2008－MPS－69 (3)
2008／5／16

リンク構造に基づいたWWWからのトピック抽出

山下 長義 †, 森山 甲一 ††, 沼尾 正行 ††, 栗原 聡 ††

† 大阪大学情報科学研究科情報数理学専攻 †† 大阪大学産業科学研究所

本論文では，Webページを分類するために，Webのリンク構造の類似性に着目する．たとえば，あるWebページと強い関連があるWebページが存在する場合には，それらを参照するページ群や，それらから参照されるページ群が似ていると考えられる．そこで，このようなことを判定するためにネットワーク分析の分野で使われている構造同値の概念を用いる．そして，クラスタ外のページとクラスタ内のページとの参照パターンを分析することで，構造同値に基づいて作成したデンドログラムにおけるクラスタの境界を個別に判定し，Webページを分類する手法を提案する．実験を行った結果，このような関係にあるクラスタを抽出することが有効であることが分かった．

Topic Detection from WWW Based on Link Structure

Nagayoshi YAMASHITA † Koichi MORIYAMA †† Masayuki NUMAO †† Satoshi KURIHARA ††

† Department of Information and Physical Sciences, Graduate School of Information Science and Technology, Osaka University

†† The Institute of Scientific and Industrial Research, Osaka University

In this paper, we focused on the similarity of link structure to classify Web pages. For example, pages with strong relation in content are often pointed from, and pointing to, the same pages. A concept of structural equivalence in network analysis is used to evaluate these structures. We propose a methodology to determine the boundary of each cluster in the dendrogram based on structural equivalence by analyzing the reference patterns on pages outside of the cluster. A preliminary experiment shows that extracting sets of clusters in this relationship is effective.

1 はじめに

近年，大量の情報から必要な情報を見つけることが困難になりつつある．そこで，検索結果の全体を把握することを容易にするために，検索の結果得られたWebページを言語処理によって分類する研究[1, 2]が行われている．一方で，リンク解析は計算量が少ないという特徴があり，Webページをハイパーリンクによるネットワークとみなすことで解析が行われている．

本論文では，Webページを分類するためにWebのリンク構造の類似性に着目する．たとえば，あるWebページと強い関連があるWebページが存在する場合には，それらを参照するページ群や，それらから参照されるページ群が似ていると考えられる．そこで，このようなことを判定するためにネットワーク分析の分野で使われている構造同値の概念を用いようと考えた．

そして，クラスタ外のページとクラスタ内のページとの参照パターンを分析することで，構造同値に基づいて作成したデンドログラムにおけるクラスタの境界を個別に判定し，Webページを分類する手法を提案する．たとえば，外部から同時に参照されていることが多い最大のクラスタや，外部に対して同時に参照していることが多いクラスタを抽出すれば，これらは互いに関連している可能性が高い．さらに，このような外部のページの中でも，特定のクラスタとの間にほとんどのリンクが存在するページは，ひとつのトピックのみにリンクを張っている可能性が高いのではないかと考えられる．そこで，このような外部ページとの間にリンクが多数存在するクラスタを複数抽出することでWebページの分類を行う．

初期実験を行った結果，提案手法の基本的な有効性が確認された．

以下，2節では関連手法について簡単に述べ，3節で構造同値について説明し，4節で提案手法を説明する．そして，5節で実験の手順を説明し，6節で評価を行い，7節にてまとめとこれからの課題を述べる．

2 関連研究

Web構造マイニングにおいては，リンク構造の参照関係から任意のWebページに対する類似サイトを発見する研究が行われている．たとえば，2つのページが共通の親を持つ，すなわち2つのページに対して同時にリンクを張っているページがある場合，これら2つのページは参照共起関係にあるページとして，これらページに対する親ページの数（参照共起度）が最も大きな値を持つ兄弟ノードを関連ページとする研究[3]や，このようなアルゴリズムを拡張し，複数のWebページに対する関連ページを発見する研究[4]もある．

本論文において導入する構造同値という概念はさまざまなネットワークに適用されている．たとえば，企業間関係の分析[5]や論文の参照関係から研究トピックを抽出する研究[6]に用いられている．言語処理による分類と比較して構造同値は，少数のクラスタへの分類には不向きであるが，より多数のクラスタへの分類では高い精度が得られることを，榊らは実証している[6]．

