

## 構造同値に基づく Web ページの分類

山下 長義<sup>†</sup> 森山 甲一<sup>††</sup> 栗原 聡<sup>††</sup> 沼尾 正行<sup>††</sup>

<sup>†</sup> 大阪大学情報科学研究科情報数理学専攻

〒 567-0047 大阪府茨木市美穂ヶ丘 8-1

<sup>††</sup> 大阪大学産業科学研究所

〒 567-0047 大阪府茨木市美穂ヶ丘 8-1

E-mail: †{nagayosi,koichi,kurihara,numao}@ai.sanken.osaka-u.ac.jp

あらまし Web 空間から必要な情報を見つけることが難しい状況において、構造同値に基づいたリンク解析を行うことで Web ページを分類し、関連する部分を特定する方法を提案する。構造同値のページとは、グラフのなかでページのラベルを入れ替えても、全く相互の関係パターンが変化しないようなページのことをいう。本論文では、構造同値の関係にあるページ同士は内容も関連していると想定し、Web ページの分類を行う。そして、ハブというページを独自に定義することで関連している部分を特定する。予備実験を行った結果、Web ページの構造的な類似度に基づく分類が有効であることがわかった。

キーワード Web 構造マイニング, 情報検索, 社会ネットワーク

## Web Page Classification Based on Structural Equivalence

Nagayoshi YAMASHITA<sup>†</sup>, Koichi MORIYAMA<sup>††</sup>, Satoshi KURIHARA<sup>††</sup>, and Masayuki NUMAO<sup>††</sup>

<sup>†</sup> Department of Information and Physical Sciences, Graduate School of Information Science and Technology, Osaka University

8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan

<sup>††</sup> The Institute of Scientific and Industrial Research, Osaka University

8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan

E-mail: †{nagayosi,koichi,kurihara,numao}@ai.sanken.osaka-u.ac.jp

**Abstract** Due to difficulty finding necessary information from the Web, we propose a novel methodology for classifying Web pages using link analysis based on structural equivalence. Two nodes are structurally equivalent if these nodes have the same relationship with other nodes. In this paper, assuming that pages with similar link relationship among the other pages are also similar in contents, Web pages are classified and then related parts are specified by defining the hub page. A preliminary experiment shows classification based on the structural similarity of web pages is effective.

**Key words** Web structural mining, Information retrieval, Social network

### 1. はじめに

近年のインターネットの急激な成長により、WWW から瞬時に大量の情報を得られるようになった。しかしその反面、情報洪水と言われるように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。たとえば検索エンジンを使用しているユーザは、調べたいことに対して漠然としたイメージしか持っていないことが多い。断片的な検索語によって得られた

検索結果から、試行錯誤の末に目的を明確にするとともに、自分に必要な情報を含む Web ページを絞り込んでいく。このような試行錯誤を解消するため、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] や、分類したものをユーザに提示するサービスが実用化されている。しかし、言語処理では言語の曖昧性や多義性を扱うことが難しいため、上手く分類できないのが現状である。

一方、WWW はハイパーリンクを有することから、WWW

をネットワークとみなすことで情報検索においてリンク解析を行う研究が行われている。この分野において Web ページ間のハイパーリンクによる隣接関係が似ているページ同士は内容も似ていると仮定し、任意の Web ページに対する関連ページを発見する研究が行われている [3]。そこで我々も社会ネットワークにおける構造同値という概念に基づいて Web ページの分類を行う研究を進めている。

構造同値のページとは、グラフの中でページのラベルを入れ替えても全く相互の関係が変化しないようなページのことをいう。分類に際して実際に存在するネットワークにおいて完全に構造同値の関係にあるノードは少ない。そこで、隣接行列における行同士と列同士の相関を計算することで構造同値の度合いを求め、この相関係数の値を基に階層的クラスタリングによってノードを分類する手法が社会ネットワーク研究分野において開発されている [4]。しかし、階層的クラスタリングにおいてどの部分までを 1 つのクラスタに分類するかという絶対的な指標がない。そこで、どの部分までが関連しているかを判断する手法を提案する。実際に検索語を入力して得られた Web ページの集合に対して初期実験を行い、提案手法の有効性を示す。

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 節で実験と評価を行い、6 節で考察を行い、7 節にてまとめとこれからの課題を述べる。

