4Q−11

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*

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

近年のインターネットの普及に伴い、WWW (World Wide Web) 上で提供される情報は多様化・複雑化・大規模化の一途をたどっている。どのような情報がどこから発信されているかを統一的に管理する機関は存在しないため、ユーザが必要な情報へ効率よくアクセスすることは非常に困難になってきている。

WWW ブラウザを用いて情報にアクセスしようとする場合、その時点で閲覧している WWW ドキュメントから直接リンクされている WWW ドキュメントしか把握できない。このため、必要な WWW ドキュメントが現在閲覧している WWW ドキュメントから数段ハイパーリンクをたどった先にあることがわかっている場合でも、ハイパーリンクをしらみつぶしに探すことになり非効率的である。また、リンク先の WWW ドキュメントの内容を表現するものは、リンク元のアンカー文字列だけである。そのため、実際にハイパーリンクをたどった先の WWW ドキュメントの内容が、想定していたものと大きく隔たっているという状況が頻繁に生じてしまう。

そこで、本研究ではユーザのブラウジング環境の向上を目的とした WWW ドキュメント間のノードリンク構造を視覚化するツールを開発中である（図1参照）。本稿ではそのツールに実装する手法として、WWW ドキュメント間を関連づけているハイパーリンクの意味を理解することにより、WWW 情報空間を意味ネットワークの観点から組織化する手法を提案する。

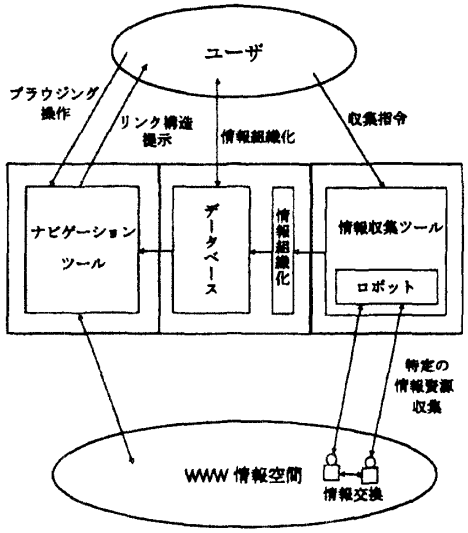

図 1: 本研究で想定するシステム概要

2 ハイパーリンクの意味理解

ハイパーリンクの意味理解が実現されることによって、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。以下では、ハイパーリンクの意味理解処理の具体的手法を説明する。処理の流れを図2に示す。

2.1 ハイパーリンクの意味解析

意味解析には、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

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。

*Semantic Understanding of Hyper-Link and Organization for Semantic Network.

†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

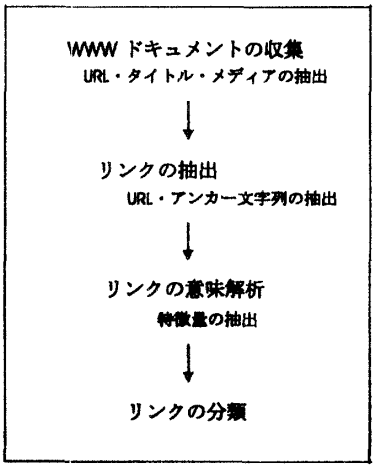

図 2: ハイパーリンクの意味理解処理の流れ

(b) HTML 文書の内容
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。その上で、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。

(d) 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。

(e) 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。

(f) 関連する情報
関連情報を提供している、ほかの 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。

3 意味ネットワークによる組織化

情報組織化は情報間の共通性を見いだし、その共通性を表現することにより、情報の質的な向上を目指す操作である。現在、様々な情報源から機械的に情報を取り込んで、それをユーザが修正するという形で組織化が行われている。

ハイパーリンクの意味理解が実現されれば、情報組織化の操作において意味ネットワーク [1] の考え方が利用できるようになる。WWW ドキュメントをノード（接点）、ハイパーリンクをリンク（有向枝）としてあらわすことにより、WWW 情報空間を意味ネットワークの形に変換することができる。その際にハイパーリンクの意味を解析し、概念の包合関係を示す "is-a" 関係や、部分-全体の関係を示す "part-of" 関係などに対応させることによって、活性化伝搬や性質継承の考え方を用いた計算機による高次利用が可能になると考えられる。

4 おわりに

本稿では、WWW 情報空間における様々な操作の効率化を目的として、WWW ドキュメント間を関連づけているハイパーリンクの意味理解を用いた手法を提案した。計算機によるハイパーリンクの意味理解が実現できれば、現在 WWW 情報空間で計算機の支援によって行われている様々な操作の効率を向上することができると考えられる。

今後は、本稿の手法を実装した WWW ドキュメント間のノードリンク構造を視覚化するツールを作成し、特定の意味を持つハイパーリンクだけを提示する機能、ユーザの履歴を用いた WWW ドキュメントへのアクセス頻度提示機能などを考慮して、研究を進めていく予定である。

参考文献

[1] Quillian, M. R. : Semantic memory. In M. Minsky (Ed.), Semantic information processing. Cambridge, Mass. : MIT Press, 1968.