3 構造同値

本節では，構造同値の例と定義を述べ，次に構造同値の度合いを求めるための相関係数の計算方法について説明する．そして，相関係数を基にして逐次ノードを融合して階層構造を得る方法について説明し，最後にデンドログラムにおいてどのレベルのクラスタを出力するかを判断する方法について説明する．

3.1 構造同値の定義

何らかの組み合わせのグラフを考えたとき，ノードAとBがグラフ内の他のノードと完全に同じ関係を持つ場合，ノードAとBは構造同値であるという[7]．たとえば，図1において，ノード1とノード2，ノード3とノード4が構造同値の関係にあるとみなすことができる．構造同値の関係にあるノードは代替可能であるがゆえに，位置の独自性がなく競争関係になりやすいという特徴がある．

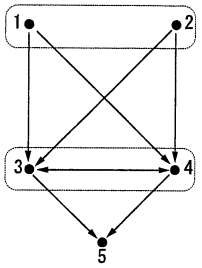

Fig. 1 構造同値の例

3.2 構造同値の度合いを求めるための相関係数

実際のグラフでは完全に構造同値であることはあまりないため，構造同値性を指標化し，連続量として捕らえるために隣接行列における行同士と列同士の相関が用いられる[8]．まず，グラフ上のノードの接続関係を隣接行列に変換する．隣接行列はグラフを表現するために用いる行列で，ある頂点vとwの間の辺の有無を行列の成分に割り当てる．辺があるとき(v,w)を1に，辺がないとき(v,w)を0にする．次に，隣接行列における相関を計算する．相関係数とは2つのデータ列の間の類似性の度合いを示す統計学的指標である．-1から1の間の実数値をとり，1に近いときは2つのデータ列には正の相関があるといい，-1に近ければ負の相関があるという．ノードiとノードj間の相関係数は式(1)のように定義することができる[9]．ただし，対角成分を除くi行の値の平均を$\overline{x}_{i+}$，同様にi列の値の平均を$\overline{x}_{+i}$とし，合計はkに対して行い，$i \neq k$，$j \neq k$である．

$$r_{ij} = \frac{A+B}{C \cdot D} \tag{1}$$

$$A = \sum (x_{ki} - \overline{x}_{+i})(x_{kj} - \overline{x}_{+j})$$

$$B = \sum (x_{ik} - \overline{x}_{i+})(x_{jk} - \overline{x}_{j+})$$

$$C = \sqrt{\sum (x_{ki} - \overline{x}_{+i})^2 + \sum (x_{ik} - \overline{x}_{i+})^2}$$

$$D = \sqrt{\sum (x_{kj} - \overline{x}_{+j})^2 + \sum (x_{jk} - \overline{x}_{j+})^2}$$

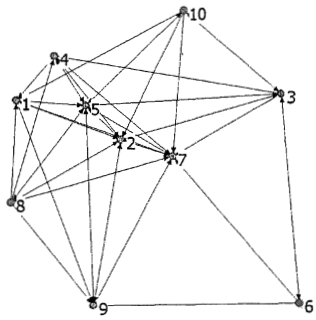

Fig. 2　グラフの例

	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.21	1.00								
3	0.12	0.06	1.00							
4	0.48	0.33	0.10	1.00						
5	0.51	0.63	0.09	0.51	1.00					
6	0.05	0.31	0.03	0.05	0.27	1.00				
7	0.30	0.52	0.31	0.53	0.43	- 0.07	1.00			
8	0.36	0.20	0.37	0.23	0.09	0.00	0.04	1.00		
9	0.47	0.21	0.50	0.35	0.24	- 0.05	0.42	0.47	1.00	
10	0.21	0.22	0.30	0.29	0.28	0.39	0.07	0.33	0.29	1.00

Fig. 3　図 2 のグラフの隣接行列における相関

例として図 2 のような 10 個のノードからなるグラフの隣接関係の相関を求めると，図 3 の行列が得られる．この行列において (i, j) は，ノード i とノード j の相関係数を表している．

次に，相関係数を基にデンドログラムを作成する．まず，1 つのページだけを含むクラスタがある初期状態を作る．そして，もっとも相関係数の大きな値を持つクラスタを逐次融合し，すべてのノードが 1 つのクラスタに融合されるまで融合を繰り返すことで階層構造を得る．この階層構造は，デンドログラムや集合に基づいて表示することができる．

