2D3-1

文書とリンク構造を考慮したベクトル空間法による Web グルーピング手法に関する研究

Grouping Web Pages using Vector Space Model for Document Contents and Link Structures

佐々木雄一*1 Yuichi Sasaki　栗原正仁*2 Kurihara Masahito

*1*2北海道大学 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

Several kinds of vector space model for analyzing document similarity for grouping web pages have been developed. However, they are not used for analyzing link structures, partly because they are complex and links do not necessarily satisfy the similarity relation. If we can device vector space models for link structures, we can combine them with those models document similarity in order to develop the unified basis for grouping web pages. In this paper, we present a vector space model for link structures by using the shortest path length between web pages. We also discuss the extension of this model to the model called content-link vector space model, which can treat document information and link information of web pages in a unified way. From the experiments with content-link vector space models, new groups that are not found with either link structures or documents are extracted.

1. はじめに

近年，ブログや SNS の開発が進んだことにより，Web ページはますます多様化し数を増やし続けている．それに伴い，現在ユーザの Web ページ閲覧における負担の軽減や効率的な情報収集を補助していく技術が強く求められている．ユーザの補助のための代表的なアプローチとして，大量の Web ページを関連したページ毎にグループ化するという方法が挙げられる．このアプローチは，Web ページのもつリンク構造や文書内容などのデータを利用して，類似性，関連性などの情報を取り出し，それらを元にグループを構築して行く手法である．

Web ページからグループを構築していく手法 [1, 2] は多岐に渡って研究されているが，中でも文書内容を用いた代表的な Web ページのグループ化手法であるベクトル空間法は，多くの Web ページのクラスタリング処理に使われている一方で，リンク構造を一切考えていない．我々はこの点に着目し，ベクトル空間法を用いて，Web ページがもつデータをより有効に使うことができる

本論文では，ベクトル空間法を用いて，文書内容と同じ様式で，リンク構造を表すベクトルを作り，それら 2 つを混ぜ合わせるシンプルでかつ効果的な手法を提案する．さらに，ブログのデータを用いた実験により，文書のみ，リンクのみ，文書とリンクの 2 つの情報を使ったグループ構築について考察する．

2. 提案手法

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 を表す文書ベクトル $\overrightarrow{D_i}$ を，

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

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．

2.2 リンクベクトル

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．無向辺で表されるリンク構造が与えられたとき，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．

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

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，基準となるページから共通の内容や興味を持つ Web ページまでは，少ないリンク数でたどり着くことができ，逆に基準となるページから内容が違うもしくは関連性がない Web ページまでは，多くのリンクを辿らなくてはたどり着くことができないという性質を利用している．このモデルを用いると，例えば複数のテニスに関する基準ページから見て，共通して少ないリンク数でたどりつける，近い位置にある Web ページはテニスの大会やテニスプレイヤー，スポーツなどの類の Web ページで，遠い位置にある Web ページは，映画や政治ニュースなどの関係のない Web ページだと判断してグループ化していくことができる．グループ化したい Web ページ集合に対し，満遍なく基準ページを設置することで適切なグループを得ることができる．

2.3 Content-link ベクトル

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

連絡先: 佐々木雄一，北海道大学情報科学研究科，
TEL 011-706-6815, FAX 011-706-7831,
E-mail:yusasaki@complex.eng.hokudai.ac.jp

ベクトル $\overrightarrow{P_i}$ を作ることができる．$\overrightarrow{P_i}$ は式 3 で定義される．

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

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．

3. 実験と考察

3.1 実験設定

実験には，2007 年に投稿されたライブドアブログの記事を使用する．実験データは，1566 個の記事と 8953 個のトラックバックのリンク構造から構成される．人手による調査から, 実験データは映画に関するトラックバックのリンク構造であり, ほとんどすべての記事が, 映画の作品をトピックとしていることがわかっている．

