

## NEWテクノロジー

# レイアウト定義不要で帳票の論理構造を認識する先進技術



富士通研究所は、見積書や納品書といった多様な帳票から、あらかじめレイアウト定義を行うことなく高精度にデータの論理構造を認識する技術を開発しました。この技術により、レイアウトが未知な帳票を扱う窓口業務においてデータ入力業務の負担を大幅に軽減できるほか、タグの自動付与機能によってタグ付き検索が可能になります。またe-文書法や日本版SOX法に対応したソリューションへの応用も期待されています。

### 求められる非定型帳票の効率的な電子保存

e-文書法の施行によって、それまで紙による保存が義務付けられていた財務・税務関係書類は、電子的に保存することが認められるようになりました。また一方、2008年3月決算期からの施行が予定されている日本版SOX法では、従来以上に財務状況の管理・監視が求められると同時に、IT活用による内部統制の重要性も示されています。

こうした環境の変化から、企業ではさかんに紙文書・帳票の電子化を進めています。スキャナで取り込んだ文書を画像データとして保存してデータベース化するには、日付や取引先名、金額のデータを検索できるようにキーワードを付与する作業が必要になってきます。

しかし見積書や納品書は、項目の論理的な構造は似ていても、レイアウトや見出し表記はそれを発行する企業毎に異なります。こうした「非定型帳票」に対して、あらかじめレイアウトを定義する必要のあるOCR技術を適用するには、異なる帳票毎、あるいはレイアウト変更が行われる毎に定義を登録しなければなりません。また、OCRを導入せず人手によってデータ入力をするとなると、帳票の枚数に比例して膨大なコストがかかってきます。

### 確率伝搬法を用いた論理構造認識技術の開発

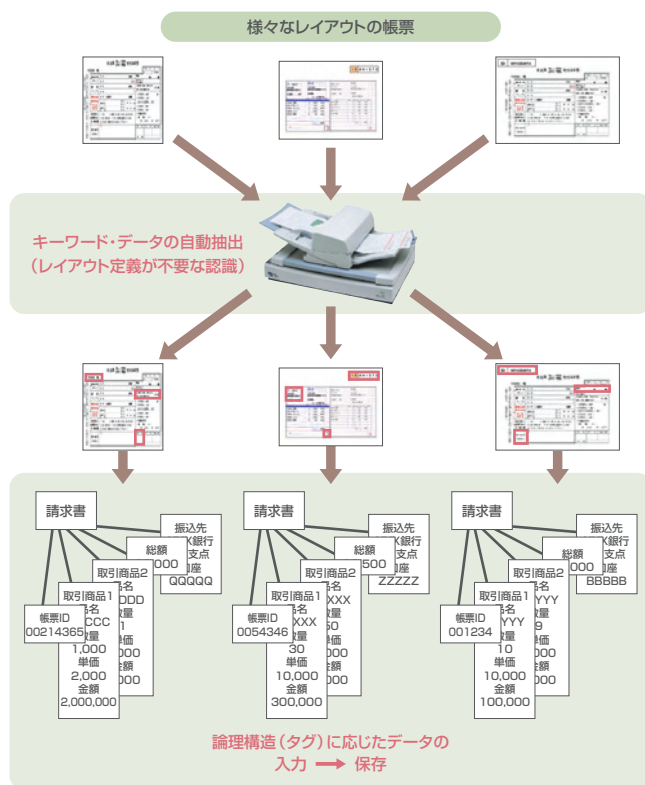
そこで富士通研究所は、文字の位置やレイアウトに依存しない技術の開発に取り組み、2006年10月、業界で初めて、多様なレイアウトの帳票から見出しやデータの論理構造を高精度に認識できる技術の開発に成功しました。

非定型帳票の論理構造認識技術は以前から存在していましたが、それは、「見出し辞書」として登録してある見出し（例えば「氏名」）のような文字列をまず帳票から認識し、次にその見

