考察

4.1 評価データ

実際に送受信された，送信者がすべて異なる日本語メール200通（signatureあり：180通，signatureなし：20通）を収集し，評価対象とした．日本語メールとは，bodyに少なくとも1文以上の日本語文を含むものである．この評価用メールはsignature検出・住所録情報抽出ルール作成には利用していない，オープンデータである．

signatureを持つメール180通において，signatureまたはheaderに住所録情報が存在するメール数（属性別），また参考として，飾り文字のあるsignature数，住所録情報以外の情報を持つsignature数を**表5**に示す．ここで日本語存在数とは，漢字（かな）表記が存在する（英語と漢字（かな）併記を含む）メール数，英語存在数とは英語表記しか存在しないメール数

☆ 姓名読み候補が存在しない場合には姓名辞書のみを用いて表記のみのマッチングにより抽出する．

表5 評価対象メールの属性存在数
Table 5 Feature count in e-mails evaluated.

属性名	日本語存在数	英語存在数
会社名	131	17
所属名	122	5
会社名+所属名	138	17
会社名+所属名（部以上）	137	17
役職	4	2
姓名	155	17
郵便番号	43	8
住所	45	8
メールアドレス	180	—
電話	128	1
FAX	118	1
飾り文字	140	—
他情報	18	4

表6 signature 検出結果
Table 6 Signature separation results.

(1)	signature 有	正解	155
(2)		境界誤り	3
(3)		未検出	22
(4)	signature 無	正解 (signature 検出せず)	17
(5)		誤検出	3
(6)	適合率	(1)/((1)+(2)+(5))	96.3%
(7)	再現率	(1)/signature 存在数	86.1%

を表す．また，数字表記である郵便，電話，FAX 番号では，英語表記の住所内に含まれる郵便番号，日本国外の番号を表す電話，FAX 番号を英語表記として扱った．姓名とメールアドレスについては，header の From field に存在した場合にもカウントした．

4.2 signature 検出の評価と考察

表6 に signature の検出結果を示す．適合率 96.3%，再現率 86.1%と高精度で検出でき，行分類の組合せパターンによる検出が有効であることが確認できた．

signature の境界を誤った（表6(2)）原因は，本文中に句読点がついていないため文末行を認定できなかったこと（2件）と，signature が 30 行と長すぎ，その途中までを signature と検出したこと（1件）であった．

signature 未検出（表6(3)）の原因のうち延べ2件以上存在したものは，signature の後に添付書類がつく，signature が 20 行を超える，signature が本文直後に飾り記号行や空行を持たない右揃えのシンプルなものであった等，本検出方法が前提とした特徴を満たさないメールであったこと（9件），signature 内で文末表現ではないのに文末行と誤判定したこと（4件），signature 内に複数の文末表現が存在したこと（2件）であった．

また，signature なしのメールを signature ありと誤検出した（表6(5)）原因は1つで，本文末尾が表形式等のために行末に句読点が存在せずデフォルト行となったため（3件）であった．

未検出，誤検出した原因のうち，signature 内の文末行の誤判定は文末行パターンの修正により，また，右揃え signature に対しては行分類に右揃え行を追加することにより対応可能である．しかし，本文末尾が表形式等のため句読点が存在しない場合には，行分類に基づく検出ルールでは signature かどうか区別できない．また，現在 signature 内には文末行を1行しか許していないため，複数の文末行を含む signature は検出できない．複数の文末行を許すと，本文を signature と誤検出する可能性が高まる．

4.3 住所録情報抽出の評価と考察

表7 に住所録情報抽出の評価結果を示す．ここで「signature 検出正解」は，住所録情報抽出に対する評価であり，signature を正しく検出したメールのみ（155 通）を対象とした．一方,「総合」はすべてのメール（200 通）を対象としており，signature 検出と住所録情報抽出をあわせた総合評価となる．また，漢字（かな）表記のみを抽出対象としているので，適合率，再現率とも日本語存在数を基準とした．

