自 然 言 語 処 理 101－8
（1994. 5. 27）

電子メール文書からの関係情報の自動抽出

河合敦夫　塚本雄之　椎野努

三重大学　工学部

文書中から、あらかじめ目的とした情報のみを取り出してくる内容抽出について述べている。従来の内容抽出研究は、印刷出版物の文からであるのに対し、本報告で対象とする電子メールでは、表、箇条書き、単語の羅列といった文以外の構成要素を考慮すべき点に特徴がある。電子メールの文書構造を、表部分（文字等の物理的（２次元的）配置が意味を持つ）と、文章（物理的配置が意味を持たない。文のみから構成される）に分類し、文字、品詞、および単語の持つ意味属性の物理的配置を利用して表部分を認識している。表部分からの内容抽出の手がかりとして、単語の字面、意味属性、文字種別、表部分の認識結果（同一ブロック内にある抽出情報の関連づけ）がある。現在、中古品売買情報の電子メール文書に適用して評価を行うとともに、処理系の開発を行っている。

Information Extraction from Electronic Mail

Atsuo Kawai, Takeyuki Tsukamoto, Tsutomu Shiino

Faculty of Engineering, Mie University
1515, Kamihama-cho, Tsu 514 Japan

This paper describes about information extraction from the documents of electronic mail. We classified the document structure of electronic mail into 2 parts: chart parts and sentence parts. Chart parts of electronic mail are recognized by using the physical (2 dimensional) arrangement of letters, parts of speech, and semantic feature of noun. The clues of information extraction from chart parts are: word itself, semantic features, classes of letters, recognition results of chart parts.

1　はじめに

大量の文書データの中から、あらかじめ目的とした情報のみを取り出してくる技術、すなわち内容抽出の技術は、文書情報の整理やデータベースの自動的な構築のためには、必須の技術である。

文書データからの内容抽出の第一目標としては、テキストからのキーワード抽出がある。キーワード抽出に関しては、例えば、どのようにして重要なキーワードを確定するかなど、さまざまな観点から研究がなされている。しかし、単なるキーワード抽出だけでは、得られた情報は構造化されておらず、キーワード間の関係は不明である。そこで、内容抽出の次のステップとして、構造化キーワードの抽出、すなわち、キーワード間の関係情報を取り出すことが考えられる。

通常のテキストベースでは、キーワード間の関連は、日本語の助詞、動詞、．．．などの文法的な機能によって表現される（例：本体とフロッピーディスク装置の売価は、それぞれ１５万、３．５万です）。こうした、主語、述語等を備えた通常の文からの内容抽出の研究は、学術論文[1]、特許[2]、製品の紹介記事[3][4]、新聞[5]等の文書を対象として研究されている。しかし、電子メール中では、主語、述語等を備えた文のみではなく、表、箇条書き、単語の羅列といった文以外の構成要素を含んでいる場合が多い。したがって、従来の文間や文内の文法的関係をもとにキーワード間の構造化を行うことはできない。また、文法的な情報以外には、従来、多くのシステムが、記述対象分野の専門知識を強化することにより、処理の高精度化を行ってきた。しかし、テキストベースを対象とした知識の拡充は、データベースの場合と比べて膨大な工数が必要であるとの指摘もあり[6]、専門分野ごとに十分な知識ベースを作成し、世の中の動きに合わせて時々刻々とこれをメンテナンスしていくことは困難である。このため、文以外の要素を含む文書に対しても対応でき、かつ、不十分な知識ベースのみを用いても、ある程度の構造化キーワードの抽出が可能な手法を導入する必要がある。そこで、本研究では、電子メールの文書構造の認識、すなわち、表や箇条書きの全体範囲を認識し、個々の項目について述べられているまとまり（ここでは、これをブロックと呼ぶ。１～数行からなる）に分割することによって、キーワード間の関係を認識する方法を提案する。

以上では、主に自然言語処理や内容抽出といった観点から本研究の位置づけを述べた。次に、本研究で提案する方式では、箇条書き、表などの文書構造を認識する処理過程が含まれている。この点について、本研究の位置づけを述べる。