3.3　どのレベルのクラスタを出力するかの判断方法

デンドログラムにおいてどのレベルのクラスタを出力するかの判断方法は，階層的クラスタリングの分野で以下のような方法がある．

クラスタの数をあらかじめ指定する方法では，クラスタの融合を逐次繰り返し，指定したクラスタ数に到達したときに融合を終了し，それまでに形成されたクラスタを出力する．

直接出力するレベルを指定する方法では，出力する相関のレベルを決め，その相関の値以上を持つページ同士を融合しクラスタを形成させる．

4　Web ページの分類

Web ページのハイパーリンクによる隣接関係の相関を基にしたデンドログラムを作成し，デンドログラムにおいてどのレベルのクラスタを出力するかを決定する．

従来の階層的クラスタリングの分野における方法では，クラスタの数をあらかじめ指定する方法や直接出力するレベルを指定する方法がある．しかし，適当なクラスタ数はあらかじめ分からない場合がほとんどである．また，Web ページのネットワークはスパースなので，優位なレベルを判定することが難しいことが多い．その上，Web ページによるネットワークの次数分布がスケールフリーであることから，相関係数を基に作られたデンドログラムにおいて，直接レベルを指定してクラスタに分類すると，出力するレベルによっては次数の低いページが多数含まれるクラスタと，次数の大きいページがひとつだけ含まれるクラスタに分類されることが予想される．次数の小さなページは，参照パターンが限られるため他のページとの相関が大きくなりやすいからである．

またこれら 2 つの方法は，ネットワーク上のすべてのノードを分類する手法であるが，Web ページのネットワークでは，同一サイトのページなどの特定のページとのつながりが強く次数が少ないページなど，どうしても適切に分類できない部分が存在し，すべてを分類しようとすると精度の低下を招く．

そこで，クラスタの外のページとクラスタ内のページとの参照パターンを解析することで，デンドログラムにおけるクラスタの境界を個別に判定し分類する手法を提案する．外部から同時に参照されていることが多い最大のクラスタと，外部に対して同時に参照していることが多い最大のクラスタは互いに関連している可能性が高い．さらに，このような外部のページの中でも，特定のクラスタとの間にほとんどのリンクが存在するページは，ひとつのトピックのみにリンクを張っている可能性が高いのではないかと考えられる．そこで，このような外部ページとの間にリンクが多数存在する極大のクラスタをデンドログラムにおけるクラスタの境界とし，このようなクラスタを複数抽出

することでWebページの分類を行う．クラスタごとにクラスタ外のページとの関係を調査するため，出力されるレベルはクラスタごとに異なり，出力されるクラスタ数はあらかじめ指定する必要はない．

手順は以下の通りである．

1. Webページを収集する．
2. 得られたリンク構造に対して，ノード対ごとに隣接行列における相関を求め，デンドログラムを作成する（3.2節）．
3. （2）で作成したデンドログラムにおいて相関係数が0から1の間で融合されるクラスタごとに定義1から定義3に基づいて極大関連クラスタかどうかを判定する．
4. （3）で求めた複数の極大関連クラスタを分類結果として出力する．

定義 1 任意のページ i とクラスタ C_k との間にリンクが存在する割合 $\Delta_{i\to C_k}$ と $\Delta_{C_k\to i}$ を以下のように定義する．

$$\Delta_{i\to C_k} = \frac{\sum_{j\in C_k} X_{ij}}{|C_k|} \tag{2}$$

$$\Delta_{C_k\to i} = \frac{\sum_{j\in C_k} X_{ji}}{|C_k|} \tag{3}$$

ただし，

$$X_{ij} = \begin{cases} 1 & \text{ページ } i \text{ からページ } j \text{ へリンクがあるとき} \\ 0 & \text{ページ } i \text{ からページ } j \text{ へリンクがないとき} \end{cases} \tag{4}$$

とする．

定義 2 任意のページ i のクラスタ C_k に対する集中度 $S_i(C_k)$ を以下のように定義した．ページ i の次数を H_i，次数 H_i のうち *1* つのクラスタ C_i 間にあるリンク数を L_i とする．

$$S_i(C_i) = \frac{L_i}{H_i} \tag{5}$$

定義 3 任意のページ i とクラスタ C_l があるとする．

$\Delta_{i\to C_l} \geq \alpha$ または，$\Delta_{C_l\to i} \geq \alpha$

かつ

$S_i(C_i) \geq \beta$

のとき，ページ i をクラスタ C_l のハブと呼ぶ．そして，ハブとなるページが存在するクラスタ C_l を関連クラスタと呼ぶ．

定義 4 C が関連クラスタ　かつ　$C \subsetneq D$ となる任意のクラスタ D が関連クラスタでないとき，C を極大関連クラスタと呼び，デンドログラムにおけるクラスタの境界とする．

