別記様式第３号　　課程博士用

平成　23年　2月28日

豊橋技術科学大学長　殿

審査委員長　　青野　雅樹

論文審査及び最終試験の結果報告書

このことについて，下記の結果を得ましたので報告いたします．

学位申請者	鶴田　雅信	学籍番号	第００１０６５号
申請学位	博士（工学）	専攻名	電子・情報工学専攻
論文題目	クローリング範囲を考慮した Web サイトからの情報抽出システム		
公開審査会の日	平成２３年２月２５日		
論文審査の期間	平成２３年１月２７日～平成２３年　２月２８日	論文審査の結果	合格
最終試験の日	平成２３年２月２５日	最終試験の結果	合格

論文内容の要旨

本論文では，人間が用いていると考えられるヒューリスティクスを組み込むことで，計算能力やネットワーク帯域などの資源を潤沢に持たないユーザでも利用できる，大規模なクローリングを必要とすることなく実行できるような情報抽出システムの構築手法について記述したものである．第１章では，ヒューリスティクスを組み込んだ情報抽出システムについての背景，および，関連研究について説明している．第２章では，Web ページから主要な部分を抽出するために，人間がアノテーションを行ったレイアウトの情報を用いる手法について述べている．第３章では，第２章の内容を発展させ，アノテーションが付与されたレイアウト情報を，より有効に利用出来る手法について述べている．加えて，クローリングを行う既存手法，および，人間の作成した広告パターンのデータを利用する既存手法と組み合わせることで，第２章で述べた手法の問題点を解決し，性能を向上させる枠組みについて述べている．第４章では，企業の基本情報属性という，異なる企業の Web サイト間で共通したフォーマットを持たない情報を，手がかり語に類似した語を持つリンクを辿ることでクローリングを行いながら抽出する手法について述べている．第５章では本博士学位論文の結論について述べている．

審査結果の要旨

Web ページからの主要部分抽出は，Web マイニングや全文検索の前処理として非常に重要な要素技術である．本論文で提案された手法は，一般的なブラウザにおける Web ページのレンダリング結果，および，人間によるアノテーションを用いることに新規性がある．さらに既存のノイズ除去手法を組み合わせることで，既存の主要部分抽出手法に比べて大きな性能向上を達成しており，工学上，意義深い成果を得ている．また，企業の基本情報属性は，異なる Web サイト間において共通したフォーマットを持たない属性であり，大規模なクローリングを行うことなく抽出する試みはこれまで行われていなかった．本論文において，提案手法は，大規模なクローリングを行った上で情報抽出を行うベースライン手法と比較して高い性能を示しており，ページの探索を行いながら情報抽出を行うというアプローチの有用性を示唆している．この研究成果は，企業の基本情報だけではなく，多くの共通したフォーマットを持たない属性情報についても適用が可能であると考えられ，工学上の意義は高い．本研究の結果は，計２編の原著論文，また，１編のレターとして論文誌に掲載されている．以上により，本論文は博士(工学)の学位論文に相当するものと判定した．

審査委員

青野　雅樹　印	石田　好輝　印	増山　繁　印

（注）　論文審査の結果及び最終試験の結果は「合格」又は「不合格」の評語で記入すること．