箇条書き、表などの文書構造を認識する研究は、２つの面からなされてきている。

１つは、ＳＧＭＬやＯＤＡといった文書の論理構造記述の観点からの研究[7]である。これらの研究では、箇条書きの認識や、文章中の記述から表への参照構造の認識は、自動的に行っている。しかし、表は１つのブラックボックス的な存在としてのみ捉えられ、表の内部構造までを認識する研究とはなっていない。

もう１つは、いわゆる文書画像理解と呼ばれる研究分野[8][9][10]で、画像認識や文字認識の延長線上の分野である。文書画像理解の分野では、表の内部構造の認識を行っている。しかし、これらの研究の多くは、個々の文書に固定された表構造のみを扱っており、本研究で取り扱う、箇条書きと表の混在といった多様な書式構造には対応していない。また、一部では、多様な帳票構造を認識する研究もなされているが、帳票という表の認識にとどまっている（すなわち先程述べたように箇条書きと表の混在といった多様な書式構造には対応していない）。また、文書画像理解の分野では、いずれも、入力が画像データであるため、水平・垂直の枠罫線も情報として入力され、それが表構造の認識に有効な情報となっている。これに対して、本研究では、入力がテキストベースの文書であるため、水平・垂直の枠罫線は論理上では存在するが、印刷出力されて文書には現ない。このため、これを推測しなければならない必要性が生じてくるという違いがある。

また、上に述べた２つの面からの研究は、いず

れも、人間が文書構造を認識する際に用いる、単語の意味や文法といった自然言語処理的な情報を用いていない。また、内容抽出の観点からの研究とはなっていない。

2　電子メールの文書構造

2.1　電子メールの文書構造

電子メールは、日本語から記述されている。しかし、いわゆる、主語、述語等を含む文のみから構成されているわけではない。我々は、電子メール中に出現する文書構造を以下の３つに分類した。

①表部分：文字等の物理的（２次元的）配置、すなわち、レイアウト情報が意味を持つ。文および単語の羅列からなる。この中には、いわゆる箇条書きも含む。

②文章：文字等の物理的配置が意味を持たない。文のみから構成される。

③その他：（①、②以外の部分、例えば電子メール全体のタイトル（表題）部分は、ここに含まれる）

図１は、パソコンの中古品売買に関する電子メールである。この文書では、表部分、文章は、それぞれ図１に示す部分となる。

```
27  FBH01433  8/03  34 RA21他売ります。
        塚本　雄之
表部分
  ◎PC9801RA21本体
    11万円で。
  ◎120ﾒｶﾞ HDD
    緑電子製、SCSI NOVA V120
    新品同様、購入後　1か月。
    多少のｱﾌﾟﾘｹｰｼｮﾝはｲﾝｽﾄｰﾙ済みです。
    4万3千円で。
  ◎3.5ｲﾝﾁFDD 緑電子製、1FDD New Little F
    1万2千円で。
  ◎内部増設4MRAM ﾒﾙｺ製 EDA4000
    1万円で。
    いずれも、箱、ﾏﾆｭｱﾙ完備。
文章
  取りにこれる方には、ｹﾞｰﾑｿﾌﾄ（ちょっと
  古いけど）さしあげます。その他は、宅
  急便で送料無料で、送ります。
FBH01433 TSUKA
```

図１　電子ﾒｰﾙの文書例

2.2　全体の処理手順

１章で述べたように、表部分からの内容抽出にあたっては、文内や文間の文法的情報を利用することができない。したがって、表部分からの内容抽出と文章からの内容抽出の方式は異なってくる。このため本研究では、図２に示すように、まず電子メール文書全体を文章と表部分に分ける。

