WWW 上のリンク構造を用いた情報検索に関する研究

大島 龍之介

北陸先端科学技術大学院大学 情報科学研究科

平成 12 年 2 月 18 日

キーワード: リンク構造, WWW, ウェブ, 情報検索, 検索エンジン.

World Wide Web(以下、Web と略する) は、多くの人にとって、情報の利用が簡単なだけでなく、情報の発信の敷居も低い。そのために、実に雑多で質もばらばらな情報が、時々刻々と追加・更新されている。しかし、Web の機構自体には、扱っている情報を分類・整理する機能が備わっていない。結果として、膨大で雑多な Web 上の情報を利用者は持て余してしまい、利用者・発信者、お互いの意図に沿った情報の利用が困難になっている。

Web の利用者・発信者が、おのおのの目的に沿って Web 上の情報を利用できるようにするためには、情報を分類・整理して利用者に理解し易い形で提示し、利用者を誘導する仕組みが必要とされている。Web 上の情報を適切な基準で分類、グループ化、階層化、順序づける仕組みは、利用者・発信者が把握・理解できる情報の量を増すための大きな助けとなるのである。

Web 上のページの分類・整理の方法は従来より、自然言語処理の分野である情報検索の分類の方法から盛んに研究されてきている。その一方で、Web の特徴であるリンクから分類をする方法はあまり研究されてこなかった。本研究ではこのリンクによる Web の分類・整理がどのように有効かを示す。そして、Web 上のリンク構造情報の抽出、及び、抽出したリンク情報構造の有効利用を目的とする。また逆に、リンクによる Web の分類・整理の有効性を示すことで、発信者がより有効にリンクを使うことが期待される。

本研究では、リンク構造の情報を Web のページより抽出する方法を提案する。構造的なリンクの性質を詳細に検討し、その結果をふまえて、Web ページの間の構造的な関係を抽象する「距離」と「影響度」という 2 つの尺度を定義する。「距離」と「影響度」によって、ページをグループや階層にわける方法も示す。

ページに含まれる、語や HTML タグによる文章情報からも、リンク構造の情報を見つけ出す。共通語の出現頻度からなる「類似度」を定義し、さらに、Web 上の各文章の関係を「関連度」「抽象度」「相対度」の 3 つの尺度で表す。これらの尺度は、リンクのアンカーが存在する位置、テキストや HTML 内での構造、他のリンクとの関連などの様々な要素を総合して決定する。

これらの抽出したリンク構造情報に基づいて、ページ群をグループ化、階層化、順序づけをおこなう方法を提案する。グループ化は近い「距離」の強連結グループを基本としている。さらにグループ内の代表ページの発見を「影響度」を使っておこなう方法を説明する。「関連度」・「類似度」を使って、グループ化はさらに強化される。「抽象度」を用いての情報の階層構造を発見や、「相対度」による情報の順序付けの求め方で、グループ全体の把握を容易なものとする。さらに実際の Web 上のページを実例として、これらのグループ化、階層化、順序づけの説明をする。

本研究では、Web 上のリンク構造情報による実際の重要な応用例として、Web 上の情報検索の改善方法を提案する。Web 上の情報検索では、検索エンジンと呼ばれる、単語をキーとする全文検索システムが従来より使われてきている。しかし、ある検索の目的に内容が適切でも検索のキーの単語が出現しないために検索の結果に表れないページや、内容には関係のない検索のキーの単語がたまたま含まれた多数のページの中に、適切なページが埋もれてしまうなどの問題がある。

単語をキーとする全文検索にリンク構造情報を加えることにより、より適切なページを取り出し、冗長なページを除去することができる方法を提案する。単語をキーとする検索結果のページ群とそのリンク先のページのリンク構造情報により、グループ化をおこなう。グループ間、及び、グループ内でそれぞれ階層化、順序づけをおこなう。グループの大きさとグループ間の順序づけを単語をキーとする検索結果の得点に加えて、新しい全体の順序を決定する。検索結果の提示では、グループはグループ内の順序づけで上位の情報を使用して、一部の冗長な情報を除去する。各情報をグループ、順序、階層や距離に応じた視覚化、色づけをおこなう。

本論文で提案した方法の実験・評価のために、プロトタイプの設計と実装をおこなった。全文検索部には、既存の全文検索システムを流用し、リンク検索部をその上に実装した。各ページの得点とリンク構造から、代表ページを決定する。代表ページ以外のグループ内のページを除去しながら、高得点順に利用者に検索結果を返す。実際の Web 上の情報と、既存の検索エンジンの結果を実例に用いて、本提案の有効性を検証した。結果として、既存の全文検索システムの検索結果を、より適合率が高く、情報量の多い結果へと改善できることが確認できた。また、Web 上にはリンク構造によるページのグループが多数存在し、本研究の手法を幅広く適用することが可能であることも確認できた。

本研究の結果、Web 上のリンク構造情報の抽出の方法を提示し、その有効性を示した。また、情報検索において抽出したリンク情報構造の有効な利用方法を示すことができた。リンク構造情報は本研究で示した情報検索にかぎらず、情報の要約、情報の新たな構造の追加や再構成など、様々な分野に応用されることが期待される。