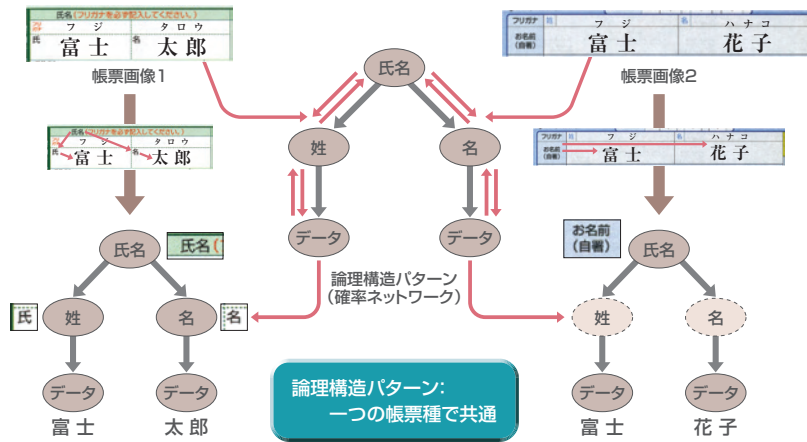
出しの位置から一定の範囲内の文字列（例えば「富士太郎」）がデータであると認識する逐次探索方式でした。しかしながらこの方式では、見出しに対応するデータを位置関係から検出するためのルール記述が膨大になり、多様なレイアウトへの対応は非常に困難でした。また、見出しとしての「氏名」が認識できなければ、「富士太郎」がデータであると正しく認識できないため、複雑な階層の見出しや、見出しの抜けがある帳票にも適用できませんでした。

富士通研究所が開発した論理構造認識技術は、スキャナで読み取ったキーワードやデータとなる文字列の意味関係に基づく推論方式を用いて論理構造を認識することによって逐次探索方式による問題を解決しました（図1）。

■ 図1 レイアウト定義が不要な非定型帳票のデータ入力技術の実現



■ 図2 論理構造認識技術の原理



本技術の原理を示します(図2)。まず、見積書や請求書等、帳票の種類毎に共通した論理要素(見出しやデータ)に対応する文字列の特徴と、それらの意味関係の可能性を記述した論理構造パターンを用意しておきます。例えば氏名欄の見出しの可能性として、「氏名」「お名前」等があり、その見出しには「氏」「名」あるいは「姓」「名」等の見出しが含まれる、といった論理要素間の関係を、確率ネットワーク上で表現します。そしてそのネットワーク上で文字列の情報を相互に伝搬させることにより、各論理要素に対応する信頼度の高いデータを認識し、逆に信頼度の低いデータを棄却していきます。また信頼度の低いデータであっても、信頼度の高いデータとの整合性があれば信頼性が高められていくといったプロセスが適用されることにより、複雑な帳票からも高精度な論理構造認識が可能となります。

こうした方式を用いることによって帳票毎の位置関係の登録が不要となり、多様なレイアウトへの対応が可能となります。加えて、文字認識誤りを類推したり、省略された論理要素を検出したりすることも可能になり、階層的な見出しを持つ複雑な帳票においても、安定した認識結果が得られるようになります。

**「紙」と「電子」のシームレスな連携の実現によって期待される様々なソリューション**

本技術を富士通指定の20の評価文書に適用したところ、9割以上の論理構造認識率(論理要素に対する文字列の対応付け正答率)が得られました。この結果から、従来は全てを手作業に頼っていた検索用キーワード付与やデータ入力コストを

約60%削減することが期待できます(コストモデルに基づく試算)。即ち、これまで5日かかっていた人手によるデータ入力・確認作業を2日に短縮でき、担当部門の作業を大幅に効率化できます。

本技術は、レイアウトが未知な帳票を扱うデータ入力作業に適していることから、帳票入力業務の支援のためのソリューションとして、富士通のOCRソフトやスキャナ製品への搭載を目指していきます。

また本技術の自動タグ付与機能は、紙だけでなく電子データにも適用できます。したがって、WordやExcel形式のファイルでタグを自動付与し、「取引日」が「2007/5/1」の伝票」といった検索をすることも可能になります。

今後さらに論理構造認識技術の性能を高めることにより、e-文書法に対応したキーワード付与や、内部統制に対応した帳票間での整合性確認や情報保証に関連するソリューションへの採用を目指します。

**お問い合わせ先**

(株)富士通研究所  
ITコア研究所 言語・メディア研究部  
TEL 044-754-2678  
E-mail lm-pr@ml.labs.fujitsu.com