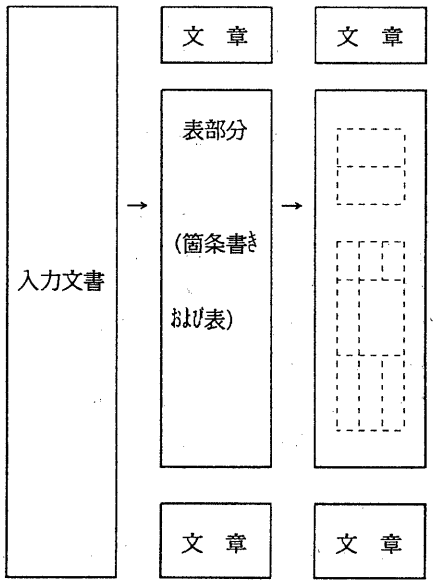
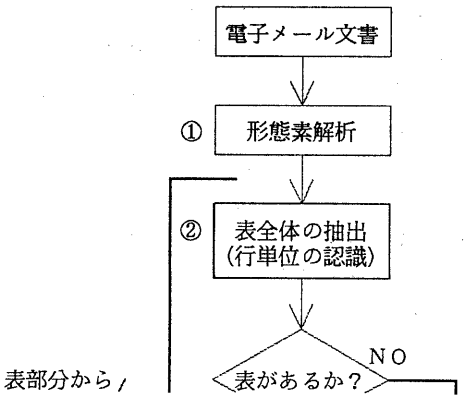


図２　文書の分割

また、全体の処理の流れは図３に示す形となっている。すなわち、



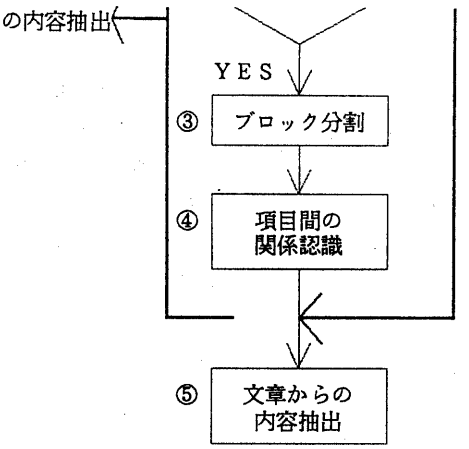


図3　全体の処理の流れ

①電子メール文書の形態素解析を行う。
②表部分全体を行単位で認識する。
③表部分を、まとまりのある行ごとに分割する。分割された結果は、1～数行からなり、ここではこれをブロックと呼ぶ。
④ブロック内に記述された単語間の関係認識を、フレームにあてはめることによって行う。
⑤文章からの内容抽出を行う。
の順で行う。

以下、本報告では、②～④で用いる規則等について、3章で詳しく述べる。

2.3　処理対象文書

電子メールには、さまざまな文書が存在する。このうち、記述対象がある程度限られた範囲に限定できる、記述内容に一般性がある（個人間の私信ではない）、文書が公開され実験対象としてまとまった数が入手できる、といった観点から、パソコン通信の中古品売買情報（売ります買います）に着目した。図4に図1の内容抽出結果を示す。ここでは、特に、中古品の中でも、パソコンを取り扱っている文書に限定し、内容抽出の目標を、パソコンの本体や付属品の名称、型番、メーカー名、価格とした。

名　称	型　番	ﾒｰｶｰ名	価　格
ﾊﾟｿｺﾝ本体	PC9801RA21	ＮＥＣ	110,000円
ﾊｰﾄﾞﾃﾞｨｽｸ	NOVA V-120	緑電子	43,000円
ﾌﾛｯﾋﾟｰﾃﾞｨｽｸ	New Little F	緑電子	12,000円
増設ﾒﾓﾘ	EDA 4000	メルコ	10,000円

図4　内容抽出結果

具体的には、Nifty-Serveの掲示板サービス中の売ります買います情報の中のＮＥＣ系パソコンの情報を用いた。数週間にわたって採録した文書、約６００件のうち、重複して投稿された文書（買い手がつかないと同一の文書を何度も投稿することになる）、ソフトウェアのみについて書かれた文書、周辺機器のみについて書かれた文書、１０年以上前についての機種について記述された文書、は人手で取り除き、残った１５０件を対象とした。