ところで現在の抽出方法では，会社名と所属名が同一スコープに存在した場合に，会社名と所属名に分けるルールが存在しない．そこで，今回は会社名+所属名を合わせたものを1属性と扱い，評価した．また signature には，担当レベルまでの詳細な所属の情報が含まれる場合があるが，住所録情報としては部（会社）や科（学校）レベルの情報で十分である場合が多いので，部，科以上の会社名+所属名に対する抽出精度評価も行った．さらに住所録の更新支援としては，属性値の一部が抽出されていれば，その分，入力負荷は軽減されるので，一部が未抽出であったもの（例：住所でビル名のみ未抽出等）も正解に含めた抽出精度も算出した．表7において，部分抽出の NG 欄は部分抽出を不正解としたもの，OK 欄は正解としたものである．

表7より，姓名，郵便番号，メールアドレス，電話番号，FAX 番号といった定型度の高い属性に対しては，適合率，再現率とも 90%以上と高精度に抽出でき，キーワードに基づく抽出が有効であることを確認した．しかし，様々な表現を持つ会社名と所属名に関しては，部分未抽出の割合が高く，抽出精度もやや低い．

誤抽出・未抽出の主な原因と属性別の内訳を**表8** に示す．

誤抽出の第1原因は，抽出ルール条件が緩く他情報

表7 住所録情報抽出評価結果
Table 7 Extraction results of sender address book information.

	signature検出正解				総合			
	適合率		再現率		適合率		再現率	
部分抽出	NG	OK	NG	OK	NG	OK	NG	OK
会社名+所属名	68.6	94.1	58.8	80.7	66.7	91.4	50.7	69.6
会社名+所属名（部以上）	85.3	94.1	73.7	81.4	82.9	91.4	63.5	70.1
役職	50.0	100	25.0	50.0	50.0	100	25.0	50.0
姓名	92.6	95.6	91.3	94.2	91.3	94.2	81.3	83.9
郵便番号	100	—	94.9	—	100	—	86.1	—
住所	92.1	97.4	87.5	92.5	92.1	97.4	77.8	82.2
メールアドレス	100	—	100	—	100	—	100	—
電話	100	—	94.0	—	100	—	85.3	—
FAX	99.0	—	94.4	—	99.0	—	85.6	—
全属性総合	93.4	98.0	88.7	93.1	92.7	97.3	81.3	85.3

表8 誤抽出・未抽出原因内訳
Table 8 Causes of extraction error and extraction failure.

		会+所	役職	姓名	郵便	住所	メール	電話	FAX	比率
誤抽出原因	他情報抽出	1		3					1	35.7%
	From field 抽出			3						21.4%
	属性未抽出	2				1				21.4%
未抽出原因	キーワードアンマッチ	40	2						1	60.6%
	抽出ルールなし			1	1	1		7	4	19.7%
	他属性として抽出	2	1	1					1	7.0%
	表記ゆれ・ミス	1		1		2				5.6%

を抽出したこととである．たとえば，珍しい姓名のためsignature内に姓名とそのふりがなが存在したメールで，ふりがなを抽出したものがあった．第2，第3原因は，From fieldに送信者の姓名以外の情報（会社名等）が記入されておりその情報を抽出したことと，本来の属性で未抽出となった部分が他の属性として誤抽出されたことである．たとえば，所属を所属として抽出できなかったために，住所（ビル名）として抽出したものが存在した．

未抽出の第1原因はキーワードにマッチしないことで，特に会社名+所属において顕著である．この一部については，キーワードの追加で抽出可能となる．たとえば役職抽出において，「所属名キーワード+“長”」という役職キーワードにより，「○○部門担当課長」というスコープから「課長」が役職として抽出された．これを正しく「担当課長」として抽出するためには，役職キーワードに「“担当”+所属名キーワード+“長”」を追加すればよい．しかし，「○○インターナショナル」という会社名のみが存在し，会社名キーワードにマッチせず，また所属名も存在しないために未抽出となった例では，追加すべきキーワードとなりうる文字列が存在しない．また，正式名称の「○○総合研究所」ならばキーワードにマッチしたが，「○総研」のように略記表現されたためにマッチしない表現も存在した．

