ウェブコミュニティ抽出における最大フローアルゴリズムの最適利用とその効果
Efficient Usage of Maximum Flow Algorithm for Extracting Web Community

D-13

今藤 紀子† 喜連川 優†
Noriko Imafuji Masaru Kitsuregawa

1. 研究の背景

ウェブコミュニティは似たようなトピックを扱うウェブページ集合を意味し、それを抽出するための種々の手法が提案されている。最大フローアルゴリズムを利用した手法もそのうちの一つで、シードとして与えたノードから数リンク離れたノードや入出次数の少ないノードも適切であればウェブコミュニティのメンバーと成り得るという利点を持つ。その一方、不適切なノードが入り込んでしまう場合が多くみられ、この手法の致命的欠点も実験結果から明確になった。そこで、本研究の目的は、最大フローアルゴリズムを利用した抽出手法をベースに、ウェブコミュニティへの不適切なノードの侵入を防ぐための処理を加えた新たな手法を提案することにある。

2. ウェブコミュニティ抽出手法

ウェブコミュニティを抽出するための手法はそのアプローチの違いから2つのタイプに大別できる。一つ目はリンク解析によるアプローチで、Kleinberg提案のハブ、オーソリティの概念を利用したHITSアルゴリズムに基づいた手法が主流である[3]。二つ目は、グラフ論的アプローチによる手法で、ウェブページ、ハイパーリンクをそれぞれノード、エッジとみなし、ウェブ全体を巨大なグラフ（ウェブグラフ）としてそのグラフ構造の解析からウェブコミュニティを得るというものである。代表的なものとしてKumarらによる完全2部グラフをウェブコミュニティのグラフ構造として利用したものなどがある[1]。本論文の主題となっている最大フローアルゴリズムを利用した手法もグラフ論的アプローチによるものである。

2.1 最大フローアルゴリズムに基づく手法

最大フローアルゴリズムに基づく手法は[2]によって提案された。[2]では、ウェブコミュニティのメンバーページを「コミュニティ内のノードへ（から）のリンク数が外のノードへ（から）のリンク数よりも多いノード」と定義し、このようなノード集合が最大フローアルゴリズムの解として得られることを証明している。この手法は以下のような手順で実現される。

入力: S; シードノードの集合
出力: C; ウェブコミュニティ
1: S の各ノードから深さ2の周辺グラフ ($G = (V, E)$) をクロール
2: 探索グラフ (G) を構築
3: G に対して最大フローアルゴリズムを施し、s から到達可能なノード集合 C を得る
4: C のノードを C 内のリンク数によってランク付けし上位ノードを S に加える
5: **1-4** を繰り返す

ここで、手順**2**における探索グラフは、$G = (V, E)$ を基に以下のように構築される；V に s,t（それぞれ仮想ソース、仮想シンクを意味する）を加える。E に s から S の各ノードへの辺（辺容量=∞）を加える。E の各辺（辺容量=シードノードの個数）を双方向にする。s,t 及び S のノードを除く各ノードから t への辺（辺容量=1）を加える。

実データを用いた種々の実験を通して上述の手法には以下のような問題点があることが明確になった。
・手順**3**において最大フローアルゴリズムが C にシードノード以外のノードを含まない、つまり新たなメンバーノードが得られない。
・探索グラフが巨大化し最大フローアルゴリズムに莫大な計算時間を費やしてしまう
・手順**5**における繰り返しの最適な打ち切り回数が与えたシードによって異なる

3. 提案手法

我々が提案する手法は前述の手法をベースにしており、その手順は以下のようになる。

入力: S; シードノードの集合
出力: C; ウェブコミュニティ
1: S の各ノードから深さ2の周辺グラフ ($G = (V, E)$) をクロール
2: 探索グラフ (G) を構築
3: G から最大のノード数で構成される完全2部グラフ $K(X, Y)$ を探索し集合 Y に属するノードとそれに付随する辺を G から削除する
4: G に対して最大フローアルゴリズムを施す
5: シードノード以外に s から到達可能なノードが存在しない場合、t への辺以外の辺の容量を増加させ**4**に戻る
6: シードノード以外に s から到達可能なノードが存在する場合、それらのノード集合を C とし C 内での入次数の大きさにおける上位ノードを S に加える
7: 手順**1-6**を繰り返す。但し、一つ前の繰り返しと同じ C が得られた場合、C 内の最上位ノードのスコアが突出して高い場合、G のサイズが不変になった場合、処理を終了する