5　動作実験

リンク解析を行うWebページを以下のようにして収集する．検索エンジンにあるキーワードを入力し，結果上位200までのWebページのURLを収集する[1]．そして，これらのページからリンクが張られているすべてのページと，これらのページに対してリンクを張っている10ページを収集することとした[2]．ただし，異なるドメイン間のリンクのみを用いる．また，広告のページやポータルサイトなどをストップページリストに加え，このリスト上に存在するページは解析対象から除外した．データに関する詳細は以下の通りである．

- 検索語　S社 (電機メーカー)
- ページ数　583

このようにして収集したページ間のリンク構造に対して相関係数を計算した．相関係数が互いに1であるページ同士を1つのクラスタに分類した場合，336個のクラスタに分類され，1つのクラスタに分類される平均の文書数は1.72ページであった．全体の約半数の262ページは，1つのクラスタに対して1つのページが分類された．

そして，提案手法によりハブを同定し極大関連クラスタを出力した．ただし，クラスタに対してハブとなるために必要なリンクが存在する割合 α を0.5，集中度の閾値 β を0として実験を行った

[1] HITSアルゴリズムの論文[10]において，検索結果の上位200ページから収集を始めているため，それに倣った．

[2] Google WEB APIsによりそれぞれのURLに対してリンクを張っているページを収集した．しかし，1回10件，1日1000回までとう制限があるため，200ページそれぞれに対してリンクを張っているを10ページ収集することに止めた．

6 評価

それぞれの極大関連クラスタ内のページが互いに関連しているかを評価するために，被験者10人に極大関連クラスタ内のそれぞれのページを見てもらい，それぞれのページを3つ以内の言葉で表現してもらった．その結果からそれぞれの極大関連クラスタ内のページが互いに関連しているかどうかを判断した．クラスタ内のページが共通の概念を持つ1つ以上の言葉で被験者によって表現されていれば，その極大関連クラスタは内容が関連しているとみなした．

完全に構造同値（相関係数が1）であるページで構成され，かつ2ページ以上を含むクラスタの精度は73%であった．

6.1 集中度に対する精度

実験で定義3におけるリンクが存在する割合の閾値 α を0.5，集中度の閾値 β を0としたが，実際にはリンクが存在する割合が1であるハブページを有する極大関連クラスタは全体の94%であった．一方，集中度は幅広い値をとった．そこで，α を0.5に固定して，β を変えたときの精度と分類されるページ数の変化を検証した．集中度は0から1までの値をとり，集中度が1のときはハブのすべてのリンクが特定のクラスタとの間に存在している．

図4は，横軸に集中度を，縦軸に精度と分類されるページ数をとったものである．横軸において0.5であるとは，β が0.5のときの極大関連クラスタの精度と分類されるページ数を表している．高

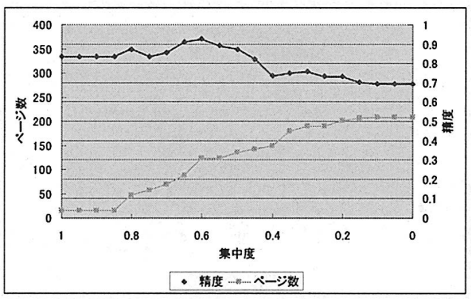

Fig. 4 集中度に対する精度と分類されるページ数

精度で多くのページが分類されることが望ましい結果である．

集中度が1から0.5の間において精度は0.85を下回らず維持しているため，β が0.5のときは，高い精度を維持しつつ，分類されるページ数が多く，β が0.5のときがふさわしい状態の1つであると考えられる．β が0.5のときの極大関連クラスタは31個形成され，そのうち内容が関連があるクラスタは27個あり，極大関連クラスタのうち内容が関連していると評価されたものの割合，つまり精度は87%であった．

以下では，β が0のときの極大関連クラスタを出力した場合，つまり集中度を全く考慮しない場合と比較することで，β が0.5のときの極大関連クラスタを出力した場合の有効性の検証を行った．