第2原因は抽出ルール不足である．たとえば，「TEL&FAX 03-1111-1111」のように，電話番号とFAX番号が同一である場合に対応していなかったため，電話またはFAXいずれか一方としてしか抽出できなかった．これらは抽出ルールを追加または変更することにより抽出可能となる．

第3原因は，他属性として誤抽出されたため，抽出対象とならなかったことである．たとえば，会社名を表す文字列が姓名として抽出されたために，会社名としては未抽出になった場合があった．

第4原因は，表記のミスとゆれである．たとえば，本来は「○○株式会社」であるものが，「社」が欠落し「○○株式会」と表現されていたり，住所で番地が省略されているものが存在した．

5. 住所録管理システム

3章で述べた電子メールからの住所録情報抽出方法を用いて，住所録管理システム[6]を構築した．住所録管理システムは，ユーザが事前に住所録情報の抽出対象とするメールの条件を設定しておくだけで，各メール到着時に，条件を満たすメールであるかを判定し，条件を満たす場合には住所録情報を自動的に抽出し，既存の住所録情報との差分が存在する場合にはその差分情報を保存しておく．そして，ユーザが指定したタ

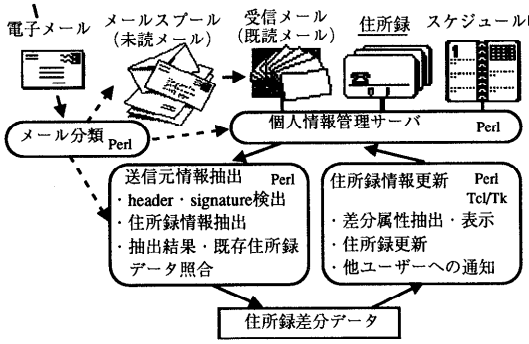


図5 電子メールを利用した住所録管理システム
Fig. 5 Address book management system for e-mail use.

イミング（ログイン時や特定時刻）での自動起動，あるいはユーザ自身の起動により，その差分情報をユーザに提示し；住所録の更新を簡単なGUI操作だけで行えるようになる．また，更新した住所録情報を特定の人やグループに通知する機能も持つ．

このシステム構成を図5に示す．本システムはPerlとTcl/Tkにより実装されSunOS4.1.4上で動作する．以下に各処理の概要を説明する．

個人情報管理サーバ[7]は，受信メール，住所録，スケジュール帳等のすべての個人情報を共通的な形式で管理するサーバであり，各利用者ごとに起動される．他ユーザの個人情報管理サーバへメールを用いてアクセスすることも可能である．住所録へのアクセスは，この個人情報管理サーバを介して行う．

メール分類は，headerの各fieldやbodyに対する特定文字列の有無条件に従って，メールを任意のプログラムに引き渡すことができる．これにより「To fieldに自分のメールアドレスが存在する場合のみ住所録情報の抽出を行う（メーリングリスト宛やCcされているメールは処理しない）」といった，ユーザの必要に応じた条件設定が可能となる．また，個人情報管理サーバ用のメールを個人情報管理サーバに引き渡す役割も持つ．

送信元情報抽出では，3章で述べたheader・signature検出を行い，住所録情報を抽出する．そして，抽出した住所録情報と既存住所録データとの差分を照合し，差分が存在した場合には住所録差分データに書き出す．

住所録情報更新では，既存住所録と住所録差分データで差分のある属性を検出し，ユーザに提示する．ユーザはGUIを用いて更新処理を行う．画面では既存住所録データと異なる情報を持つ属性を色分けして表示するので，ユーザは簡単に変更部分の確認，場合によっては修正をして，住所録の更新を行える．また通知機能により，他ユーザへ更新した住所録情報を知らせることができる．

