Max-Flowアルゴリズムを基にした効率的なWebコミュニティ取得方法*

松葉 潤†　陳 漢雄‡　古瀬 一隆§

†,‡,§,¶ 筑波大学大学院システム情報工学研究科　〒305-8577 茨城県つくば市天王台 1-1-1

1 はじめに

近年、Web空間の拡大に伴い、Webページの数が膨大になってきている。このため、ユーザにとって興味のある情報を見つけることが難しくなっている。そこで、密にリンクされたWebページの集合を取得することは有益であると言える。このようなWebページの集合をWebコミュニティと言う。Webコミュニティはユーザが興味のあるWebページをシードページとすると、シードページと関連する内容を持つWebページの集合である可能性が高い。本研究ではこのWebコミュニティの抽出方法を提案する。

既存のWebコミュニティ抽出方法の一つにMax-Flowアルゴリズム[1]がある。しかし、Max-Flowアルゴリズムの特性上、十分な大きさのコミュニティを得るために、抽出したコミュニティをシードページに加えて再度このアルゴリズムを行うといった反復作業が必要となる。そのため、多くの時間がかかってしまうと推測できる。そこで本研究ではこのMax-Flowアルゴリズムに改良を加える。Webグラフの構築の段階でコミュニティに含まれそうなページを優先的に選択することで反復を無くし、より効率的にWebコミュニティを抽出する手法を提案する。

2 提案手法

2.1 Sスコア

先に述べた既存のWebコミュニティ抽出方法であるMax-Flowアルゴリズムは密に繋がっているページ集合を抽出するにはとても有効であると言える。しかし、必要なサイズのWebコミュニティを得るにはグラフ構築とコミュニティ抽出の作業を何度も繰り返さなければならない。この繰り返しの作業に多くの処理時間がかかってしまうと想定できる。

本研究ではWebグラフ構築の段階でコミュニティに含まれる可能性が高いページを優先して取得することで繰り返しの作業を無くし、一度のコミュニティ抽出でWebコミュニティを抽出する手法を提案する。

本研究ではコミュニティに含まれる可能性が高いページを発見するために、フローの流れ出しやすさを表現する値としてページにSスコアという値を付与する。ページuのSスコアは以下の計算式で求める。

$$S(u) = \frac{1}{\sum_{\omega \in W} \frac{1}{S(\omega)/out\_degree(u)}} \quad (1)$$

Wはuへのリンクを持つページ集合とする。

この式は「ページuのSスコアはuのリンク元のSスコアをuの出リンク数で割った値の逆数」ということを意味する。

このSスコアが小さい程フローがそのページの先の辺で詰まり、そこでボトルネックになっている可能性が高い。よってフローが流れ出しやすいと考えられるSスコアの大きいページのリンク先をクロールすることとする。Sスコアはクロールと同時に計算される。そして次にクロールする際はSスコアの大きい一定数のページからのみクロールする。

2.2 アルゴリズム

具体的な手法を説明する。まず、各シードページへ向かう辺を持つ仮想始点を与える。仮想始点のSスコアは1とし、これを元にシードページのSスコアを計算しておく。それからシードページから距離1のページをクロールし、クロールしたページのSスコアを計算する。クロールしたページをSスコアの降順でランキングして上位の一定数のページを選択する。選択したページのみからクロールを行い、再度クロールしたページのSスコアを求める。ここで以前選択されなかったページも含めSスコアの降順でランキングし、同様にページ選択を行い、クロールする。以上の作業を指定した十分な大きさのWebグラフが構築されるまで行う。

図1はSスコアによるグラフ構築の例である。この例ではページ選択の際3つのページを選択している。

その後Max-Flowアルゴリズムと同様に構築されたWebグラフに仮想終点を加え、フローを流せるだけ流し、フローが飽和している辺でグラフを切り離しコミュニティを抽出する。

