リンク構造を考慮したベクトル空間法によるWebグラフ分割の実験的解析

佐々木 雄一† 栗原 正仁‡
北海道大学 大学院情報科学研究科§

1 はじめに

近年，ブログやSNSなどの従来とは違った種類のWebページが現れたことにより，ますますWebページは多様化し，その数を増やしている．このような中，ユーザのWebページ閲覧における負荷の軽減や効率的な情報収集のための解決策として，多量のWebページを関連したページ毎にグループ化して情報を提供する方法が挙げられる．しかしながら，文書の内容やリンクの構造を用いたグループ構築に関する研究はそれぞれ別々に行われている．Webページからユーザにとって有益なグループを発見するためには，ページの文書内容とリンク構造の両方とも活用するべきである．そこで，本研究ではそれら両方を取り入れたグループ構築手法を提案する．文書とリンク両方の内容を反映させてグループを抽出することで，片方の内容を使うだけでは得られない，ユーザにとって有益な情報の提供を目指す．

2 関連研究

2.1 ベクトル空間法

Webページの集合から文書内容に基づきグループの構築を行うため，文書間の類似度を算出する手法として，一般的にベクトル空間法[1]が用いられている．ベクトル空間法では，Webページ文書d_iを表す特徴ベクトル$\overrightarrow{D_i}$を，

$$\overrightarrow{D_i} = [ws(d_i, w_1), ws(d_i, w_2), \ldots, ws(d_i, w_N)] \quad (1)$$

と定義している．ここでNは単語の総数，$ws(d_i, w_j)$は文書d_i中の単語w_jの出現頻度やtf-idfなどの特徴量である．2つの文書ベクトル$\overrightarrow{D_i}$と$\overrightarrow{D_j}$からコサインを求めることで，文書d_iと文書d_jとの類似度を算出する．類似度が1に近い値であるほど，d_iとd_jは類似した内容を持つ文書である．

Emperical analysis of web page grouping with vector space model for link structures
†Yuichi Sasaki
‡Masahito Kurihara
§Graduate School of Information Science and Technology, Hokkaido University

2.2 完全二部グラフ構造とコミュニティ

Kumarらは同じ興味を持ったWebページ集合fansは，興味の対象となる共通のWebページ集合centersへとリンクを張ることで，完全二部グラフ構造をもつコミュニティを形成するという考えに基づいて，Webの大規模データからサイズを固定した完全二部グラフ構造を高速に探索する手法を提案し，実験を行った．[2]実験により抽出されたコミュニティをランダムサンプリングして人手によって調査を行ったところ，それらのWebページ間には高い関連性があるという結果を得た．

3 提案手法

ベクトル空間法の特徴ベクトルを用いて類似性を算出するという考え方は，文書内容だけにしか適用できないわけではなく，リンク構造のように特徴量として考え難いものでも応用することが可能である．リンクから適切な類似度を与えるようなベクトルの特徴量として，基準となるWebページから周りのWebページへの最短パス長が挙げられる．Webページp_iのリンクベクトル$\overrightarrow{L_i}$を式2のように提案する．

$$\overrightarrow{L_i} = [spl(p_i, b_1), spl(p_i, b_2), \ldots, spl(p_i, b_k)] \quad (2)$$

kは基準となるページの総数，b_jは基準となるページ，$spl(p_i, b_j)$はページb_jからページp_iへと移動する際に辿る最短リンク数である．このモデルは，多くのWebページからほぼ同じパス長でたどり着くことができるWebページは類似するという考えに基づいている．さらに，リンクベクトルを用いて得られる類似度に着目すると，高い類似度を与えるリンク構造の1つにKumarらの提案する完全二部グラフが含まれる．最短パスの中にfans側のWebページが1つでも含まれる場合，centers側のWebページは同じ特徴値を持ち，高い類似度が与えられる．このように，リンクベクトルはcentersをグループとして構築させる性質をもち，さらに制約が強い完全二部グラフ構造以外のグループ構築にも対応した，より柔軟な性質をもつ．

上記のリンク構造を表す特徴ベクトル$\overrightarrow{L_i}$を，文書内

容を表す特徴ベクトル $\overrightarrow{D_i}$ に加えることで，両方を考慮に入れた Content-Link ベクトル $\overrightarrow{P_i}$ を作ることができる．$\overrightarrow{P_i}$ は式 3 のように定義される．

$$\overrightarrow{P_i} = \left[\alpha\overrightarrow{D_i}, (1-\alpha)\overrightarrow{L_i}\right] \quad (3)$$