6. おわりに

電子メールからheaderとsignatureを検出し，住所録情報を抽出する方法を提案した．そして，定型度の高い電話番号等の属性は90%以上の適合率，再現率で情報を抽出することを確認した．また，この住所録情報抽出法を用いた住所録管理システムを構築し，ユーザが非定期に起こる住所録の変更の必要性やそのタイミングを留意している必要なく，簡単なGUI操作のみで新規データの登録や更新，他ユーザへの通知を行うことを可能とした．これにより従来の住所録管理の負荷が大幅に軽減されることが期待できる．

今後は，本抽出方式では維持管理の煩雑さから各種の辞書は最小限の利用にとどめているが，メールアドレスのドメインと会社名等を対応づけたドメイン辞書の導入など，これらを積極的に用いていく方式との比較検討も行っていく予定である．

参考文献

1) 松尾比呂志，木本晴夫：抽出パターンの階層的照合に基づく日本語テキストからの内容抽出方法，情報処理学会論文誌，Vol.36, No.8, pp.1838–1844 (1995).
2) 江里口善生，木谷　強：パターンマッチング手法による名称特定処理の有効性の検討，情報処理学会研究報告，NL115-10, pp.9–15 (1996).
3) 佐藤　円，佐藤理史，篠田陽一：電子ニュースのダイジェスト自動生成，情報処理学会論文誌，Vol.36, No.10, pp.2371–2379 (1995).
4) 斎鹿尚史，中村安久，北村義弘，森田敏昭：名刺読み取りシステム，電子情報通信学会技術報告，NLC93-23, pp.9–16 (1993).
5) Saiga, H., Nakamura, Y., Kitamura, Y. and Morita, T.: An OCR system for business cards, *Proc. 2nd International Conference on Document Analysis and Recognition*, pp.802–805 (1993).
6) 浅野久子，加藤恒昭，高木伸一郎：電子メールを利用した住所録管理システム，第55回情報処理学会全国大会論文集，Vol.4, pp.327–328 (1997).
7) 加藤恒昭，浅野久子，中野有紀子：情報交換に電子メールを用いる個人情報サーバの実現，第55回情報処理学会全国大会論文集，Vol.3, pp.302–303 (1997).

表9 正規表現表記法
Table 9 Notation of regular expressions.

表記	意味
[···]	··· 中の任意の1文字
[^···]	··· 中にない任意の1文字
<···>	··· 中の任意の文字種☆ 1文字
<^···>	··· 中にない任意の文字種☆ 1文字
(···)	グループ
x-*y*	文字 *x* から *y* までの任意の1文字
.	任意の1文字
*	0回以上の繰り返し
+	1回以上の繰り返し
?	0回または1回の繰り返し
{*n*,*m*}	*n*回以上 *m*回以下の繰り返し
{*n*}	*n*回の繰り返し
|	選択
^*r*	先頭が正規表現 *r*
r$	末尾が正規表現 *r*
\	エスケープシーケンス
x	上記以外の任意の文字 *x* 1文字

☆ C：漢字，K：カタカナ，J：ひらがな，N：数字，A：英字，S：スペース，T：タブ，D：漢数字，P：句読点文字（任意設定可能），I：情報記号（任意設定可能），V：長音，M：記号

表10 キーワード
Table 10 Keywords.

キーワード名	種類	キーワード例	マッチ例
メール (ML)	名	E-?MAIL, 電子メ[一イ]ル	E-Mail, 電子メール
メールパターン (MP)	値	[A-Z0-9_\- \.] +@[A-Z0-9_\-\.]+\.jp	foo@test.jp
電話 (TL)	名	TEL[A-Z]*	telephone, TEL
電話副 (TS)	名	\(直通 \),\(代表 \)	(直通), (代表)
内線 (EX)	名	内線, EXT (ENSION| \.)?	内線, Ext.
FAX (FX)	名	FAX, ファッ?クス	Fax, ファクス
番号パターン (NP)	値	03<MVS>{1,2}<N>{4} <MVS>{1,2}<N>{4}	03-1111-1111
内線パターン (EP)	値	[0-9]{4,4}	1234
郵便番号 (ZI)	名	〒, 郵便番号	〒, 郵便番号
郵便番号パターン (ZP)	値	[0-9]{3,3}(-[0-9]{2,2})?	123-45
市町村 (CT)	値	市, 区, 郡, 町, 村	千代田区 1-1
会社名 (OZ)	値	<CJHA>+(株)	AAA（株）
所属名 (DV)	値	部, 課, 科	営業部
役職 (OT)	値	部長	部長

表11 キーワードパターンマッチのパラメータ
Table 11 Parameters of keyword pattern match.

