情報処理学会論文誌　数理モデル化と応用　Vol. 2　No. 1　133–144 (Feb. 2009)

リンク構造に基づいたWWWからのトピック抽出

山下長義[†1]　森山甲一[†2]
沼尾正行[†2]　栗原聡[†2,†3]

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,[†1] Koichi Moriyama,[†2]
Masayuki Numao[†2] and Satoshi Kurihara[†2,†3]

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 大阪大学情報科学研究科情報数理学専攻
Department of Information and Physical Sciences, Graduate School of Information Science and Technology, Osaka University
†2 大阪大学産業科学研究所
The Institute of Scientific and Industrial Research, Osaka University
†3 JST CREST

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，Web ページをハイパーリンクによるネットワークと見なすことで解析が行われている[3),4)]．Web ページ上のハイパーリンクは，Web 文書内に埋め込まれた他の文書などの位置情報であり，さらにその参照先のページ内容に対して何らかの関連があり，そのページの情報に対して価値を認めているという特長もある．

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，1 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 章では関連手法について簡単に述べ，3 章で構造同値について説明し，4 章で提案手法を説明する．そして，5 章で実験の手順を説明し，6，7 章で評価し，8 章で考察を行い，9 章でまとめとこれからの課題を述べる．

2. 関連研究

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．

デンドログラムにおいてどのレベルまでに形成されたクラスタを出力するかを決定する従来の方法は，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，適当なクラスタ数や出力するレベル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に対して E-I index[7]（3.3.2 項）を用いても有意なレベルを判定することが難しいことが多い．

Web 構造マイニングにおいては，リンク構造の共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[3] や，この研究を拡張した複数の Web ページに対する関連ページを発見する研究[4] も行われている．また，サイト間の参照頻度の行列において高い相関を持つサイトどうしは類似していると見なし，得られた相関行列を可視化することで，リンク構造による分類の有効性を検証している研究[8] がある．

Web 上の検索エンジンによって得られた結果を言語処理や統計処理によって分類する研究[1),2)] や，Clusty [⋆1] や Grokker [⋆2] などサービスとして実用化されているものもある．

3. 構 造 同 値

本章では，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デンドログラム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までに形成された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

何らかの組合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[9]．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見なすことができる．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．

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どうしと列どうしの相関が用いられる[10]．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[11]．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．

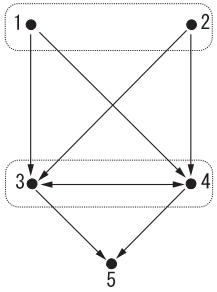

図 1　構造同値の例
Fig. 1　Example of structural equivalence.

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

$$A = \sum (x_{ki} - \overline{x}_{+i})(x_{kj} - \overline{x}_{+j})$$
$$B = \sum (x_{ik} - \overline{x}_{i+})(x_{jk} - \overline{x}_{j+})$$
$$C = \sqrt{\sum (x_{ki} - \overline{x}_{+i})^2 + \sum (x_{ik} - \overline{x}_{i+})^2}$$
$$D = \sqrt{\sum (x_{kj} - \overline{x}_{+j})^2 + \sum (x_{jk} - \overline{x}_{j+})^2}$$

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．

⋆1 http://clusty.jp/
⋆2 http://www.grokker.com/

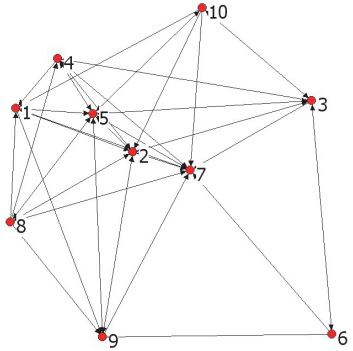

図 2　グラフの例
Fig. 2　Example graph.

	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

図 3　図 2 のグラフの隣接行列における相関
Fig. 3　Correlation for graph shown Fig. 2.

