1Z－3

# 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*

小野田 浩平† 土肥 浩† 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) 情報空間の大きな特徴として、HTML 文書をノードとするネットワークにより構成される点が挙げられる。そのため、情報発信者により作成されたハイパーリンクの「意味」を解析することにより、WWW 情報空間における情報収集、ブラウジング、情報組織化など様々な操作の効率化が可能になると考えられる [1]。

そこで本研究では、ハイパーリンク情報の解析を行なうことによってハイパーリンクをその意味から分類し、WWW 情報空間のノードリンク構造を視覚化するツールの開発を行なった（図 1 参照）。

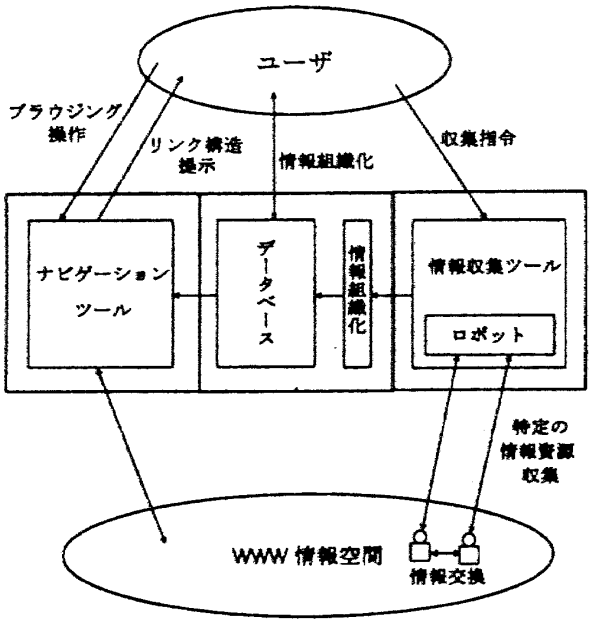


図 1: システム概要

*Semantic Classification of WWW Hyper-Link and Visualization of Node-Link Structure.

†Kohei Onoda, Hiroshi Dohi, Mitsuru Ishizuka

‡University of Tokyo,7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan

e-mail : onoda@miv.t.u-tokyo.ac.jp

## 2 ハイパーリンクの意味解析・分類

以下では、ハイパーリンクの意味解析、分類処理の具体的手法を説明する。

### 2.1 ハイパーリンクの意味解析

ハイパーリンクの意味解析処理においては、以下の 2 つの情報を使用する。

**(a) リンクの URL・アンカー文字列**

HTML 文書から、明示的なタグによってマークされているハイパーリンクの URL・アンカー文字列の抽出を行い、解析を行う。

URL ・アンカー文字列の解析によって得られる情報には、データ量としては少ないが重要な情報が含まれている。とくにリンク先が非 HTML 文書の場合は、リンク先の拡張子を抽出することによってリンクの意味解析が可能である。また、アンカー文字列にはリンク先の情報資源を端的に表す言葉が含まれているため、それを解析することによって有用な情報が得られる。

**(b) HTML 文書のタイプ分類**

リンク集のリンクなどは、URL ・アンカー文字列を用いた方法では解析が困難である。その場合には、HTML 文書のタイプ分類を用いたハイパーリンクの意味解析が必要になる。

そこで、その HTML 文書の持つ「固有の情報量」、「リンク情報量」をテキスト量やハイパーリンクの数などを用いたヒューリスティックにより決定して分類を行う。あるトピックについての情報へのリンクを集めた HTML 文書は大きな「リンク情報量」を持つのに対し、論文のアブストラクトなどのようなそれ自体が情報を持つ HTML 文書は大きな「固有情報量」を持つ。