No.	exec-cond	target	keyword	feature	value	comp-stg
4-1	—	S(all)	*MP*	MAIL	M(*MP*)	M(*MP*)
4-2	*MP*	S(*MP*, cp)	*ML*	MAIL	—	M(*ML*)
7-1	—	S(all)	*NP*	TEL, FAX	M(*NP*)	M(*NP*)
7-2	*NP*	S(*MP*, cp|lp)	*TL*, *FX*	—	—	S(*TL*), S(*FX*)
7-3	*NP*	S(*NP*, cn|ln)	*EX* & *EP*	EXTN	M(*EP*)	S(*EP*), S(*EX*)
8-1	—	S(all)	*ZI* & *ZP*	ZIP	M(*ZP*)	S(*ZI*), S(*ZP*)
8-2	—	S(all)	*CT*	ADD	S(*CT*)	S(*CT*)
8-3	*CT* & ! *ZP*	S(*CT*, cp|lp)	*ZP*	ZIP	M(*ZP*)	S(*ZP*)
12-1	—	S(all)	*OZ*	ORG DIV	M(*OZ*) S(*OZ*, cn)	S(*OZ*)
12-2	*OZ*	S(all)	*DV*	DIV	S(*DV*)	S(*DV*)
12-3	! *OZ*	S(all)	*DV*	ORG DIV	S(*DV*, p+t, s) S(*DV*, p, s)	S(*DV*, lp, s) S(*DV*)
12-4	*OZ*	S(*OZ*, n, s)	<CKJ>+	DIV	S(*OZ*, n, s)	S(*OZ*, n, s)
12-5	—	S(*OZ*), S(*DV*), S(all)	*OT*	TITLE	M(*OT*)	M(*OT*)
14-1	*CT*	S(*CT*, nn)	—	ADD	S(*CT*, nn)	S(*CT*, nn)

M(***keyword***)：***keyword***にマッチした文字列.
S(***keyword***)：***keyword***にマッチしたスコープ．ただし ***keyword***=all の場合は，すべての未処理スコープ.
S(***keyword***, pos)：M(***keyword***) を基準にして位置 pos にあるスコープ.
S(***keyword***, pos, s)：S(***keyword***, pos) のうち，スペース，改行，連続飾り文字（プロセス5）からなるスコープ区切り文字で区切られるスコープ.
pos の値=p：前方，n：後方，lp：同一行前方，ln：同一行後方，nn：直後スコープ，
cp：同一スコープ前方，cn：同一スコープ後方，+t：条件を満たす最遠スコープ.
!：否定. &：論理積. |：論理和.
exec-cond の ***keyword***はそれが以前にマッチしている/していない（否定（!）の場合）ときに真である.
7-1 の feature は，7-2 で M(*FAX*) がマッチした場合には FAX，それ以外の場合には TEL となる.

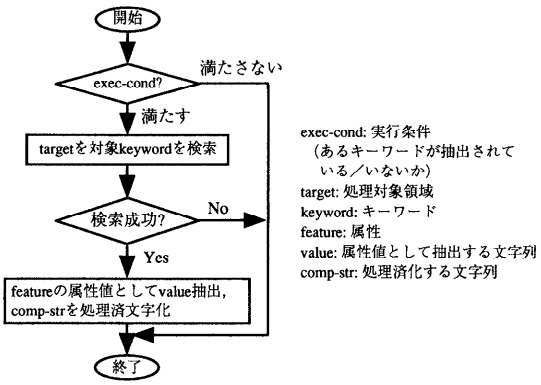


図6 キーワードパターンマッチの抽出フロー
Fig. 6 Extraction flow of keyword pattern match.