次に，相関係数を基に完全連結法[⋆1]によって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，最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**図 4** のように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判断する方法には以下のようなものがある．

⋆1 最長距離法とも呼ばれ，クラスタ間の類似度（距離）をそれぞれのクラスタ間の類似度（距離）のうち最小（最大）のものとして決定する方法

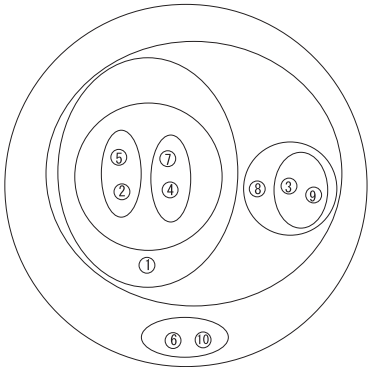

図 4　集合に基づいた表現方法
Fig. 4　Example of hierarchical clustering.

3.3.1　クラスタの数をあらかじめ指定する方法

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．

3.3.2　直接出力するレベルを指定する方法

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どうしを融合しクラスタを形成させる．

相関係数によるデンドログラムでは，そのレベルを決定するために E-I（External-Internal）index[7] という手法が用いられる．E-I index は式 (2) で計算できる．ただし，任意のレベルで形成されるクラスタにおいて，クラスタの外のノードに対するリンクの数を L_{out}，クラスタの内のノードに対するリンクの数を L_{in}，同一ネットワーク内のすべてのリンク数を L_{all} とする．-1 から $+1$ の間の値をとり，すべてのリンクがクラスタの外のノードへのリンクであるときは $+1$ となり，すべてのリンクがクラスタ内へのリンクであるときは -1 となる．

$$E - Iindex = \frac{L_{out} - L_{in}}{L_{all}} \tag{2}$$

E-I index が大きな値を保持しつつ，かつ少ない数のクラスタに融合されたレベルを出力することが望ましい．なぜなら相関係数では，クラスタの内側で高い類似度を示す一方の外のノードとは明確に区別されるクラスタを抽出することが目的であるからである．

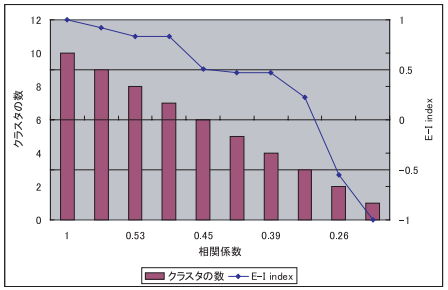

図 5　E-I index の例
Fig. 5　Example of E-I index.

たとえば，3.2 節で扱ったデータに対して E-I index の値を求める（**図 5**）．

相関係数が 0.39 のときを出力するレベルとすると，E-I index の値が 0.47 で，4 つのクラスタに分類される．相関係数が 0.39 より小さくなると E-I index の値は減少するため，相関係数を 0.39 のときを出力レベルとし，4 つのクラスタに分類することが有力な候補となる．

4. Web ページの分類方法

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に形成されるクラスタを出力するかを決定する．

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，1 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，このような外部の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．

手順は以下のとおりである．

(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．

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

$$\Delta_{C_k\to i} = \frac{\sum_{j\in C_k} X_{ji}}{|C_k|} \tag{4}$$

ただし，

$$X_{ij} = \begin{cases} 1 & \text{ページ } i \text{ からページ } j \text{ へリンクがあるとき} \\ 0 & \text{ページ } i \text{ からページ } j \text{ へリンクがないとき} \end{cases} \tag{5}$$

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_k との間にあるリンク数を L_i とする．

$$S_i(C_k) = \frac{L_i}{H_i} \tag{6}$$

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

$S_i(C_l) \geq \beta$

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．

たとえば，**図 6** においてクラスタ 1（ページ 2，4，5，7）に対してページ 10 のようにクラスタ 1 に対して多数のリンクを張っていて，ページ 10 のクラスタ 1 に対するリンクがページ 10 のすべてのリンクであるとする．この場合定義からページ 10 をハブとし，クラ

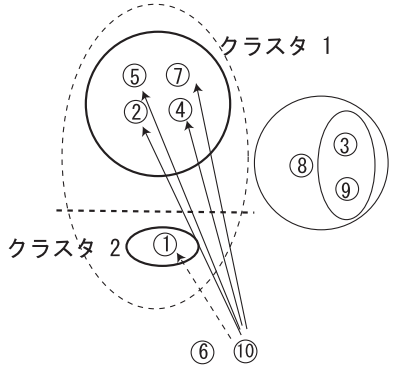

図 6　ハブの例（この例ではページ 10 がハブとなる）
Fig. 6　Example of hub page (In this figure, page 10 is a hub).

