修　士　論　文　要　旨

開放環境科学　専攻	学籍番号 80224679	氏名（フリガナ）	朝吹（アサブキ）　拓（タク）

（論文題目）

スタイルシートによりWebページの意味的構造を抽出する方法

（内容の要旨）

近年，インターネットを介した情報公開が進められており，今後も多くの文書がWWW上にWebページとして公開されると予想される．しかし，WWWに統一的な索引は存在しておらず，検索サイトなどで使用される情報検索技術がますます重要になると考えられる．

Webページの記述に使われるHTMLは，表題や見出しなどの論理的な構造により要素を定義しており，検索エンジンはWebページをその論理構造とテキストから解析している．しかし，検索エンジンとは異なり，ユーザはWebブラウザを用いて閲覧し，論理構造ではなく，画面上に表示された視覚的な印象によって内容を理解している．また，1つのWebページ内で関連性の少ない内容が複数扱われていることがあり，現状の検索サイトでは，離れた場所にあるキーワード同士によって，ユーザの望まないページが提示されることもある．

Webページの表示や視覚効果はスタイルシートによって指定できる．本研究では，このスタイルシートを解析することによって，Webページの意味的構造を視覚効果に基づいて抽出する方法を提案する．意味的構造とは，そのページ内で扱われている意味内容の構造である．論理構造ではなく，スタイルシートを用いることで，実際に閲覧するユーザに近い基準で処理を行える．意味的構造の抽出では，空白や罫線，文字の大きさや太さ，色などの情報を基準とし，HTMLで指定されている視覚効果もスタイルシートに変換することで構造抽出に用いる．

そして，抽出した意味的構造によってWebページをテーマごとに分割する方法を提案する．また，意味的構造に基づき，Webページを特徴付けるキーワードを抽出する方法も提案する．なお，テーマごとの分割では，テキスト内容も併用して同一のテーマを扱っている部分を識別する．

提案する方法は，HTMLの論理構造や特定の記述パターンを前提としておらず，Webページ全般に対して適用することができる．本研究で提案した意味的構造抽出をWWW検索システムで使用することにより，ユーザに近い基準となる．検索システムを利用するのは人間であるため，これを用いた検索システムの精度は向上することが期待される．

本研究での提案を実装したシステムで評価実験を行った結果，テーマ区切りでは被験者と多くの区切り位置が一致した．また，キーワード抽出は頻度順抽出と同程度の精度となった．課題は残るが，正しくテーマを分割していることから，意味的構造の抽出ができたと考える．よって，本研究の提案である，スタイルシートによるWebページの意味的構造抽出が可能であることが分かった．

慶應義塾大学大学院理工学研究科　前期博士課程

（内容の要旨は約25行程度で記入のこと）