また、Max-Flowアルゴリズムと違い、距離2までのページでは無く十分な大きさのWebグラフを構築するので一回のコミュニティ抽出で処理を終了する。

*Improving Max-Flow Algorithm for Efficient Discovery of Web Communities

†Jun Matsuba, Graduate School of SIE, University of Tsukuba

‡Hanxiong Chen, Graduate School of SIE, University of Tsukuba

§Kazutaka Furuse, Graduate School of SIE, University of Tsukuba

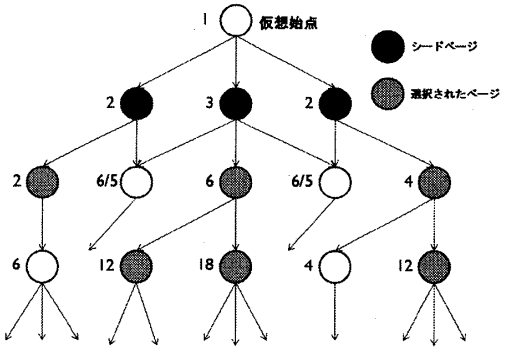

図 1: S スコアによる Web グラフ構築

2.3 係数の付与

S スコアはフローが流れ出しやすいページを発見するには有効であると言えるが、その特性上シードページから離れる程値が大きくなる。そのため S スコアにおけるページ選択では値が大きいページを選択していくので、一度選択されなかったページは以降のページ選択でも選択されない可能性が高い。そこでシードページから距離が遠い程 S スコアを小さくするような係数を付与する。

$$S(u) = \frac{1}{E^d \sum_{\omega \in W} \frac{1}{S(\omega)/out\_degree(u)}} \quad (2)$$

d はシードページからの距離である。この係数により E の値が大きい程シードページからの距離が遠いページの S スコアが小さくなる。つまりシードページから近いページの S スコアが大きくなり選択されやすくなる。従って E の値が大きい程浅く、広いコミュニティが抽出されることになる。逆に E の値が小さい程深く、狭いコミュニティが抽出される。

3 実験

3.1 実験概要

本研究の実験では Max-Flow アルゴリズムと提案手法の比較実験を行った。取得したコミュニティの一致率、処理時間、クロールしたページ数について比較を行う。

各種設定について説明する。シードページは同じトピックに沿ったページを選択する必要がある。そこで、本研究では Yahoo!カテゴリ内の「トップ→趣味とスポーツ→テニス」から 3 ページを選択する。また、本研究で経験的に良い結果を得ることができた以下の設定を用いる。まず、Max-Flow アルゴリズムにおいて再度 Web グラフを構築する際に増加させるシードページの数は 3 ページとする。そして、提案手法における S スコアによるページ選択は、クロールリストを S スコアの降順に並び替えた時の上位 10 ページを選択することとする。

3.2 実験結果

Max-Flow アルゴリズムと提案手法の処理時間による比較実験を行う。取得するコミュニティのサイズを変化させて各々の処理時間を比較した結果を図 2 に示す。

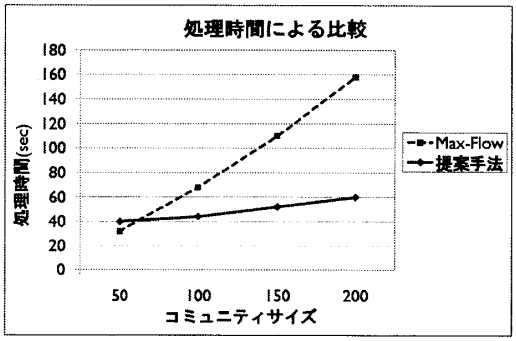

図 2: 処理時間による比較

提案手法により処理速度の向上が確認できた。これは Max-Flow アルゴリズムではコミュニティサイズが大きくなる程、Web グラフ構築とコミュニティ抽出の反復をする回数が多くなり処理時間が増加している一方で、提案手法では反復を行わないためだと考えられる。

4 まとめと今後の課題

本研究では既存の Web コミュニティ抽出方法である Max-Flow アルゴリズムを基により効率良くコミュニティを抽出する手法を提案した。

そして、各種比較実験を行い、Max-Flow アルゴリズムと同様のコミュニティを効率良く取得できていることが確認できた。

今後の課題としては各パラメータを変えて詳細な比較実験を行いたい。また、行った実験の結果を基に Max-Flow アルゴリズムで取得できるコミュニティとの一致率を向上させることを目指す。

参考文献

[1] G. Flake, S. Lawrence, C. L. Giles, "Efficient Identification of Web Communities", Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.150-160, 2000.

[2] L. R. Ford Jr. and D. R. Fulkerson, "Maximal Flow Through a Network", Canadian Journal of Mathematics 8, pp.399-404, 1956.

[3] N. Imafuji, M. Kitsuregawa, "Finding Web Communities by Maximum Flow Algorithm Using Well-Assigned Edge Capacities", IEICE Transactions on Information and Systems. Vol. E87-D No2, pp.407-415, 2004.

[4] Yahoo!カテゴリ, http://dir.yahoo.co.jp/