論文の内容の要旨

論文題目　A METHOD FOR INFORMATION EXTRACTION FROM TABLES AND LISTS
（表形式と箇条書き形式からの情報抽出手法）

氏名　吉田　稔

WWW 上の表形式と箇条書き形式に対する解析手法について論じる。表と箇条書きは、WWW 文書内にしばしば見られる表現形式であり、その解析は、WWW 文書の理解には不可欠である。表や箇条書きは、自己紹介における「性別」属性と「女性」属性値の如く、ある物事の属性と属性値による表現と捉えることができる。この属性と属性値が、表あるいは箇条書き内のどの部分に相当するかを決定することが、本論文で扱う主な問題である。

WWW 文書内の単語は、単語そのものの意味の他に、それが表示される位置、文字の大きさ、色といった、様々なレイアウト情報を持つ。特に、表や箇条書きに代表される文章以外の表現に対しては、文章の場合と異なり、一般的な文法規則が存在しないため、レイアウト情報がより重要な役割を果たす。そのため、本論文で提案する解析手法は、レイアウト情報を積極的に利用するという方針に基づいている。

本論文は主に２つの内容で構成される。すなわち、表の解析と箇条書きの解析、である。表の解析では、表形式内に出現する単語間における属性・属性値関係（オントロジー）を抽出するための手法を提案する。ここでは、単語の位置を手がかりとして、EM アルゴリズムによる語彙情報の推定を行っている。ここで語彙情報とは、ある単語が属性として用いられる確率の値、あ

るいは、属性値として用いられる確率の値のことを指す。

箇条書きの解析は、表解析によって得られたオントロジーに基いて行われる。様々なレイアウトを持つ箇条書き形式が、オントロジーの利用により解析できる。箇条書き形式の解析結果は、レイアウト情報の推定にも用いることができる。ここでレイアウト情報とは、WWW 文書のソースファイル中における各 HTML タグの出現確率の値を指し、これは EM アルゴリズムにより推定することができる。推定された確率値は、箇条書き形式の解析精度を向上させることにも役立つ。

これらの手法を、WWW 上に存在する実際のページに対して適用し、レイアウト情報が表や箇条書きの解析に於いて有用であることを実験結果を通じて示す。