投稿された文書中、文章のみからなる文書は全体の約４０％、残りは文章と表部分からなる。全体的な傾向としては、売りたい品目数が少ない場合は文章のみからなり、記述量も少ない。多い場合には、主要な情報は表部分にまとめて記載されることが多い。文章と表部分が混在する文書では、本研究で、抽出目標とした名称、型番、メーカー名、仕様、各品目の価格の情報は、ほとんどの場合、表部分に記載されている。

3　内容抽出処理方式

3.1　形態素解析

形態素解析部では、形態素解析システムJUMANを用いる。また、単語の字面、品詞、意味属性の他に、単語の文書中における二次元的位置情報（行番号と列番号）を出力する。単語に付与した意味属性としては、メーカー名、名称（ＣＰＵ，ＲＡＭカード、etc.）、型番、および仕様（５ｲﾝﾁ、外付け、ＳＡＳＩ、etc.）がある。

また、電子メールでは、未知語が頻出する。本報告の処理対象文書では、未知語の推測法として、以下の①～③を考えることができる。
①単語を構成する文字列の種別から推測

例：電子メールのユーザーＩＤ（３桁の英文字とそれに続く５桁の数字）
電話番号（０で始まる９～１０桁の数字、途中に "-、(、)" などの記号が入ることがある）

②表の内部構造の認識結果の利用

一つの表においては、同一列にならぶ単語は、同じ意味属性を持つ。このことから、例えば、メーカー名が並んでいる列中に、未知語が出現すれば、その未知語はメーカー名であると推測できる。

③手がかりとなる単語（または意味属性）が、前方または後方に出現(例：型番 PC9801FX、メーカー名 NEC、値段:780k)

3．2 表部分全体の認識規則

表部分の認識は、文字、品詞、および意味属性の物理的配置の特徴を利用して行っている。表部分は、大別すると、箇条書きと表に分かれる。したがって、それぞれに対応した認識規則が必要になる。以下に、それぞれの規則と規則が適用される文書例を示す。

3．2．1 箇条書きの認識規則

・同じカラム列（行の先頭に限る）に昇順の数字等((1),ｲ,ﾛ,ﾊ)や同一の特殊記号(○☆※・)が存在する。

以下の製品を売ります。
１．NECのPC-9801 DA2を80,000円。
２．三菱のモニタ（XC-1498C2)を25,000円。
３．IOデータのメモリ(PIO-DA134)を10,000円。

3．2．2 表の認識規則

・同じカラム列に同一の特殊記号(:;)か空白部分が存在する（文字列の配置）

EPSON　：　PC-386NARX
Logitec　：　80MBHDD
AIWA　：　PV-A24V5

その他、必要な附属品すべてあります。

・(表部分＝複数行が)特殊記号のみで構成される行で囲まれる（文字列の配置：表部分の開始および終了の判定条件）

以下の製品を処分します。

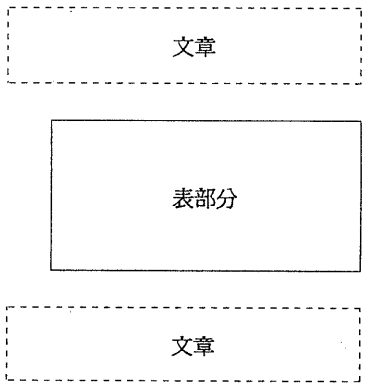
＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