スタ 1 を関連クラスタとする．そして，クラスタ 1 を包含する関連クラスタが存在しない場合，クラスタ 1 を極大関連クラスタとし，クラスタ 1 内に属するページは互いに関連しているものとする．

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分類対象から除外した．データに関する詳細は以下のとおりである．

- 検索語　S 社（電機メーカ）
- 収集した全ページ数　1,969
- 分類対象となったページ数　583

- 検索語　ジョギング
- 収集した全ページ数　2,368
- 分類対象となったページ数　524

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．

検索語が S 社の場合，相関係数が互いに 1 であるページどうしを 1 つのクラスタに分類すると，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．また，検索語がジョギングの場合，相関係数が互いに 1 であるページどうしを 1 つのクラスタに分類すると，358 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 1.46 ページであった．302 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．

6. 構造同値の関係にあるページの評価

構造同値の関係にあるページを 2 つ取り上げて，Web ページとの内容の関係を検証した．

はじめに充電池関連の商品に関する 2 つの記事について検証を行った．この 2 つのページは，S 社に関するはてなブックマークのページと，充電池に関する他の記事からリンクが張られていて，S 社のメインページ，S 社サイト内の充電池に関するページと S 社の充電池に関する記事にリンクを張っていた．

次に，S 社の当時の会長についての 2 つのページについて検証した（**図 7**）．これら 2 つのページは，同一の趣旨によって文書が書かれているページであり，これらのページに対して，S 社の当時の会長について書かれていた 2 つのブログから同様にリンクが張られていた．このような構造同値の関係にあるページが同じトピックに関して書かれているページ集合は，Web ページによるネットワークにおいても多数存在した．たとえば，検索語が S 社の場合では，携帯電話，デジカメやカーナビなどのページ集合があり，検索語がジョギング

⋆1 HITS アルゴリズムの論文 12) 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，それにならった．
⋆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 1,000 回までという制限があるため，200 ページそれぞれに対してリンクを張っている 10 ページを収集することにとどめた．

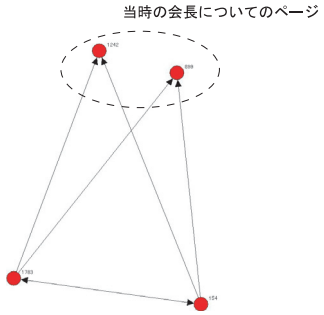

図 7　当時の会長についてページの例
Fig. 7　Pages on the former president.

の場合では，マラソン，ダイエット，筋トレや Wii Fit によるジョギングなどのページ集合が存在した．

このように，構造同値の関係にある Web ページは，内容において関連がある場合が多いことが分かった．

7. E-I index による評価

E-I index によって，どのレベルに形成されるクラスタを出力するかを判断しようと試みた．しかし，このデータにおける E-I index の値は直線に減少したため，どのレベルに形成されるクラスタが有意であるかを判断することは困難であった（**図 8**）．

8. 提案手法の評価

提案手法によるデンドログラムにおける境界と相関係数の関係を示し，次に被験者による評価の結果を示し，集中度や相関係数やハブの本数などのパラメータに対する精度の変化の検証を行った．

8.1 デンドログラムにおける境界と相関係数

図 9 は，提案手法によるデンドログラムにおける境界と相関係数の関係を示している．提案手法が一律に閾値を定める手法でないことが分かる．

8.2 被験者による評価の詳細

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

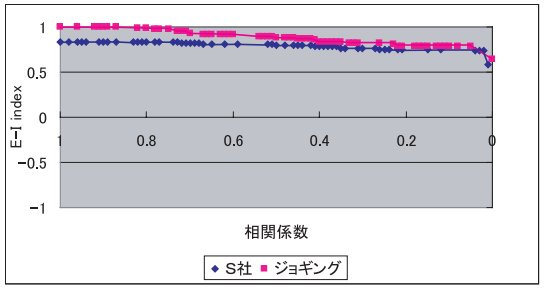

図 8　E-I index の値と相関係数
Fig. 8　E-I index and correlation coefficient.