6.2 相関係数の値に対する精度

相関係数の値と内容の関連性について検証した．図5は，横軸に相関係数の値を，縦軸に精度と分類されるページ数をとったものである．

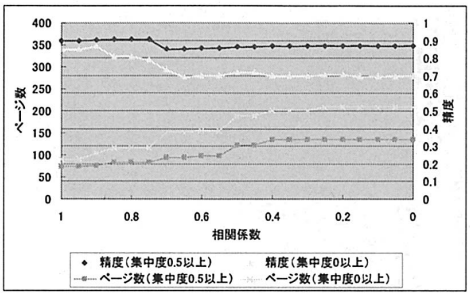

Fig. 5 相関係数に対する精度と分類されるページ数

たとえば，横軸において，0.4とは，0.4以上の相関係数の値をもつ極大関連クラスタの精度と分類されるページ数を示している．β が0のときの極大関連クラスタの精度は，相関係数が小さくなるにつれて減少した．一方，β が0.5のときの極大関連クラスタの精度は，相関係数が0.75から0.7の間でわずかに減少するが，それ以外ではほぼ一定であった．β が0.5のときの極大関連クラスタは，β が0のときの極大関連クラスタと比較して，相関係数の値にかかわらず，分類されるページは減少したが精度は高かった，

6.3 ハブの本数に対する精度

極大関連クラスタに対するハブの本数に対する精度と分類されるページ数の関係を検証した（図6）．β が 0 のとき，ハブの本数に対する極大関連

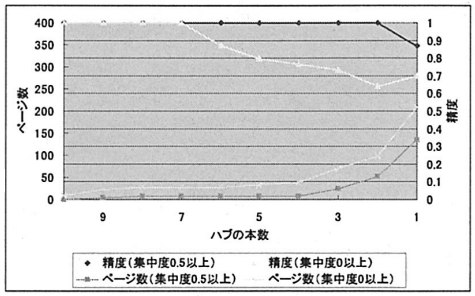

Fig. 6　ハブの本数に対する精度と分類されるページ数

クラスタの精度は比例した．また β が 0.5 のとき，β が 0 のときより高い精度で分類ができた．いずれの場合もハブの本数に精度は比例したが，集中度が 0.5 のときの極大関連クラスタの精度の方が常に高い精度を維持した．

7　まとめとこれからの課題

隣接関係における相関が高いページで構成されるクラスタとの間にほとんどのリンクが存在するページを特定し，このようなページが存在する隣接関係の相関が高いクラスタを複数抽出しすることで Web ページの分類を行った．そこで，特定のページだけと強く結びついているページや，さまざまなトピックに対してリンクを張っているページの影響を除外することができ，高い精度で分類を行うことができた．

今後の課題は，関連したページで構成されたクラスタであったにもかかわらず，極大関連クラスタ集合に含まれなかったクラスタの特徴を調べ原因を探求することである．また，解析対象のページ集合の中に検索語と関連のないページが含まれることを最小限にするなど，提案手法に適したデータ収集方法とはどのような方法であるかを検討する必要がある．

参考文献

1) H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma and J. Ma: “Learning to cluster web search results”, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2004).

2) O. Zamir and O. Etzioni: “Web document clustering: A feasibility demonstration”, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998).

3) J. Dean and M. R.Henzinger: “Finding related pages in the world wide web”, Computer Networks(Amsterdam,Netherlands) (1994).

4) 原田, 風間, 佐藤：“参照共起分析の web ディレクトリへの適用”, 情報処理学会研究会報告, 2001-FI-61-7, pp. 45–52 (2001).

5) 渡邊, 小坂：“日本における企業間関係の社会ネットワーク分析”, 経営情報学会春季全国研究発表大会, pp. 356–359 (2005).

6) 榊, 松尾, 市瀬, 武田, 石塚：“論文データベースからの研究トピック抽出”, 人工知能学会第19回全国大会 (2005).

7) 安田：“実践ネットワーク分析”, 新曜社 (2001).

8) 安田：“ネットワーク分析”, 新曜社 (1997).

9) S. Wasserman and K. Faus: “Social Network Analysis”, Cambridge University Press (1999).

10) J. M. Kleinberg: “Authoritative sources in a hyperlinked environment”, Journal of the ACM (1999).