論文内容の要旨

論文題目　リンク情報を利用したWeb文書クラスタリングに関する研究

氏　名　正田　備也

本研究では、ハイパーリンクを利用したWebページのクラスタリング手法を提案する。そして、提案手法によって得られたクラスタリングを、ネットサーフィンをナビゲートするために応用した場合、および、Web検索での検索結果の順序を付け直すために応用した場合の実験結果を提示する。

Webは、一つの巨大な有向グラフとみなすことができる。そこでは、Webページが頂点、ハイパーリンクが有向枝に対応する。本研究では、Webのリンク構造が与えるこのグラフ上で、Webページ間の相互参照が遍在する事実、つまり、ハイパーリンクがいたるところで有向閉路を構成している事実に着目する。学術文献やネット・ニュースなどの文書集合も、Webと同様、文書間の参照関係を含んではいる。しかし、参照関係が時系列順に生成されるため、相互参照はほとんど観察できない。したがって、本研究は、Webという文書集合に固有の特徴に着目していると言える。

クラスタリング・アルゴリズムを設計するにあたっては、ハイパーリンクに非負実数の重みが与えられていることを前提とする。リンクの重みとしては、リンクの始点となっているWebページの出次数、オーソリティ・スコア、ハブ・スコアなど、Webのリンク構造から得られる数値を使用する。さらに、リンクで結ばれた二つのWebページ間のテキスト内容上の類似度も、リンクの重みとして使用することを試みる。そして、有向閉路の長さ、すなわち、Webページ間の相互参照の長さを、それに沿って存在するハイパーリンクの重みの総和と定める。

本研究では、Webページ間に次のような新しい距離を導入する。二つのWebページ間の距離を、それらを経由する様々な有向閉路の長さの最小値として定義する。これを相互リンク距離と呼ぶ。なお、二つのWebページを全く有向閉路が通過しない場合、それらの間の相互リンク距離は無限大とする。このように、相互リンク距離は、Webに遍在する相互参照に基づいて、Webページ間の近さを定量化している。

本研究の提案するクラスタリング・アルゴリズムは、次のようにして、相互リンク距離を利用する。任意に選ばれたWebページを中心として、相互リンク距離において予め定められたパラメータの値 τ 以下だけ離れているページを、同じクラスタにまとめる。このパラメータを閾値パラメータと呼ぶ。ところで、相互リンク距離は、相互参照に基づくWebページ間の近さの定量化であった。したがって、閾値パラメータを大きくしていくと、提案手法の与えるクラスタリングは、強連結成分分解に近づいてゆく。なぜなら、どのような長さであれ、とにかく相互参照でつながりあっているWebページを一つにまとめ上げてゆくと、強連結成分分解を得るからである。つまり、閾値パラメータを適度な値に設定し、提案のクラスタリング手法を実行させると、強連結成分分解の細分化として、クラスタリングを得ることができる。そして、クラスタの粒度は、閾値パラメータの値を大きくすれば粗くなり、小さくすれば細かくなる。閾値パラメータによって制御された粒度でもって強連結成分を細分化していく、という意味で、提案手法を、パラメータ化された連結性に基づくクラスタリングと呼ぶことができる。

提案のクラスタリング・アルゴリズムでは、具体的には、クラスタの中心として選ばれたWebペ

ージから、SSSP (single source shortest paths) 問題を解くことによって、クラスタの構成員となるWebページの枚挙を行っている。SSSP問題とは、有向グラフ上で定義される問題であり、与えられた一つの頂点から、他のすべての頂点への最短距離を算出するという問題である。SSSP問題は、ダイクストラのアルゴリズムを、フィボナッチ・ヒープ等の効率的なヒープとともに用いるならば、O(n log n + m)の計算量で解くことできる。ここで、nはWebページの総数、mはハイパーリンクの総数である。そして、クラスタの中心として選ばれるページの総数は、高々O(n)個である。したがって、クラスタリングに必要な計算量はO(n^2 log n + mn)となる。だが、実際の計算量は、この上界よりもさらに少ない。なぜなら、提案のアルゴリズムでは、クラスタの中心として選ばれたページから、SSSP問題を完全に解く必要はなく、閾値パラメータで制限された範囲内にあるWebページについてのみ、最短距離を計算すればよいからである。

提案手法によって得られたクラスタは、その直径、つまり、クラスタに含まれるWebページ間の相互リンク距離の最大値が、必ず閾値パラメータの2倍以下となっている。この事実は、相互リンク距離が三角不等式を満たすことより証明できる。クラスタの大きさの上限が、閾値パラメータという、予めその値の分かっている数値によって決定されることは、提案手法の与えるクラスタリングの粒度が、直感的に理解しやすい仕方で制御されていることを意味する。

本研究では、提案手法によって得られたクラスタリング結果を、Webから得られる情報の有効利用へと応用した実験結果を示す。いずれの実験も、NTCIR-3ワークショップに向けて公開されたデータNW100G-01を用いている。このデータは、約11,000,000のWebページと、約55,000,000のハイパーリンクを含んでおり、提案手法がWeb情報の有効利用に対してどれだけ寄与できるかを正当に評価するためには、十分な規模を持つ。また、実際にNTCIR-3 Webタスクで使用されたデータであるため、提案手法の妥当性をより客観的に評価することができる。

本研究が想定する第一の応用は、ネットサーフィンのナビゲーションである。そこでは、同じWebページに戻ってくることなく、できるだけ多くの新しいページを連続して見続けるネットサーフィンの実現を目指す。このようなネットサーフィンを、新奇探索型ネットサーフィンと呼ぶ。本研究では、ネットサーフィンをランダム・ウォークによってシミュレーションすることで、評価実験をおこなった。実験の結果によれば、ネットサーファが、できるだけ多くのクラスタ境界を越えつつ、できるだけ多くの相異なるクラスタを横断できるようにナビゲートされるとき、ナビゲーションの無い場合に比べて、より長く新奇探索型サーフィンを続けられることが分かる。

第二の応用は、Web上での情報検索における、検索結果の順位の付け直しである。そこでは、テキスト情報のみに基づく検索で得られた文書の順位付けを、提案手法の与えるクラスタリングの結果を利用して修正し、検索性能を向上させることを目指す。個々のクラスタは、次のようなかたちで、Webページ同士がテキスト情報を部分的に交換し合う場として機能させる。つまり、クラスタが検索質問により良く適合するWebページを多く含むほど、そのクラスタ内のページが、他のクラスタ内のページに比べて、より質問に適合することになるように、各Webページから抽出された特徴量を変更する。そして、変更された特徴量を用いて、適合順位を計算し直す。実験結果は、クラスタリングを行うにあたって、Webページ間のテキスト内容上の類似性を考慮せず、Webページの出次数など、リンク構造から得られる情報だけを利用した場合も、検索性能の改善が得られることを明らかにしている。