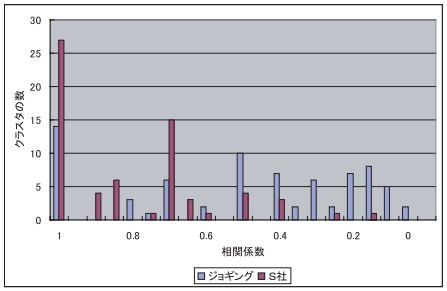

図 9　提案手法によるデンドログラムにおける境界と相関係数
Fig. 9　Border of the dendrogram and correlation.

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見なした．たとえば，検索語が S 社の場合は以下のような言葉は共通の概念を持つ言葉であると見なした．

- 「商品名」と「製品名」
- 「ナビ」と「カーナビ」
- 「充電池」と「バッテリー」
- 「会長辞任」と「トップ辞任」
- 「ビデオカメラ」と「デジタルムービーカメラ」と「digital video camera」
- 「企業情報」と「株式情報」と「経常利益」

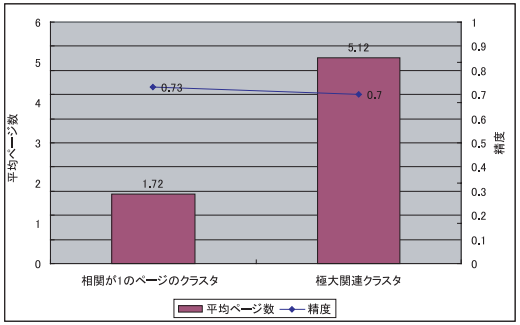

図 10　相関が 1 のページのクラスタと極大関連クラスタの比較（S 社）
Fig. 10　clusters with a correlation of 1 and maximal related clusters (S company).

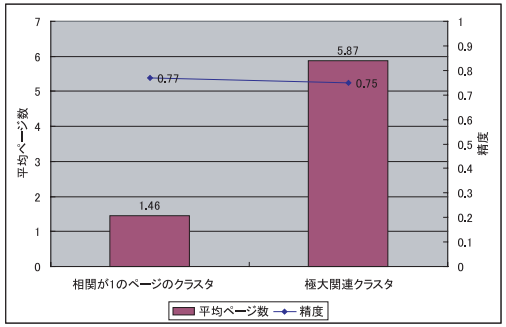

図 11　相関が 1 のページのクラスタと極大関連クラスタの比較（ジョギング）
Fig. 11　clusters with a correlation of 1 and maximal related clusters (jogging).

- 「半導体」と「電子デバイス」
- 「太陽電池」と「電池」

8.3　相関が 1 のページのクラスタと極大関連クラスタとの比較

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と極大関連クラスタの比較を行った（**図 10**，**図 11**）．

その結果両方の検索語において，完全に構造同値（相関係数が 1）であるページで構成されるクラスタに属する平均ページ数に対する極大関連クラスタに属する平均ページ数は 3 倍以上に増加したにもかかわらず，精度はほぼ変わらずに分類できた．

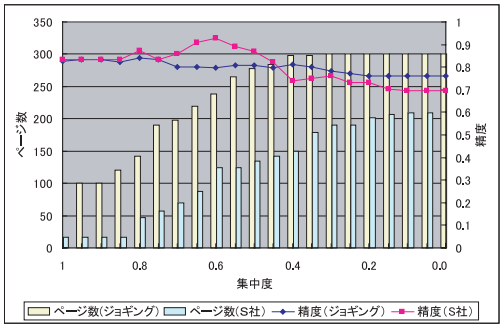

図 12　集中度に対する精度と分類されるページ数
Fig. 12　Precision and number of pages within each cluster with degree of concentricity.

