様式：課程博士用

平成 23 年　1 月 12 日

電子・情報工学専攻	学籍番号	001065
申請者 氏名	鶴　田　　雅　信	

指導 教員	増山　繁 磯田 定宏

論　文　要　旨　(博士)

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

(要旨　1,200 字程度)

現在，Web 上には膨大な数の Web サイトが存在する．これらの Web サイトに含まれる情報を有効活用するためには，情報を抽出し，整理する技術が非常に重要となる．また，一般的な Web からの情報抽出に関する技術では，クローリングを行い，Web サイトに含まれる Web ページをあらかじめ収集する必要がある．クローリングの規模は技術の種類によって異なるが，一般的に，有用な情報を抽出するためには，非常に多くのページを対象とすることが重要であると考えられている．しかしながら，大規模なクローリング，および，それによって収集した Web ページの処理にかかるコストは非常に大きく，ネットワーク帯域，および，計算機といった資源を潤沢に持たないユーザが Web サイトから情報抽出を行う際には困難を伴う．一方，一般的なユーザが Web サイトから必要とする情報を探索する際には，非常に少ないページを探索するだけで情報を発見できることがある．このような事例は，Web からの情報抽出システムを運用する際に，必ずしも大規模なクローリングが必要ではない場合が存在することを示している．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，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を提案する．

本論文の内容は，3 つの研究によって構成されている．2 章では，人間がアノテーションを行ったレイアウトの情報を集約し，Web ページから主要な部分を抽出する手法について提案する．この手法は，ベースラインにくらべて良い性能を得たが，抽出結果にノイズを含むという問題点を残していた．3 章では，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を提案する．さらに，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，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提案を行う．また，提案手法，および，その組み合わせ手法の性能について比較実験，および，詳細な考察を行った．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提案する．また，評価実験において，提案手法の性能が，サイトの大半をクローリングするベースライン手法を上回っていたことを示した．