このような HTML 文書のタイプ分類を行った上で、その中に含まれるハイパーリンクの意味解析を行う。

### 2.2 ハイパーリンクの分類

以上の解析により抽出した特徴量を用いて、ハイパーリンクの分類を行う。ハイパーリンクの分類カテゴリーを以下に示す。

(a) 非 HTML 文書
リンク先の URL を解析することにより分類。

(b) ページ内移動
リンク先の URL を解析することにより分類。

(c) リンク集
固有の情報が少なく、リンク情報が多いと判断された HTML 文書に含まれるハイパーリンクが該当。他サーバーへのリンクである。

(d) 関連する情報
関連情報を提供している、他の WWW サーバに存在する WWW ドキュメントへのハイパーリンクが該当する。HTML 文書の一部がリンク集のような構成になっている場合、そこに含まれるハイパーリンクがこのカテゴリーに分類される。

(e) 語彙説明
HTML 文書中の単語がアンカー文字列となっており、その単語の説明がリンク先に存在する場合が該当。特徴としては、アンカー文字列が文中の単語であること、同一サーバー内へのリンクであることなどが挙げられる。

(f) 詳細化
WWW サーバのトップページなどに存在し、目次的な使われ方をするリンクである。WWW 情報空間に存在するハイパーリンクとしては、もっとも多く見られる形態であろう。特徴は、リンクが単独で存在、トップページに存在、同一サーバー内へのリンクであることなどが挙げられる。

(g) ショートカット
同一サーバー内の WWW ドキュメント間の移動のために作られたリンクが該当する。トップページなど目次的な WWW ドキュメントへ戻りたいときなどによく利用される。特徴は、リンクが単独で存在、WWW ドキュメントの最初または最後に存在、同一サーバー内へのリンクであることなどが挙げられる。

## 3 WWW 情報空間の視覚化

本研究では、以上述べたハイパーリンクの意味解析・分類処理用ツール、及びその結果を用いて WWW 情報空間を視覚的に表示できるツールを開発した（図 2 参照）。使用した言語は、Tcl/Tk および Perl である。

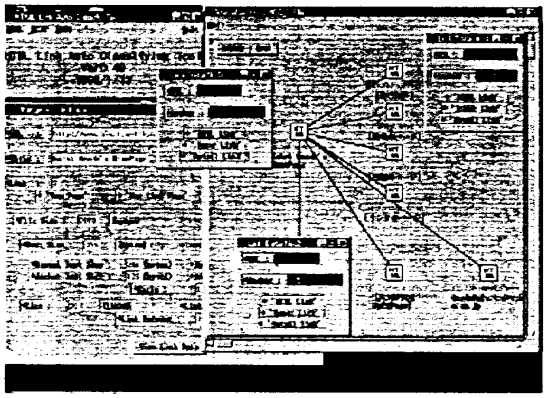

図 2: ツールの概観

WWW 情報空間視覚化ツールで実装した機能は以下のものである。

1. **ある WWW ドキュメント周辺ノードリンク構造の提示**
   リンク先の見通しを良くするという目的で、現在表示している WWW ドキュメントからリンク数段分の範囲のノードリンク構造を表示する。

2. **ハイパーリンクの意味に応じた色分け**
   直観的に WWW 情報空間のノードリンク構造を把握するために、ハイパーリンクの意味解析・分類によって得られた情報を用いてリンクの色分けを行なう。

3. **特定の意味のリンクでつながれた WWW ドキュメントのみ表示**
   少しでも複雑なノードリンク構造を表示すると画面が見にくくなってしまうため、特定の意味をもつハイパーリンクとそれにつながる WWW ドキュメントのみを表示する。

## 4 おわりに

本稿ではハイパーリンクの意味による分類手法を提案し、分類結果を用いて WWW 情報空間のノードリンク構造を視覚化するツールを作成した。

今後はハイパーリンクの意味解析・分類処理の効率化、ユーザのアクセス履歴情報の利用、および WWW 情報空間の意味ネットワークによる組織化等の検討を進めていきたい。

## 参考文献

[1] 小野田, 土肥, 石塚 “ハイパーリンクの意味理解と意味ネットワーク形状への組織化”, 第５５回情処全国大会, Sep. 1997.