8.4　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 12 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．

β が 0.5 の極大関連クラスタは高い精度を維持しながら分類されるページ数が多く，ふさわしい状態の 1 つであると考えられる．

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，つまり集中度をまったく考慮しない場合と，β が 0.5 のときの極大関連クラスタを出力した場合とを比較することで集中度の有効性の検証を行った．

8.5　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 13** と**図 14** 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．図 13 は検索語として S 社を用いた場合，図 14 は検索語としてジョギングを用いた場合である．

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持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 0.5 のときは，β が 0 のときと比較して相関係数

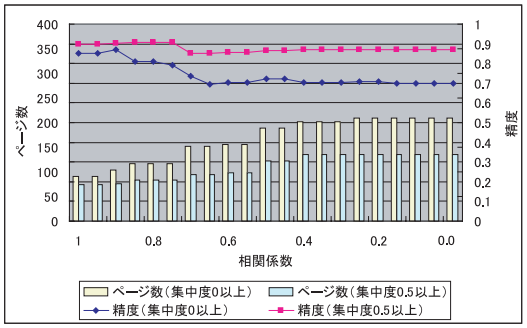

図 13　相関係数に対する精度と分類されるページ数（S 社）
Fig. 13　Precision and number of pages within each cluster with correlation coefficient (S company).

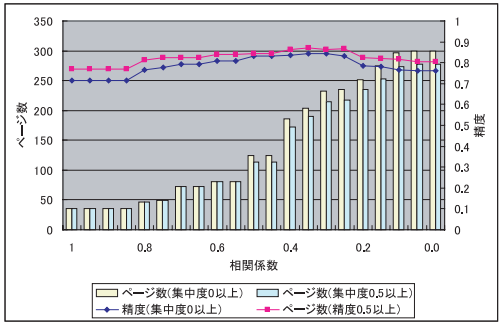

図 14　相関係数に対する精度と分類されるページ数（ジョギング）
Fig. 14　Precision and number of pages within each cluster with correlation coefficient (jogging).

の値にかかわらず精度は高かった．このように，単に同時に参照されているクラスタよりも，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がハブである極大関連クラスタの方が精度が高いことが分かった．

8.6　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 15**，**図 16**）．

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つねに高い精度を維持した．

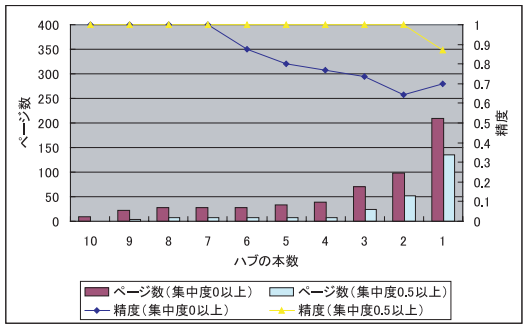

図 15　ハブの本数に対する精度と分類されるページ数（S 社）
Fig. 15　Precision and number of pages within each cluster with number of hub (S company).

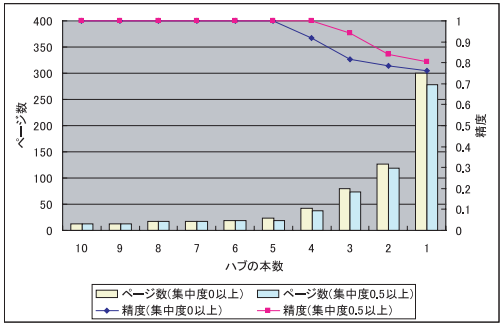

図 16　ハブの本数に対する精度と分類されるページ数（ジョギング）
Fig. 16　Precision and number of pages within each cluster with number of hub (jogging).

