5F－8

WWW情報空間におけるコアページの抽出と弱い構造化 *

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) の作る情報空間は、近年のインターネットの爆発的な普及・浸透に伴い、多様化・複雑化・大規模化の一途をたどっている。我々はこの広大な WWW 情報空間を利用する際はブラウザを用いるが、通常のブラウザは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ず、手がかりの少なさの問題がある。これを解決する手法として WWW 情報空間を視覚化する研究 [1] が成されているが、これらのほとんどは、画面の大きさの制約から周辺の情報のみを視覚化するにとどまっている。

そこで、我々は、より広い視点で WWW 情報空間をとらえ、物理空間の位置に相当する情報を WWW 情報空間上に構築することを目指している。本稿では、その基礎技術となる、WWW 情報空間のリンク結合構造からある地域の中心となるコアページを抽出し、それを基に弱い構造化を行う手法を提案する。

2 弱い構造化

WWW 情報空間は、WWW ドキュメントをノードとするネットワーク構造を成している。そこで、我々は WWW の特徴であるネットワーク構造を利用して、WWW 情報空間を弱い枠組みで構造化する手法を提案する。具体的には、リンク結合構造から抽出したコアページを中心に複数のグループを形成し、さらに、関連するグループ同士で新たなグループを作成していくことによりボトムアップに弱い構造化を行う。(図 1)

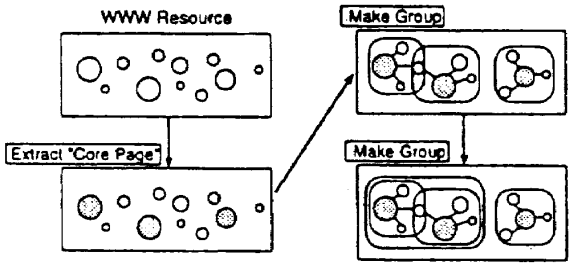

図 1: 弱い構造化

2.1 コアページの抽出

コアページは、周辺ページの中心的存在であることを考えると、参照リンク数の多いページがその候補となり得る。しかし、単に参照リンク数でのみコアページの抽出を行った場合、「目次の戻る」や、個人のリンク集による検索エンジンへのリンクなどにより、コアページとして適当でないページが抽出されてしまう可能性がある。そこで、本研究では以下の手順によりノードの重み付けを行ない、重みの大きいノードをコアページとして抽出する。(図 2)

(1) リンクの重み付け
外部より参照されるリンクの重みを大きく、内部より参照されるリンクの重みを小さく設定する。これにより、「目次に戻る」のリンクの影響を小さくできる。

(2) ノードの重み付け
参照されるリンクの重みの和のより、ノードの重みを決定する。

---

*Extracting Core Pages for Weak Organization 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

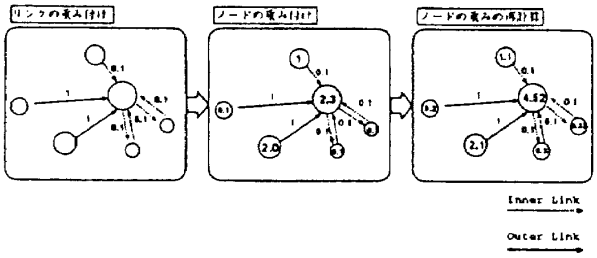

図 2: コアページの抽出

(3) ノードの重みの再計算
参照されるノードとリンクの重みの積をノードに加算する。これにより、個人のリンク集の影響を小さくすることができると考えられる。なぜなら、個人のリンク集は、内部リンク一つで結ばれることが多く、そのためノードの重みが小さくなるからである。

2.2 グループ化

コアページ抽出後、そのページを中心にリンクで結ばれたページで一つのグループを生成する。この際、コアページからどれぐらいの距離にあるページまでを一つのグループに含めるかが問題となるが、本研究では以下のような制限を設ける。

(1) コアページの直接参照、被参照リンクページ

(2) コアページから内部リンクで結ばれているページ

(1) の条件を設けることにより、一つのグループがいたずらに増大し、グループ内の情報のまとまりがなくなるのを防ぐ。(2) は、どのグループにも所属しないページが現れるのを極力避けるための条件である。但し、グループ内にコアページが出現した時は、そのコアページから内部リンクで結ばれているページは含めない。その理由は、(1) の条件同様に一つのグループが巨大化するのを防ぐためである。また、一つのページが複数のグループに属することも考えられるが、その点に関しては特に制限は設けず、複数のグループに属するページが現れてもよいとする。

2.3 グループ間の関連度

以上より複数のグループが生成されるので、次はグループ間の関連度を調べ、関連度の高いグループ同士をまとめて新たにグループを生成する。つまり、物理空間において東京と大阪を日本という一つの枠組みでとらえるのと同じように、複数のグループを一つの大きなグループでまとめることにより、広い視点で WWW 情報空間をとらえることが可能となる。

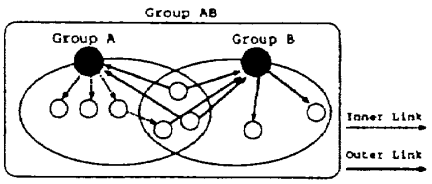

図 3: グループ間の関連度

具体的には、2つのグループ間で相互に参照するページが多い場合や2つのグループに属するページが複数存在する場合、つまり、コアページをつなぐページが多数存在する場合に2つのグループの関連度が高いと判断する。これは、関連するページは同一のページから参照されることが多いというヒューリスティックに基づく。[2] (図 3) 例えば、研究分野の似ている研究室があった場合、両研究室間でリンクが張られたり、両研究室へリンクを張るページが存在する可能性が高いと考えられる。

3 おわりに

本稿では、従来の研究では考慮されていなかった広い視点で WWW 情報空間をとらえ、その基礎となる技術として、WWW のリンク結合構造を用いたコアページの抽出と WWW 情報空間の弱い構造化を提案した。

今後は、本稿で提案した手法を基に広い範囲を視覚化するツール（エリアビュー）を作成する予定である。

参考文献

[1] 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 5 6 回情処全国大会, Mar. 1998.

[2] T.Joachims, T.Mitchell, D,Freitag, R. Armstrong :WebWatcher: Machine Learning and Hypertext, Fachgruppentreffen Maschinelles Lernen, Dortmund, Germany, Aug. 1995.