表形式や箇条書きを含む文書の構造認識

2R－8

山本勝紀、河合敦夫、加藤親貴、椎野努
三重大学　工学部

1．まえがき

通常の技術文書やビジネス文書は、小説と異なり、いわゆる主語、述語を含む文のみから構成されることは少なく、図、表、箇条書き等を含むことが多い。最近、こうした文書が、電子メールによってやりとりされることも多くなってきた。画像、音声を含むデータのやりとりも存在するが、文字コードのみを介した情報交換は、依然として多くの割合をなしている。この場合、文書中の図や表は、文字コードの集合として表現しなくてはならない。図を文字コードのみを使って表現することはむずかしいため、電子メール中の文書には含まれないことが多い。しかし、表は、枠罫線無しの表として文書中に頻繁に出現する。

一方、ビジネス文書や技術文書の構造認識は、文書からの内容抽出、ＳＧＭＬ等への文書記述変換等に必要となる。文書構造の認識については、箇条書きの範囲や、文書中の記述から表への参照構造を、自動的に認識する研究[1]は行なわれている。しかし、表はブラックボックスしてのみ捉えられ、表の内部構造までは取り扱っていない。また、表の内部構造の認識は、いわゆる文書画像理解と呼ばれる研究分野[2][3]で行われている。しかし、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必須の情報となっている。これに対して、本研究は入力が文字データ（文字コード）の文書である。したがって、枠罫線は論理上では存在するが、出力される文書には印刷されない。このため、架空に存在する枠罫線を、文字列の２次元配置や単語の持つ意味をもとに推測しなければならないという違いがある。

2．文書の構成要素

本発表では、文書の各部分を２つに分類する。

(1)文章：文字等の２次元的配置が意味を持たない。文のみから構成される。

(2)表部分：文字等の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。具体的には、箇条書き、および、文字コードのみで表現される表からなる。この中には、５章に示すように、不完全な表も含む。

3．　全体の処理手順

全体の処理の流れを図１に示す。すなわち、

(1)文書の形態素解析を行う。

(2)文書を文章と表部分を切り分け、行単位で認識する。表部分が複数の表から成り立つ時は、複数の表の切れ目を認識する必要もある。

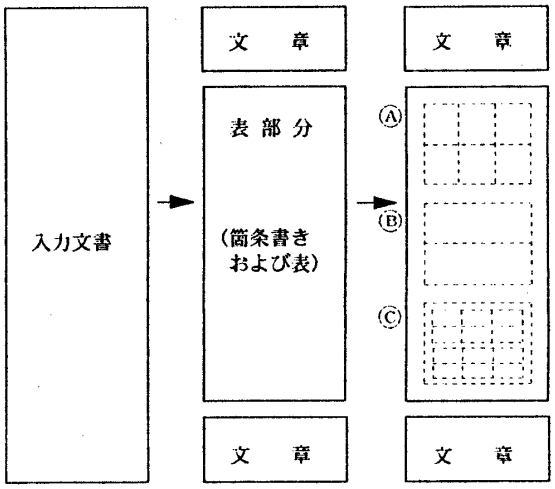

図１　構造認識処理の全体の流れ

Recognition of Document Structure Including Tabular Forms and Itemization

Katsunori Yamamoto, Atsuo Kawai, Yoshitaka Kato, Tsutomu Shiino
Faculty of Engineering, Mie University,
1515, Kamihama-cho, Tsu-shi　514, Japan

(3)個々の表部分の内部構造を認識する。
(2)と(3)については、それぞれ4章、5章で述べる。

4．個々の表部分の範囲の認識

個々の表部分の範囲の認識は、空白行／空白列等の空白部分、特殊記号、単語、品詞、意味属性の、2次元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、それぞれの規則の一部を示す。

・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（箇条書きの認識規則）。
・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)(表の認識規則)
・表部分の品詞構成
名詞(一部の転成名詞を含む)のみから構成される。助動詞、助詞、用言、読点、句点を含まない。

5．表の内部構造の認識

電子メール文書中に出現する表は、関係型データベースで取り扱うことのできる、または、表計算ソフトで扱える、理想（標準）的な表とは異なることが多い。具体的には、
(1)枠罫線の出力がない。
(2)表としてのレイアウトが必ずしも完全でない。

すなわち、理想的な表は行と列から構成され、1つの行は同一の事象に関連するデータ要素の集まりを示しており、1つの列は各データ要素の同じ属性が出現する縦の欄を表している。言い替えれば、理想的な表におけるレイアウトでは、同一の属性を持つ項目は同一列に出現し、同一の事象に関連するデータは同一行に出現する。これに対して、電子メール中の文書では、この原則は必ずしも成立せず、図2の破線部に示すように、同一行の記述順序のみが保存されていることが多い。

1　品目名等

本体	エプソン	386M－STD
プリンター	NEC	PC-PR201H
ハードディスク	TEAC	120MB

2　住所等

住所	愛知県岡崎市三嶋町
氏名	加藤勝紀
USER-ID	KFG30260

図2　表の内部構造の認識(実線および破線)

(3)論理的な関係が、表全体に対して成立していない(図2の386M-STDは型番,120MBは仕様)。

具体的には、理想的な表である図1のA、箇条書きの構造化である図1のB、箇条書きの中に表が存在する図1のC、などがある。図2に図1Cの具体的な文書例を示す。

理想的な表のみであれば、表の内部構造の認識は困難ではない。しかし、電子メール中では、今まで述べたような不完全な表の割合が大きい。このため、それぞれの表のタイプに対応したアルゴリズムが必要になる。

6．むすび

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。

謝辞

形態素解析システムJUMANを提供していただいた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)土井,福井,山口他：“文書構造抽出技法の開発”,信学論,**J76-D-Ⅱ**,9,pp.2042-2052(1993-9)
(2)駱琴,渡辺,杉江：“多種帳票文書の構造認識”,信学論,**J76-D-Ⅱ**,10,pp.2165-2176(1993-10)
(3)山田満：“文書画像のＯＤＡ論理構造化文書への変換方式”,信学論,**J76-D-Ⅱ**,11,pp.2274-2284(1993-11)