8.7　隣接行列における被参照関係のみの相関を用いた分類

最後に，隣接行列における被参照関係のみの相関を用いて分類を行った．いい換えるとどのページに対してリンクを張っているかは考慮せず，どのページからリンクを張られているかということのみを考慮して分類した．しかし，被参照関係のみの相関で分類することの有意性を見い出せなかった．Google WEB APIs における 1 日に使用することができるクエリの回数の制限によって，最初に収集した 200 ページに対してリンクを張っているページをそれぞれ 10 ページずつしか収集できなかったことが原因の 1 つであると考えられる．

9. 考　　察

検索語が S 社の場合おける極大関連クラスタ間の関係を表示し，次に提案手法のスケーラビリティに関して検討を行い，最後に Cocitation[3] という手法との比較を行った．

9.1 極大関連クラスタ間の関係

主な極大関連クラスタとそのクラスタに対するハブの関係を**図 17** に示す．ここでのノードは極大関連クラスタまたはハブページとする．ハブが他の極大関連クラスタ内のページに属していれば，極大関連クラスタ間を矢印で結んだ．図 17 のように図示することで，携帯電話関連のページ，医療関係のページ，カーナビのページ，マーケット関連のページや前会長のページなど互いに内容が関連している極大関連クラスタ間でハブが存在した．このように表示することによって，分類された検索結果全体を 1 つの画面で見渡すことが可能となり，将来情報検索に応用できると考えている．

9.2 スケーラビリティに関する検討

提案手法での主な処理は以下のものがあげられる．

(1)　Web ページを n 個収集する．
(2)　Web ページの隣接関係を表す隣接行列の相関を計算する．
(3)　これらの相関によるデンドログラムにおいて極大関連クラスタを同定する．

Web ページを収集するための処理数は対象とするページ数の増加とともに線形的に増加するため，処理 (1) における計算量は $O(n)$ となる．しかし，Google や Yahoo のような検索サービスでは，事前にデータを収集し解析することで検索の高速化を図っている．本手法においても，同様に事前にデータを収集することで処理を高速に行うことが可能であると考えられる．

次に，処理 (2) における隣接行列の相関を計算では，ページの組合せの数だけ相関を計算するため，計算量は $O(n^2)$ となる．ただし，個々の相関の計算は，他の相関の計算に影響を与えないため並列処理が可能である．

また，処理 (3) における極大関連クラスタを同定する処理では，対象となるページ数にかかわらず相関の値が 0 から 1 で融合されるクラスタごとに極大関連クラスタかどうかを判定するため，対象となるページ数が増加しても計算量はほとんど変わらない．

このように，ページの増加とともに処理 (2) における計算量が最も増大するが，計算量はたかだか $O(n^2)$ であるため十分にスケーラビリティのあるシステムであると考えられる．

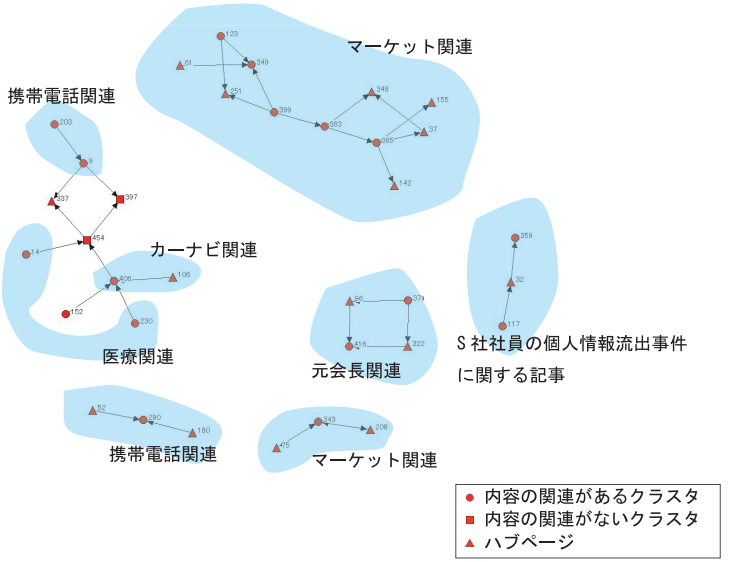

図 17　縮約されたグラフ
Fig. 17　Reduced graph.

