【18】

氏名	松本俊子（まつもと としこ）
博士の専攻分野の名称	博士（情報科学）
学位記番号	第 25292 号
学位授与年月日	平成24年3月22日
学位授与の要件	学位規則第4条第1項該当 情報科学研究科マルチメディア工学専攻
学位論文名	企業情報システムにおけるデータの抽出の効率化に関する研究
論文審査委員	（主査） 教授 薦田 憲久 （副査） 教授 細田 耕　教授 西尾章治郎　教授 藤原 融 教授 下條 真司　准教授 原 隆浩　准教授 秋吉 正徳

論文内容の要旨

本論文は，筆者が2000 年から現在まで（株）日立ソリューションズならびに2010年から現在まで大阪大学大学院マルチメディア工学専攻在学中に行ってきた，企業情報システムにおけるデータ抽出の効率化に関する研究をまとめたものである。

企業情報システムによる業務効率化の進展に伴い，効率化の対象は，状況に応じて様々な内容が記載される定型性の低いデータへと移りつつある。近年データの大規模化がますます顕著になる中，データの中から，ユーザが種々の業務をこなす中で着目するべき部分を抽出する作業の効率化に対するニーズが高まっている。この点に関し，企業情報システムにおいて利用されるデータの主なものとして挙げられる数値データおよび文書データにおける課題を解決する手法について提案する。

数値データからの着目するべき箇所の抽出に関しては，データ量の増大に伴い自動処理の重要性が高まるとともに，データの測定原理に由来する内在的法則性を利用する高精度な処理が求められている。内在的法則性について蓄積された専門家の知見は定性的な形で表されることが多いため，法則性が典型的に現れているシンプルなデータを集めて傾向を調べることで知見を定量化し，着目するべきデータを抽出する手法を提案する。

文書データに関しては，内部統制の監査における迅速な提出のため，タイトル，顧客名などのメタデータを文書データから抽出し整理分類して管理するニーズが高まっている。しかし従来のメタデータ抽出技術は抽出にあたり着目するべきキーワードやレイアウトを抽出用ルールとしてあらかじめ設定しておくことを前提としている。そこで，ビジネス文書の記載上の傾向に基づき，サンプル文書における正解メタデータの記載から抽出用ルールを生成する手法を提案する。

さらに，コンプライアンス違反の防止のため，法令，社内規則などの多数の文書データからユーザの業務遂行上参照が必要なものの抽出効率改善が求められている。そこで，業務上利用するアプリケーションの表示文字列の例とそれぞれの状況における業務情報の参照要否を入力として，参照要否の判別条件を構成・維持する手法を提案する。

論文審査の結果の要旨

企業情報システムによる業務効率化の進展に伴い，効率化の対象は状況に応じて様々な内容が記載される定型性の低いデータへ移り，ユーザが種々の業務をこなす中で着目するべき部分の抽出の効率化に対するニーズが高まっている。本論文は，数値データおよび文書データについて，データに内在する性質に基づいて抽出を効率化する手法として(1)数値データからユーザの業務において着目するべき箇所を専門家の知見に基づいて抽出する技術，(2)ビジネス文書の記載上の特徴に基づいてメタデータ抽出用ルールを生成する技術，(3)文書データの参照必要性の判断基準に基づいて業務上必要な文書データの抽出効率を改善する技術についての研究成果をまとめたものである。その主要な成果を要約すると次の通りである。

(1) 数値データのデータ量の増大に伴い着目するべき箇所の抽出の自動処理の重要性が高まるとともに，データの測定原理に由来する内在的法則性を利用する高精度な処理が求められている。法則性は定性的知見として表されることが多いため，法則性が典型的に現れているシンプルなデータを集めることで知見を定量化し着目するべきデータを抽出する手法が求められる。ヒトゲノム解読以降バイオテクノロジー市場拡大を牽引しているDNAデータを対象とした手法を提案し、平均94％の精度で真のデータを識別できることを示している。また，シンプルなデータを用いて法則性を定量化する手法について，企業内の他の数値データへの適用性を述べている。

(2) 内部統制の監査における迅速なビジネス文書提出に向け，タイトル，顧客名などのメタデータ抽出技術の導入におけるボトルネック解消のため，サンプル文書とその正解メタデータを入力として，キーワードやレイアウトの記載上の特徴を抽出用ルールとして自動生成する手法を提案している。6つのビジネス案件における営業文書および5つの研究プロジェクトにおける週次作業報告書を用いて，人手で設定したルールと同等の再現率を達成する抽出ルールを高速に自動生成できることを示し，提案手法の有効性を示している。

(3) 法令，社内規則などの多数の文書データからユーザの業務遂行上参照が必要なものの抽出効率改善のため，業務上利用するアプリケーションの表示文字列の例とそれぞれの状況における業務情報の参照要否を入力として，参照要否の判別条件を構成・維持する手法を提案している。高頻度な形態素列を「言い回し表現」として抽出し，決定木の構成および提示を抑止するキーワードの選択に利用する。3種類の業務情報を用いて，提示要否の正確な判別と不要な業務情報の提示の防止が両立できていることを確かめ，提案手法の有効性を示している。

以上のように，本論文は企業情報システムにおけるデータの抽出の効率化に関する先駆的研究として，情報科学に寄与するところが大きい．よって，本論文は博士（情報科学）の学位論文として価値あるものと認める．