ページ−コミュニティ間の関連性を考慮したWebコミュニティ抽出

斉田直幸 † 梅沢晃 ‡ 山名早人 †‡
早稲田大学理工学部情報学科 † 早稲田大学大学院理工学研究科 ‡

1 はじめに

膨大な情報を有する WWW において、WWW 上に存在する Web ページを意味的な集まりに分類、整理することは、Web の利用者への的確な情報の提供や Web の意味的な単位での取り扱いを可能にする点において有用である。

このような Web ページの分類作業は、従来、Web ディレクトリサービスなどにおいて人手によって行われてきた。しかし、人手で扱えるデータ量には限界があり、Web のほんの一部しかカバーできていない。

そこで、この作業を自動化するために、Web コミュニティという「同一の話題に対して興味を持つ Web ページの集合」を抽出する手法が考えられている。本稿では、Web ページと Web コミュニティが扱う話題との関連性によって Web コミュニティのメンバーを取捨選択することで、従来手法と比較して精度、及び網羅性の高い Web コミュニティ抽出手法を提案し、これについて評価を行う。

2 関連研究

WWW 上に存在する Web コミュニティを可能な限り全て抽出することを目的とした手法には、Web 上から全ての完全二部グラフ構造を抽出し、これをコミュニティのコアと考える Trawling[1] や、密な二部グラフ構造 (DBG:Dense Bipartite Graph) を抽出することでコミュニティの抽出を行う手法 [2] が存在する。

また、既存の Web ディレクトリなどであらかじめ分類された Web ページ集合を利用し、それらと関連性のある Web ページを抽出する手法として、最大フローアルゴリズム [3] や、MultiCocitation など [4] の手法が提案されている。

これらの手法の中で、特に [1],[2] のような Web コミュニティの抽出数を目的とする手法においては、抽出される Web コミュニティの精度や網羅性はあまり考慮されていない。しかし、Web の利用者にとって、有用であり理解できる形のコミュニティを抽出するためには、コミュニティ抽出の精度や網羅性は重要な要素である。そこで、本稿では、[2] の手法を拡張することにより、抽出される Web コミュニティの精度と網羅性を高める手法を提案する。

A Web Community Extraction Scheme Considering Page-Community Relationship.

†Naoyuki Saida, ‡Akira Umezawa, ††Hayato Yamana
†Department of Information and Computer Science, Science and Engineering, Waseda University
‡Graduate School of Science and Engineering, Waseda University

2.1 密な二部グラフ抽出手法 [2]

[2] の手法 (以下 DBG) では、まず、シードページ s を初期値とするページ集合 F を用意する。その後、F からリンクされているページを参照しているページ群を、outlink,inlink をたどることで探し出し、F に追加する。このような探索を一定回数繰り返すことで得られたページ集合 F と、F からリンクされているページを用いて二部グラフを作成し、outlink 数,inlink 数が与えられた閾値 p, q に満たないページを二部グラフから除くことによって密な二部グラフ構造を抽出し、コミュニティとする。

3 Web ページとコミュニティとの関連性を用いたコミュニティ抽出手法

3.1 Web ページの持つコミュニティの中心からの距離量の定義

提案手法 (PDBG:Plus DBG) では、Web ページが持つリンク情報を、あるページが別のページに対して興味を持っていることを示す情報と考える。つまり、Web ページ p 中に出現するリンクの 5 割が、あるトピック T に関する Web ページ集合からリンクされているページを参照している場合、p はそのページ中でトピック T に関して 5 割程度扱っていると考える。

ここで、Web ページ p と Web ページ集合 F との関連性 $Sim(p, F)$ を、p が持つリンクの中で、Web ページ集合 F の持つリンクと同じページを参照しているリンクの割合と定義する。

この関連性を利用して、ページ集合 F の中心と Web ページ p の間の距離量 $Dis(p, F)$ を次式で定義する。

$$Dis(p, F) = (1 - Sim(p, F)) + Dis(q, F)$$

ここで、q は、p と最も多くのページを共に参照している F 中のページとする。また、シードページ s を F の中心とし、$Dis(s, F) = 0$ と置く。

3.2 Web コミュニティの抽出

3.1 で定義した距離量 $Dis(p, F)$ を利用して、Web コミュニティの抽出を以下の手順で行う。

1. シードページ s と Web ページ集合 F を定義し、s を F のメンバーとする。
2. F にページが追加されなくなるか、設定した繰り返し回数をオーバーするまで、以下の手順を繰り返す。
   (a) F がリンクしているページにリンクしている F 中に存在しないページを抽出する。
   (b) 抽出されたページの持つ F との距離量を計算し、距離量が閾値 d 以下のページを F に追加する。
3. F と F からリンクされているページによって二部グラフを作成する。
4. 二部グラフから inlink 数、outlink 数が与えられた閾値 p, q に満たないページを繰り返し削除する。
5. 得られた二部グラフのうち、被リンク側をコミュニティとして出力する。