9.3 提案手法と Cocitation の比較

任意のページに対して Cocitation を適用することで得られた関連ページと，提案手法を適用することによってこのページが属する極大関連クラスタの他のページとの比較を行った．ただし，極大関連クラスタは，α と β がともに 0.5 のときに出力されるものを用いた．

Cocitation とは，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ページを発見する手法で，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存在する場合，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，これらの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手法である．

はじめに，S 社の充電池を紹介するページ（**図 18** の星印）について比較を行った．このページに対して Cocitation を適用すると，2 つ以上のページから参照共起関係にある 10 ページが得られ，このうち，3 ページが同一トピックである S 社の充電池を紹介するページであった．同一トピックではないページが 7 ページ含まれていたのは，S 社に関するはて

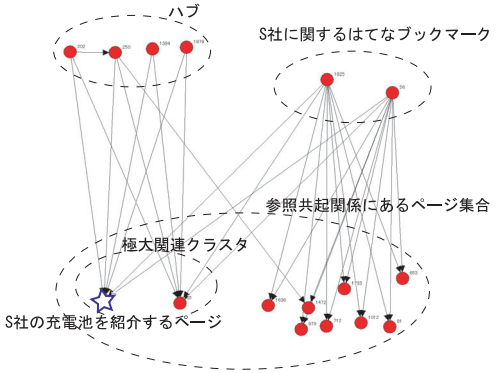

図 18　充電池を紹介するページの例
Fig. 18　Pages on rechargeable batteries.

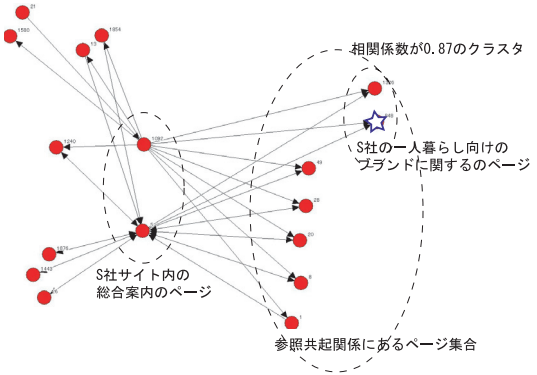

図 19　S 社の 1 人暮らし向けのブランドに関する S 社のサイト内のページの例
Fig. 19　Pages on S company's products for single life.

なブックマークの 2 つのページが，S 社に関するさまざまなトピックに対してリンクを張っていたためであった．一方，提案手法によってこのページは充電池を紹介する他のページとともにクラスタを形成した．充電池に関する他の 2 つのページは集中度が 0.67 でハブになり，このクラスタは極大関連クラスタとして出力された．S 社に関するはてなブックマークの 2 ページのこのクラスタに対する集中度はそれぞれ 0.08 と 0.17 で，このクラスタ以外のページにも多数リンクを張っていたため，これらのページはハブとはならず，その上これら参照共起関係にあった 10 ページは隣接関係において正の相関を示さなかったため，デンドログラムにおいて 1 つのクラスタとして見なされることはなかった．このように，同じ極大関連クラスタ内に関連のないページが含まれることを防ぐことができた．

次に，S 社の 1 人暮らし向けのブランドに関する S 社のサイト内のページ（**図 19** の星印）について比較を行った．このページに対して Cocitation を適用すると，2 ページ以上から参照共起関係にある 7 つのページが得られた．このページは S 社サイト内の総合案内の 2 つのページと参照共起関係にあったため，7 ページすべてが S 社サイト内にある他の商品を紹介するページで，共通のトピックについて書かれたページではなかった．一方このページは，提案手法によって他の商品を紹介しているページとクラスタを形成していたが，このクラスタに対する S 社サイト内の総合案内のページの集中度は，それぞれ 0.2 と 0.18 となりハブとはならず，この S 社の 1 人暮らし向けのブランドに関する S 社のサイト内のページは，どの極大関連クラスタにも含まれなかった．S 社サイト内の総合案内の 2 つのページは，S 社サイト内のページとのつながりが強い一方，S 社サイト内のそれぞれのページの隣接関係における相関は高くなかったため，ハブにはならなかった．このようなページは参照共起関係にあるページを特定したとしても，同一トピックの関連ページを発見できなかった．そこで，このようなページを分類の結果から除外することは，Cocitation と比べて分類の精度を向上させる可能性が高いと考えられる．

