

WWW 情報空間の弱い構造化と エリアビュー機能*

3P-6

福島 伸一

石塚 満†

†東京大学工学部電子情報工学科

1 はじめに

WWW (World Wide Web) 情報空間上は、そのオープンな性格から多種多様な情報が無秩序に散在している。このような WWW 情報空間を利用する上で、閲覧している WWW ページを中心に表示する既存のブラウザでは、有益な情報を効率良く得ることは困難になっている。そこで、より広い視点で WWW 情報空間をとらえるための新たな技術が必要であると考えられる ([1])。

本稿では、WWW 情報空間を弱い枠組みで構造化することにより実現されるエリアビュー機能を紹介する。

2 弱い構造化

WWW 情報空間は WWW ページ間がリンクで結ばれたネットワーク構造をなしている。WWW 情報空間上のリンクは、参照している WWW ページの作成者により意図的に作成されたものである。作成者によりリンクの作成理由は異なるが、参照先の WWW ページの存在を作成者が既知であることは共通している。つまり、多数のリンクにより参照されている WWW ページは、その参照元のページの作成者が自分以外である場合、多くの人に既知でそして意味のあるページであると考えられる。そこで、我々はその意味のあるページ（コアページ）を抽出し、コアページを基に WWW 情報空間上のページをグループ化する手法を開発した。一つの WWW ページが複数のグループに属することもあるため、この

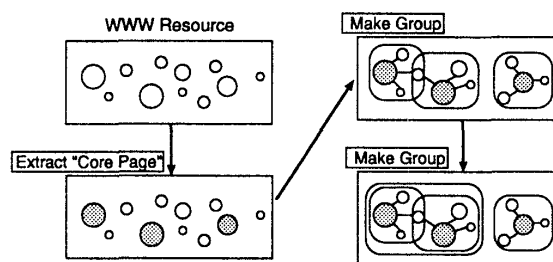


図 1: 弱い構造化

手法を WWW 情報空間の弱い構造化と呼び、以下に具体的な手法を説明する (図 1)。

コアページの抽出

各 WWW ページは、参照リンク数に応じて重みを付け、重みの大きい WWW ページをコアページとして抽出する。具体的には、他のサーバへのリンク（外部リンク）が大きくなるようにリンクに重みを付け、各 WWW ページは参照されているリンクの重みの和をそのページの重みとする。

グループ化

まず、抽出されたコアページが参照しているページと、コアページから参照されているページでグループを形成する。次に、どのグループにも属さないページは、関連度の高いグループに分類する。関連度は各 WWW ページ内から抽出した単語の特徴ベクトルを用いて算出される。

グループ間の関連度

グループ間の関連度は、WWW ページとグループとの間の関連度と同様に、単語の特徴ベクトルを用いて計算される。グループの特徴ベクトルはそのグループに属する各 WWW ページの特徴ベクトルの和で定義される。

*Weak Organization and Area-view Function of WWW Information Space

†Shinichi Fukushima, Mitsuru Ishizuka

†University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail : shin@miv.t.u-tokyo.ac.jp

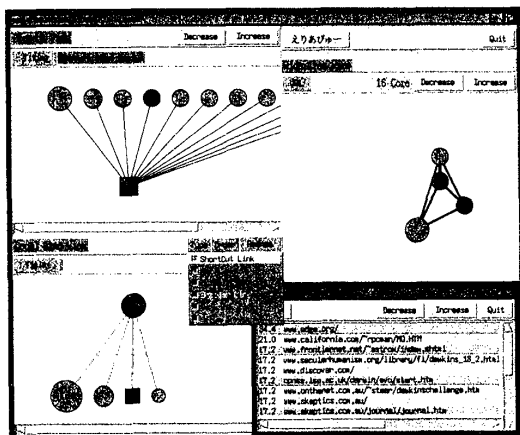


図 2: システムの概観図

3 エリアビュー機能

WWW 情報空間の視覚化は、閲覧している WWW ページの周辺情報を提示することによりユーザのブラウジングを支援することを目的とする技術である。従来の視覚化の研究では、多数の WWW ページやリンクが存在する場合に画面が複雑になるという欠点があった。そこで、本研究ではこの問題を解決することを目的として、上述した情報空間の弱い構造化を用いた視覚化システムを開発した(図 2)。

3.1 システム概要

本システムでは、すべての WWW ページを視覚化するのではなく、コアページを中心に表示することにより、多数のページが同一画面上に表示されて複雑になってしまう問題を解決している。以下、システムの具体的な機能を紹介する。

所属グループの視覚化

現在閲覧しているページが所属しているすべてのグループのコアページを表示することができる。(図 2左上)この機能により、閲覧している WWW ページが属するグループを視覚的に把握できる。コアページを選択し、グループ内の他の WWW ページを参照したり(図 2左下)、次に紹介する Wide Area View に選択したコアページを配置することができる。

コアページの視覚化

コアページを視覚化する Wide Area View(図 2右)では、コアページ間の関連度を視覚的に把握することができる。円で表されたコアページは、重みの

大きさと円の半径が決められる。また、コアページ間を結ぶ線はそのグループ間の関連度を表わしており、関連度が大きいほど太く表示されている。新たなコアページが配置される際には、関連度の高いグループのコアページの近くに配置されるようになっている。

グループ内の視覚化

グループ内の WWW ページを表示する。この際、当研究室の研究 [2] によるリンク分類情報により、ある意味のリンクで結ばれたページのみを表示することができる。また、ブラウザと連携しており、選択したグループ内の WWW ページをブラウザで表示されるようになっている。

関連度の高いグループを検索

コアページを選択し、そのグループと関連度の高いグループを検索し、Wide Area View に配置することができる。この機能はユーザが興味のある未知のグループを発見する支援を目的としている。

3.2 実装

現在、エリアビューは PC の Linux 上で実装されている。インタフェースは Tcl/Tk で実装され、WWW 情報空間の弱い構造化をは Perl で処理されている。理想的にはすべての WWW ページを保持し、弱い構造化を行うことが望ましいが、国内だけでも 1800 万ページ以上もあるとも言われている WWW ページを蓄積するのは困難であるため、本システムでは、ある程度の WWW ページを蓄積しておき、必要に応じてデータを取得する方法が用いられている。

4 むすび

本稿では、WWW 情報空間の弱い構造化を用いた視覚化システム(エリアビュー機能)を提案し、実装したシステムを紹介した。このシステムにより従来の研究では実現が難しかった WWW 情報空間を広い範囲で視覚化することが可能となった。今後は多くの WWW ページを収集して、検索システムとしての機能を設けたい。

参考文献

- [1] 福島, 石塚 “WWW 情報空間におけるコアページの抽出と弱い構造化”, 第 57 回情処全大, No.5F-08, 1998.10.
- [2] 小野田, 土肥, 石塚 “WWW ハイパーリンクの意味による分類とノードリンク構造の提示”, 第 56 回情処全大, No.1Z-03, 1998.3.