Web ページ p_i,p_j 間の類似度はベクトル $\overrightarrow{P_i}$，$\overrightarrow{P_j}$ 間のコサインを求めることで計算できる．式 3 の α は文書とリンクのどちらの情報に重みをおくかを調整するパラメータである．

4 実験

実験には，アメリカの政治に関するブログのリンク構造のデータを使用する [3]．実験データの内，リンクがつながっている 1222 個の Web ページを用いて実験を行う．それらの内，586 個が民主党，636 個が共和党寄りの Web ページであり，91% のリンクは同じコミュニティへのリンクであることがわかっている．

実験では，式 3 の $\alpha = 0$ とし，リンクベクトルが Web ページに対してどのようなグループ構成を与えるかについて調べる．リンクベクトルの計算は基準となる Web ページを複数個選ぶことで行われるが，基準となるページの選び方によって結果が大きく異なるため，すべての Web ページを基準となるページとする．また，リンクは無向辺として扱う．

リンクベクトルから得られた Web ページ間の類似度に対し，完全連結法によるクラスタリングを適用してグループを構築する．

まずは，クラスタリングによって得られた結果として，民主党派と共和党派の 2 つのメイングループが 1 つになる直前のグループ構成表 (図 4) とその状態 (図 1) を示す．表 1 は，図 1 の各グループの民主党派と共和党派の数の割合を黒と白の棒グラフで示している．図 1 において，四角は民主党の Web ページ，丸は共和党の Web ページを意味し，同じ色で塗られた頂点が抽出された同一のグループを示している．

表 1: リンクベクトルの実験におけるグループ構成表

グループ名 (グループサイズ)	グループ構成 (数と党派の割合)		多数派を正解とした精度
	民主	共和	
Group1 (594)	547	47	92.1%
Group2 (628)	39	589	93.8%

クラスタリングでは，ハブとなっている中心ページの周りに存在するもの同士からグループが構築される傾向が見られた．表 1 に示した結果に限らず，クラスタリングの過程で得られたグループのほとんどは，民主党もしくは共和党のどちらかが大半を占めていた．

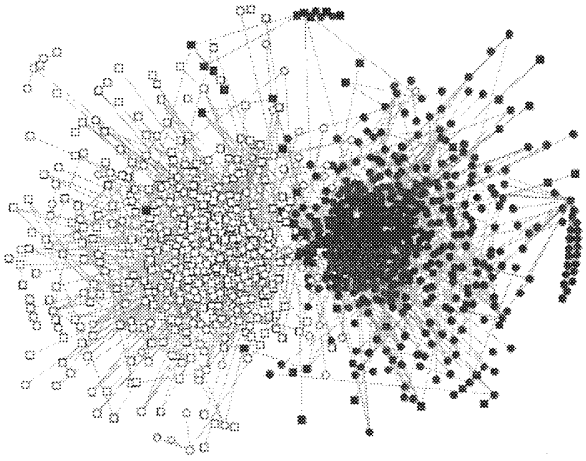

図 1: リンクベクトルの実験におけるグループ構成図

リンクベクトルから計算される類似度について調べると，その平均は 0.965，分散は 2.690×10^{-4} で，1 に近いほぼ同程度の値となっている．これは，ブログネットワークのスモールワールド特性により，他の Web ページへの最短パス長が 0 から 10 程度の間でしか変動しないという理由から生じた結果である．このことにより，文書内容との組み合わせを考えた際，リンク構造がグループ構築結果に反映されない結果につながる可能性がある．

5 おわりに

本研究では，ベクトル空間法のシンプルな考え方を利用することで，グループ構築にリンク構造を反映させたリンクベクトル提案した．さらに，文書とリンクのベクトルを組み合わせることで，両者の情報を利用した Content-Link ベクトルを提案した．ブログのリンク構造に対し，リンクベクトルを用いて実験を行い，結果として適切なグループを抽出することができた．今後の課題は，リンクベクトルのスモールワールド特性を考慮に入れた改善である．

参考文献

[1] Salton, G., "The Vector Space Model, Automatic Text Processing. " Addison Wesley Publishing, pp.312-325 (1985).

[2] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, "Trawling the Web for Emerging Cyber-Communities. " Proc. of WWW8, pp.403-415, (1999).

[3] L. A. Adamic, N. Glance, "The political blogosphere and the 2004 US Election.", Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005).