リンクベクトル $\overrightarrow{L_i}$ 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．次に，文書ベクトルについて説明する．形態素解析器 SEN を使い，記事内の名詞，未知語および，それらを組み合わせた単語を取り出す．次に，単語の長さが 3 以上で，かつ記事に出現した回数が 2 回以上 200 回未満ものを選出し，各単語 w_i のすべての記事における出現回数の合計 tf_i，w_i を含む記事の出現回数 df_i，記事数 $N = 1566$ としたとき，グループを構築するための単語の重要度 A_i を $A_i = tf_i \times log(N/df_i)$ の式により求める．求めた重要度 A_i が高い順に単語を N 個選び出し，文書ベクトル $\overrightarrow{D_i}$ を構築していく．また，文書ベクトル $\overrightarrow{D_i}$ の各要素には tf-idf の値を使用する．以上のリンクベクトル，文書ベクトル，$\alpha = 0.5$ として組み合わせた Content-link ベクトルの 3 つを用いて実験をして，グループ構築の考察をする．Content-link ベクトルから 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．クラスタリングによって得られたグループの評価は，同一のトピックを持つグループと，得られたクラスタリング結果のグループとの近さを表す Hubert statistic [3] を用いた．

3.2 結果と考察

文書ベクトル，リンクベクトル，$\alpha = 0.5$ とした Content-link ベクトルを用いた結果, いずれも高い評価値を与えるグループを構築した．代表的な結果として,$\alpha = 0.5$ とした Content-link ベクトルのグループ結果を示す (図 1)．図 1 において, 同じ色で塗られたノードは抽出された 10 以上のサイズのグループを意味し, 同じ形で表示されたノードはブログ記事の同一トピックを意味する．

リンクベクトルによって得られた結果では，他トピックへのリンクがなく，周りの Web ページにも他トピックへのリンクが存在せず，またリンク数が少ない傾向にある Web ページは，同一トピック同士の Web ページでも別々のグループを構築した．このような Web ページのベクトルの特徴量は，周りの Web ページのベクトルのすべての特徴量と 1，2 だけずれた値になる．同一トピックの Web ページ同士でも類似度が低くなり，グループが構築されないという結果につながっている．

文書ベクトルによって得られた結果では，同じブログ名をもつ記事同士のグループや，同一トピックの Web ページグループの中でもいくつかのグループが得られた．これらのグループは，ブログ内でよく使われる固有名詞が tf-idf により大きく特徴付けられていることや，グループ内でも文書の長さや単語の

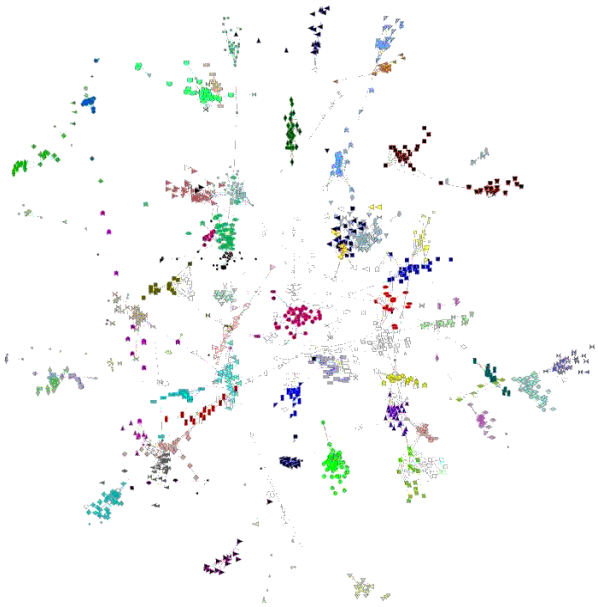

図 1: Content-link ベクトルの実験におけるグループ構成図

差異など，様々な文書内容があり，それを完全連結法でグループ化したことが原因となった結果である．

Content-link ベクトルの結果では, リンクベクトルと，文書ベクトルを用いた際に，両者の性質をカバーし合ったために, グループ化できていなかった部分のグループ構築をすることができた．しかし, 早い段階で違ったトピック同士の Web ページがグループ化したため，リンクベクトルのみ，文書ベクトルのみの結果よりも，高い評価値を与えるグループは構築できなかった．今後，この原因と α の値は適切であったかなどの考察を進めていく必要がある．

4. おわりに

本研究では，文書内容の類似性を算出する上で頻繁に利用されるベクトル空間法を，リンク構造にも同様に利用することで，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，Content-link 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，基準ページの選択や単語の選択，式 3 の α がグループ結果への影響について，詳しく実験をして考察を行うことである．

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] François Boutin, Mountaz Hascoët, “Cluster Validity Indices for Graph Partitioning. “, Proceedings of the Conference on Information Visualization IV’2004