4 評価

提案手法によって得られるコミュニティの精度と網羅性の評価を行う。

評価に用いるデータセットとして NTCIR-4[5] Web タスクのデータセット (2001 年に jp ドメインから収集され、11,038,720 ページを含む) に含まれるリンクデータを使用する。

inlink 数,outlink 数の閾値 p, q をそれぞれ 3 とし、距離量の閾値 $d = 1$ と置き、outlink を持つ全ての Web ページをシードページとして提案手法 (PDBG) を適用し、データセットより Web コミュニティを抽出した。また、[2] の手法 (DBG) によっても同様にコミュニティ抽出を行い、提案手法と比較した。ここで、[2] の手法における繰り返し回数は 1 とし、inlink 数,outlink 数の閾値はともに 3 とする。

4.1 比較方法

比較に際して、DBG、PDBG の両手法によってコミュニティの抽出が可能であり、コミュニティのサイズが増加している 20 のシードページから得られたコミュニティを、DBG,PDBG 双方によって抽出されたページ集合 (*core*) と PDBG のみで抽出されたページ集合 (*add*) とに分けることにする。

この *core* と *add* のそれぞれからランダムに 50 個の Web ページ (*core*, *add* のサイズが 50 以下の場合はすべての Web ページ) を選択し、コミュニティの持つトピックと一致しているかを、人手により、以下の基準を用いて評価した。

- トピックそのものを扱う Web サイト (同一の作者によって作られた Web ページ集合) のなかで、入り口となるようなページ (例：日本の大学というトピックで、大学のトップページ)
- ページ中にトピックについて扱ったコンテンツを含むページ、またはこれらのページを含む Web サイトの入り口となるページ (例：デジカメというトピックで、各社の製品比較を行っているページ)

ここでは、ページの持つ有用性などは考慮しないことにし、単純にトピックやそれに関連するキーワードが、ページ中の主要なコンテンツ内に出現するかを元に判断する。

4.2 評価結果

まず、DBG,PDBG における網羅性の比較を行う。データセット全体から抽出されたコミュニティの大きさの分布と平均精度を図 1 に示す。

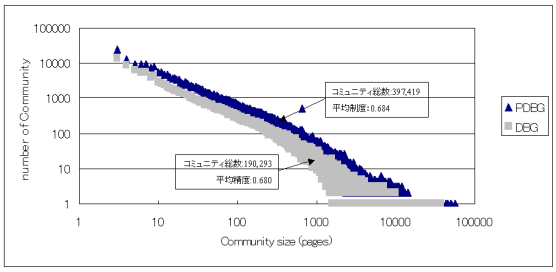

図 1　コミュニティサイズの分布と平均精度

図 1 より、抽出されたコミュニティのサイズ、総数共に PDBG が DBG を上回っていることがわかる。また、データセット全体で抽出されたコミュニティのサイズの合計は DBG で 55,372,789.191、PDBG で 154,089,281.775 となり、約 2.8 倍となっている。

次に、DBG,PDBG それぞれの精度の比較を行う。ここで、コミュニティ中でそのコミュニティの持つトピックと一致しているページを有効なページと定義する。DBG 中の有効なページ数は *core* の精度 × *core* のサイズ であり、PDBG の有効なページ数は (*core* の精度 × *core* のサイズ) + (*add* の精度 × *add* のサイズ) とする。また、PDBG の精度は (PDBG 中の有効なページ数)/(*core* のサイズ + *add* のサイズ) とする。

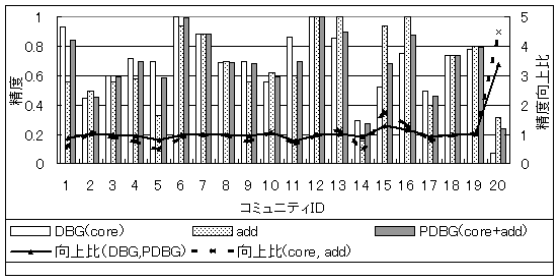

図 2　DBG と PDBG の精度比較

図 2 に、ランダムに選択した 20 個のコミュニティについて、*core*,*add* それぞれの精度を示す。DBG を基準としたときの PDBG の精度の向上は 0.8〜1.2 倍にとどまっており、精度の増減はコミュニティによって異なる。また、図 1 の平均精度を見ても、DBG,PDBG の両手法によって得られるコミュニティの精度には決定的な差はないと判断できる。

このように、提案手法によって従来手法 [2] よりも精度を落とすことなく、約 2.8 倍の Web コミュニティのメンバーを発見することが可能であることがわかった。

5 おわりに

本稿では、Web ページ－コミュニティ間の関連性によって定義された距離量を用いて、シードページと関連性のあるページのみを収集し、そこから密な二部グラフ構造を抽出することで、従来手法と比べて精度を落とすことなく、より網羅性のあるコミュニティを抽出することに成功した。

謝辞

本研究の一部は文部科学省「e-Society 基盤ソフトウェアの総合開発」によるものである。

参考文献

[1] R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Trawling the web for emerging cyber-communities", 8th International World Wide Web Conference, 1999.

[2] K. Reddy, M. Kitsuregawa, "An approach to relate the web communities through bipartite graphs", The 2nd International Conference on Web Information System Engineering, 2001.

[3] G. Flake, S. Lawrence, and C. Giles, "Efficient identification of Web communities", 6th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp.150-160, 2000.

[4] 原田 昌紀, 風間 一洋, 佐藤 進也, "参照共起分析の Web ディレクトリへの適用" IPSJ 研究報告「自然言語処理」No.142-007, 2001.

[5] NTCIR-4 Workshop, http://research.nii.ac.jp/ntcir/workshop/work-en.html