†東京大学 生産技術研究所
{imafuji,kitsure}@tkl.iis.u-tokyo.ac.jp

4. 実験結果とその評価

実験には、30個の相異なるトピックを扱うウェブページをシードノードとして採用した。メジャーなトピックの場合、特別な手法を利用しなくてもウェブコミュニティに相当するようなウェブページの集合を得ることは容易であることから、実験で利用するトピックとしてはマイナーなもの、つまり、そのトピックを扱うウェブページの存在があまり明確でないものとなるように選択した。全ての実験は、2000年にクロールした国内のウェブスナップショットを利用して行った。提案手法の効果の解析は、同じシードノードに対して、もとの最大フローアルゴリズムを利用した手法 (**手法 1**)、提案手法-但し手順 **3** を行わない (**手法 2**)、提案手法 (**手法 3**) によってウェブコミュニティ抽出を試み、その結果の比較により行った。

4.1 実験結果

(1) 表1は、それぞれの手法によってどの程度の割合で適切なウェブコミュニティが得られているかを示している。(a) は、シードノード30個のうち、(サイズやメンバーページの内容等が) 妥当だと見なせるウェブコミュニティが結果として得られた個数の割合、また (b) は、妥当だと見なしたウェブコミュニティにおいて、その上位10のメンバーページが各シードノードに対して、ページタイトルやそのページ上の頻出語などを基に設定されたトピックキーワードを含んでいる割合の平均値を示している。

表 1: ウェブコミュニティ抽出における各手法の効果

	手法 1	手法 2	手法 2
(a)	23.3 %	56.6 %	76.6 %
(b)	74.2 %	85.9 %	88.3 %

(2) 図1は、実験を行った30個のシードノードの一つ、www.alive-net.net/ に関して手法 1,2,3 によって得られたウェブコミュニティの各メンバーページにおけるトピックを大別し、その個数の推移を示している。このシードは、動物保護に関するサイトである。3つの手法を通してこの実験での最終的な周辺グラフのノード数は約4800であった。

4.2 評価

妥当なウェブコミュニティが抽出実現の可否は手法の持つ問題点を直接示すものである。その点で、実験結果 (1) における表1(a) よりその実現割合が飛躍的に向上していることから提案手法によって、手法1の持つ根本的な問題点は解決できたと言える。また、抽出されたウェブコミュニティの質としての問題、つまりウェブコミュニティへの関連の無いウェブページの侵入は、表1(b) の結果より上位ノードにおいては、どの手法によっても誤差範囲内であると見なせるが、提案手法の方が良い結果

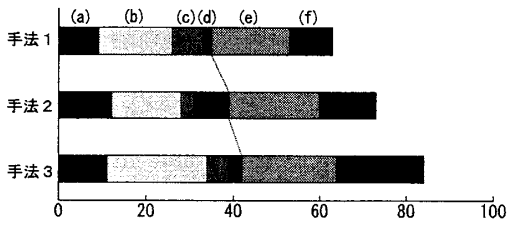

図 1: ウェブコミュニティメンバーページのトピック推移 (a) 総合情報が得られるサイト (b) ある特定の動物についてのサイト (c) 動物関係のリンクを少数持つサイト (d) 動物園、公園のサイト (e) 関連が無いサイト (f) 確認不可能なサイト

を返している。実験結果 (2) における図1のグラフより、手法3では、シードノードのトピックに深く関連しているが、入出次数の少ないある意味「マイナー」なページ（つまり、(b) のノード）が目立った。これは、探索グラフから完全2部グラフをあらかじめ取り除くことで、本来ならば最大フローアルゴリズムによって切断されてしまう部分での探索が可能になったために得られたページである。マイナーではあるがトピックとして深く関連しているページの抽出は既存の手法では限界があったが、そのようなページが抽出できたということは非常に意味のあることだと言える。

5. まとめ

本論文では、既に提案された最大フローアルゴリズムを利用した手法をベースに、新たなウェブコミュニティ抽出手法を提案した。実データを利用した実験を行い、その実験結果から提案手法によって既存の手法に見られたような問題点を回避しこれまででは得ることが不可能であったノードを含んだウェブコミュニティの抽出が可能となったことが実証された。しかしながら、ウェブコミュニティ内にはまだまだ関連の無いサイトの侵入が少なくはなくこのようなサイトの侵入を防ぐための処理を考える必要がある。純度の高いウェブコミュニティ抽出アルゴリズムを開発しそれを利用して抽出されたウェブコミュニティの成長過程の解析を行うことを今後の課題としている。

参考文献

[1] R.Kumar, P.Raghavan, S.Rajagopalan, and A.Tomkins, "Trawling the web for emerging cyber-communities," In Proc. 8th WWW Conference, 1999.

[2] G.W.Flake, S.Lawrence, and C.L.Giles, "Efficient Identification of Web Communities," In Proc. KDD 2000, 2000.

[3] J.M.Kleinberg, "Authoritative Sources in a Hyperlinked Environment," In Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.