付　　録

A.1 正規表現表記法

3.1 節で述べた文末表現と行分類の signature 検出パターン，3.2.3 項で述べたキーワードを記述する正規表現の表記法を**表 9** に示す．この表記法は，基本的には Perl の正規表現と同一であるが，<…>，<^…> については拡張している．また，大文字と小文字，全角文字と半角文字を区別せずに扱える．

A.2 パターンマッチによる住所録情報抽出

ここでは，3.2.3 項のスコープを対象とした☆キーワードのパターンマッチによる抽出の詳細を述べる．**表 10** に，キーワードとその種類（属性名/属性値），キーワード表現の例とそれにマッチする文字列の例を示す．

キーワードのパターンマッチによる抽出は，**表 11** に処理順に示したプロセス No. の 1 つ 1 つに対して，それぞれのパラメータを用いて**図 6** のフローにより行う．

ここで，プロセス 4 のメールアドレス抽出 (2) に対して，“E-mail yamada@yoyogi.aaa.co.jp”（1 スコープ）を例に抽出の流れを説明する．表 11 のプロセス No.4-1 において，**target**=S(all) であるので，全未処理スコープを対象に，**keyword**=*MP*，メールパターンキーワードを検索する．ここで “yamada@yoyogi.aaa.co.jp” がメールパターンキーワードにマッチし，メールアドレスとして抽出，処理済化される．次にプロセス No.4-2 において，実行条件 **exec-cond**=*MP* は，メールパターンが抽出されており満たされているので，**target**=S(*MP*, cp)，すなわち，メールパターンにマッチした同一スコープ前方を対象に，**keyword**=*ML*，メールキーワードを検索する．ここで，“E-mail” が属性名として抽出され，処理済化される．

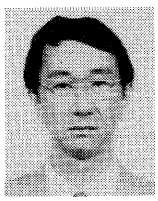
このように，属性名キーワードと属性値キーワードのマッチは完全に独立に行われるわけではなく，必要に応じてそれぞれの結果を利用することで精度を上げるようにしている．また，属性名，属性値とも処理済化することで以降の処理対象を削減している．

(平成 9 年 5 月 23 日受付)
(平成 10 年 4 月 3 日採録)

☆ 番号パターンマッチにおいては，例外的にスコープ区切り文字であるスペースを抽出対象に含める．

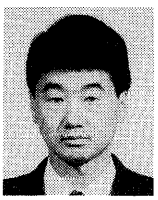
浅野　久子（正会員）

1991 年横浜国立大学工学部電子情報工学科卒業．同年，日本電信電話（株）入社．自然言語処理，テキスト処理の研究に従事．現在，NTT 情報通信研究所知的通信処理研究部勤務．言語処理学会会員．



加藤　恒昭（正会員）

1959 年生．1981 年東京工業大学電気電子工学科卒業．1983 年同大学総合理工学研究科電子システム専攻修士課程修了．同年，日本電信電話公社（現 NTT）横須賀電気通信研究所に入所．自然言語処理，対話処理，マルチモーダルコミュニケーションに関する研究に従事．1998 年 4 月より，NTT コミュニケーション科学研究所知識処理研究部主幹研究員．工学博士．電子情報通信学会，人工知能学会，言語処理学会，ACL 各会員．



高木伸一郎（正会員）

1979 年金沢大学工学部電気工学科卒業．1981 年同大大学院修士課程修了．同年，日本電信電話公社（現 NTT）横須賀電気通信研究所に入所．以来，日本語形態素解析を応用した日本文校正支援システムの研究開発を経て，現在，合成音声による日本文読み上げ技術を応用した知的支援サービスの開発に従事．現在，NTT 情報通信研究所知的通信処理研究部主幹研究員．電子情報通信学会会員．