10.　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付いているページや，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，大規模なページ集合を扱うために，データ収集方法や前処理の方法を検討する．さらにリンク解析による Web ページを分類する他の手法との比較を行うことである．

謝辞　本研究の一部は，総務省 SCOPE プロジェクト「インターユビキタスネットワーク情報基盤の研究（課題番号 071607001）」からの補助により遂行されたものである．

参 考 文 献

1) Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y. and Ma, J.: Learning to Cluster Web Search Results, *Proc. 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2004).
2) Zamir, O. and Etzioni, O.: Web Document Clustering: A Feasibility Demonstration, *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1998).
3) Dean, J. and Henzinger, M.R.: Finding Related Pages in the World Wide Web, *Computer Networks*, Amsterdam, Netherlands (1994).
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，情報処理学会研究会報告，2001-FI-61-7, pp.45–52 (2001).
5) 渡邊　剛，小坂　武：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　剛史，松尾　豊，市瀬龍太郎，武田英明，石塚　満：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) Hanneman, R.A. and Riddle, M.: Measures of similarity and structural equivalence. http://www.faculty.ucr.edu/˜hanneman/nettext/C13_%20Structural_Equivalence.html
8) Larson, R.R.: Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace, *Proc. ASIS Annual Meeting*, Vol.33, pp.71–78 (1996).
9) 安田　雪：実践ネットワーク分析，新曜社 (2001).
10) 安田　雪：ネットワーク分析，新曜社 (1997).
11) Wasserman, S. and Faus, K.: *Social Network Analysis*, Cambridge University Press (1999).
12) Kleinberg, J.M.: Authoritative Sources in a Hyperlinked Environment, *J. ACM* (1999).

(平成 20 年 4 月 17 日受付)
(平成 20 年 6 月 18 日再受付)
(平成 20 年 6 月 26 日採録)

山下　長義（学生会員）

2006 年大阪大学大学院情報科学研究科情報数理学専攻修了．修士（情報科学）．現在同大学院情報科学研究科博士後期課程に在学中．Web マイニングの研究に従事．

森山　甲一

1998 年東京工業大学工学部情報工学科卒業．2003 年同大学院情報理工学研究科計算工学専攻博士課程修了．博士（工学）．同年同専攻助手．2005 年大阪大学産業科学研究所助手．2007 年同助教．現在に至る．人工知能，特にマルチエージェントシステムにおける強化学習の研究に従事．人工知能学会会員．

沼尾　正行（正会員）

昭和 62 年東京工業大学大学院博士課程修了．工学博士．同大学工学部情報工学科助手，講師，助教授，同大学院情報理工学研究科計算工学専攻助教授を経て，現在，大阪大学産業科学研究所知能システム科学研究部門教授，（独）日本学術振興会学術システム研究センター主任研究員．人工知能，機械学習の研究に従事．人工知能学会，日本認知科学会，日本ソフトウェア科学会，電子情報通信学会，AAAI，ACM 各会員．

栗原　　聡（正会員）

1992 年慶應義塾大学大学院理工学研究科計算機科学専攻修士課程修了．同年日本電信電話株式会社入社．基礎研究所を経て未来ねっと研究所に所属．1998 年から慶應義塾大学大学院政策・メディア研究科専任講師（有期）．現在，同大学環境情報学部非常勤講師．2004 年から大阪大学産業科学研究所知能システム科学研究部門准教授（同大学大学院情報科学研究科情報数理学専攻准教授兼務）．マルチエージェント，ネットワーク科学等の研究に従事．著書『社会基盤としての情報通信』（共立出版，共著）．翻訳『スモールワールド』（東京電機大学出版局，共訳）．編集『Emergent Intelligence of Networked Agents』（Springer in Computational Intelligence Series）等．博士（工学）．人工知能学会，日本ソフトウェア科学会，ESHIA，各会員．