本体	NEC	PC9801FA2
モニタ	三菱	XC-1498C2
ハードディスク	ICM	HC-100ES
メモリ	メルコ	EMJ-4000L

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

以上、フルセットで10万円でお譲りします。

・複数行の先頭列が文章部の先頭行より右側に存在する（文字列の配置：表部分の開始および終了の判定条件）



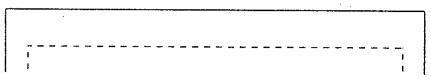
・表部分の品詞構成

①名詞(一部の転成名詞を含む)のみから構成される。

②助動詞、助詞、用言、読点、句点を含まない。

・同一表内は同一の分類の意味属性を持つ単語によって構成される（意味属性の配置）

売ります。



本体　PC-386M STD
　箱、マニュアル、附属品すべて有
HDD 40M SASI
　箱、マニュアル、附属品すべて有
CRT 12インチ・カラー

以上　手渡し希望

価格は相談に応じます。

3.3 ブロック分割の手がかり

箇条書きのブロック分割

・先頭列の数字や特殊記号

1.NECのPC-9801 DA2を80,000円。
（マニュアル付き）

2.三菱のモニタ(XC-1498C)を25,000円。

3.IOデータのメモリ(PIO-DA134)を9,000円

表のブロック分割（文字列の配置）

・特殊記号

EPSON　:　PC-386NARX

Logitec　:　80MBHDD

AIWA　:　PV-A24V5

・空白部分

日本電気
　PC9801DA2　(386DX,20MHz)
　80,000円。

シャープ
　CU-14BD　(400/200ライン対応)
　25,000円。新品同様。

緑電子
　DOODA!A-120 (外付けﾊｰﾄﾞﾃﾞｨｽｸ)

表のブロック分割（単語の羅列）

3.2節で表部分と認定された場合でも、上に示したブロック分割の手がかりとなる特殊記号や空白部分等が存在しない場合がある。その時は、1行1ブロックとして処理する。

・1行1ブロック

エプソン　486GR+

ハードディスク　ICM内蔵180メガバイト

メモリー　IOデータ16メガバイト

3.4　記述項目間の関係認識

パソコンの中古品売買情報の電子メールでは、3.3節で分割された1つのブロックが、本体/付属品の1品目の記述（名称、型番、メーカー名、価格、仕様、等）に対応している。そこで、図5に示すように、文書中の1つのブロック（破線で囲まれている）と、フレームとを対応づける。そして、ブロック中に存在する各名詞の字面や意味属性にもとづいて、フレーム中のどのスロットを埋めるかを決定してゆく。

電子メールでは、帳票のフォーマットに記入した場合と異なり、抽出すべき項目（スロット）よ

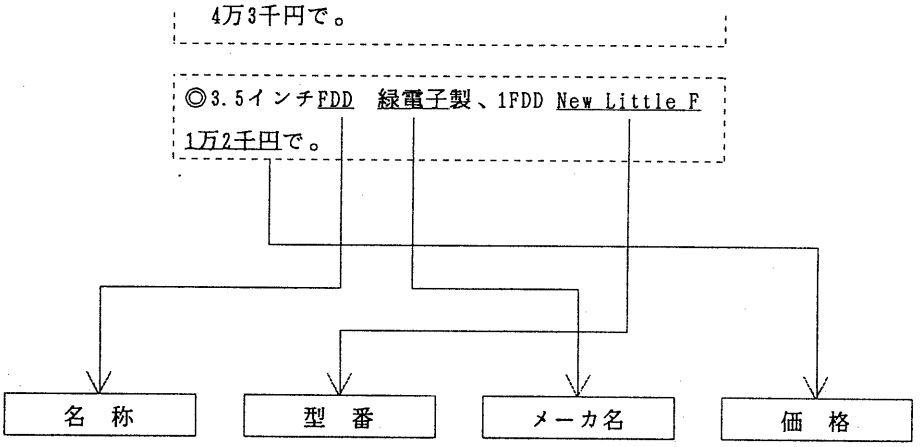


図5　記述項目間の関係認識
（同一ブロック内での単語の関係付け）

り多くの記述項目が含まれている場合や、記述項目が不足している場合がある。不足している場合は、フレーム中の埋まらないスロットを推論する必要が生じる。例えば、型番(PC9801RA)のみの記述から、名称(パソコン本体)やメーカー名(NEC)を推論する場合である。推論には、対象分野についての知識が必要となる。完全な推論を行うことはむづかしいが、実現効果の高い知識から実装していく。

3.5　文章からの内容抽出

この部分は、現状では、形態素解析の結果を用いたキーワード抽出を行っているだけである。

4　実験

3章で述べた5つの処理過程のうち、3.2～3.4節の3つの処理の机上分析を行った（3.2節の入力である形態素解析は正しく行われているものとした）。また、処理対象件数は、収録した電子メール150件中、人手で判断して表部分が含まれるとした88件である。結果を以下に示す。

分析結果

・表部分全体の認識　　73/88(83%)
・表部分のブロック分割　67/73(92%)
・記述項目間の関係認識　52/67(78%)

表部分全体の認識に失敗した文書例としては、

EPSON、ノートパソコン売ります。
型　式：PC386NAR2
定　価：¥288000
附属品：箱以外有り（登録者カードは除く）
　　93/6購入（ほとんど使っていません）
　　＋
40M内蔵ハードディスク
　　＋
FM音源（FMStation,定価¥29800）
以上の3点セットで、¥155000。
送料当方負担。
まずは、メールを。

JCC03763

がある。この文書では、特殊記号”＋”がブロックの区切りとして使用されている。こういった使用法は、他の文書では全くみつからない例であるため、対応できていない。

記述項目間の関係認識に失敗した文書例としては、

1．CPU　NEC PC-9801 DA2
（箱、付属品有り）
2．モニタ　三菱 XC-1498C2
（箱有り）
3．メモリ　I･O DATA PIO-DA134(8M)
（本体に取り付け済）
4．その他　CYRIX CX486DLC33GP
CYRIX CX83D8733GP
プリンタリボン（黒、カラー）

がある。この文書では、その他のブロックには、１品目ではなく、３品目が記述されているためである。

5　おわりに

現在、多くの電子メール文書に適用して評価を行うとともに、処理系の開発を行っている。今後の課題として、以下の２点を考えることができる。

・得られた抽出結果は、ユーザーの関心により、さまざまな形で検索される。具体的には、製品名から検索する場合、おおまかな製品種別のみを指定して値段から検索する場合、等さまざまな要求がある。こうした種々の検索要求に対応できるデータ構造、すなわちデータベースへの格納形式について考えていく必要がある。

・本実験において使用した１５０件の電子メール文書は、それぞれが別々の人間によって書かれている。したがって、本研究で提案したアルゴリズムは、書式等の記述方法の個人差を、ある程度吸収できることが確認できた。しかし、記述内容がパソコンの中古品情報という限られたものであるので、これが変わってきた時に、今までと同様の文書解析アルゴリズムで良いかどうかを、検討してゆく必要がある。

謝辞

形態素解析システムJUMANを提供していただいた京都大学の長尾真先生、奈良先端科学技術大学の松本裕治先生および開発グループの方に感謝いたします。

参考文献

1)猪瀬博、斉藤忠夫、堀浩一：シナリオを用いる論文抄録理解・作成援助システム、情報処理論文誌、Vol.24,No.1,pp.22-29(1983)
2)高松忍、日下浩次、西田富士夫：技術抄録文からの関係情報の自動抽出、情報処理論文誌、Vol.25,No.2,pp.216-224(1984)
3)松尾比呂志：抽出パターンの階層的照合に基づく内容抽出法、情処ＮＬ研究会、Vol.99,No.2,(1994)
4)小松英二、加藤安彦、安原宏、椎野努：要約支援システムＣＯＧＩＴＯ（文書の構造解析）、情処ＮＬ研究会、Vol.64,No.11(1987)
5)Paul Jacobs and Lisa Rau,"SCISOR: Extracting information from on-line news", Comm. ACM, Vol.33, No.11(1990)
6)秋山幸司：テキスト情報の知的検索における諸問題、情処ＤＢ研究会、Vol.64,No.3,(1988)
7)土井美和子、福井美佳、山口浩司、竹林洋一、岩井勇：文書構造抽出技法の開発、電子通信学会論文誌、Vol.J76-D-Ⅱ,No.9,pp.2042-2052(1993)
8)絡琴、渡辺豊英、杉江昇：多種帳票文書の構造認識、電子通信学会論文誌、Vol.J76-D-Ⅱ,No.10,pp.2165-2176(1993)
9)山田満：文書画像のＯＤＡ論理構造化文書への変換方式、電子通信学会論文誌、Vol.J76-D-Ⅱ,No.11,pp.2275-2284(1993)
10)山下晶夫、天野富夫：モデルにもとづいた文書画像のレイアウト理解、電子通信学会論文誌、Vol.J76-D-Ⅱ,No.10,pp.1673-1681(1992)