## 2. 関連研究

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]。

また、巨大で複雑なネットワークの中の似た結合パターンをもつサブネットワークを EM アルゴリズムを用いて発見する方法 [7] が提案されている。

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 つのページを参照共起されているページとし、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] がある。また、収集した数千の Web ページの中から人手によって選び抜かれたページをサイト別にまとめ、これらサイト間の参照頻度の相関を求める。そして、高い相関を持つサイト同士は類似しているとみなし、得られた相関行列を可視化し、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によって分類する研究も行われている。たとえば、Snipet (検索語にヒットした箇所の周辺のテキスト) とタイトルを手がかりに、機械学習に基づいてラベル候補となるフレーズを発見

し、そのフレーズを含む Web ページをグループ化するクラスタリング手法 [1] や、共通の単語やフレーズを含む Web ページ集合をグループ化し、クラスタとそのラベルを同時に生成する Suffix Tree Clustering (STC) 手法 [2] がある。サービスとして実用化されているものには Clusty<sup>(1)</sup> や Grokker<sup>(2)</sup> などがある。Clusty は、検索結果を自動的にフォルダ分けして表示するシステムで、言語解析と統計分析を組み合わせた独自のアルゴリズムを用いている。Grokker は、検索結果を種類分けしたものを可視化するシステムで、自然言語処理と Human Computer Interface の研究者によって開発されている。

これらに対して本研究は、言語処理を用いずにリンク構造の類似度から Web ページの集合を分類する手法に区分される。

## 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を説明し、最後に階層的クラスタリングについて説明する。

### 3.1 構造同値の例と定義

小学校の徒競走やオリンピックの 100m 走などで選手は、スタートラインに一列に並んで、審判の合図に合わせてスタートを切る。このとき、第一レーンにしようが第三レーンにしようが、審判と競技者の関係には変わりはない。審判から見ればどのレーンの選手もみな対等な選手であり、どの選手から見ても審判は審判であり、選手たちとは異なった位置を占める。このような状況にある選手たちは構造同値であるという [9]。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。すなわち、グラフの中でノードのラベルを入れ替えても、全く相互の関係パターンが変化しないようなノードの集合を構造同値のノードという [10]。このように、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。

### 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 [9]。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 [4]。ただし、対角成分を除く  $i$  行の値の平均を  $\bar{x}_i$ 、同様に  $i$  列の値の平均を  $\bar{x}_i$  とし、合計は  $k$  に対して行い、 $i \neq k$ 、 $j \neq k$  で

(注1) : <http://clusty.jp/>

(注2) : <http://www.grokker.com/>

ある。

$$r_{ij} = \frac{A+B}{C \cdot D} \quad (1)$$

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

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

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

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

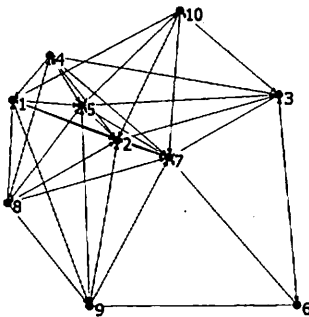


図1 グラフの例

例として図1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2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。

### 3.3 階層的クラスタリング

隣接行列における相関を基にノードを分類する方法として階層的クラスタリングを用いる。3.2節の図1の例では、結果を図3のように集合に基づいた方法で表現することができる。相関係数が1であるページ集合がクラスタを形成するとき(レベル1)、10個のノードそれぞれが1つのノードのみを含むクラスタを形成している。相関係数が0.63以上のページ集合がクラスタを形成するとき(レベル2)、ノード2と5とが1つのクラスタを形成し、この2つのノードは常に1つのクラスタ内に留まって次のレベルに引き継がれる。相関係数が0.45以上のページ集合がクラスタを形成するとき(レベル3)は、ノ

ドはそれぞれ  $(1,2,4,5,7)$ ,  $(3,8,9)$ ,  $(6,10)$  の部分集合に分類される。このように隣接行列における相関を基に連続的にクラスタを融合して、各レベルでのクラスタを形成していく。階層的クラスタリングは、手続きがわかりやすく解釈が明確であるため広く用いられている。しかし、どのレベルまでに形成されたクラスタを結果として用いるかという決定的な基準が存在しない。

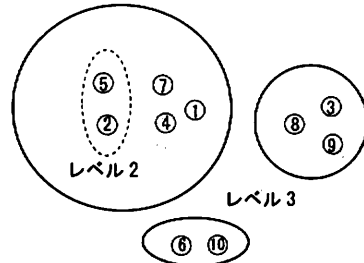


図3 集合に基づいた表現方法

## 4. 提案手法

構造同値の度合いが高いページ同士は、内容も関連していると仮定する。すなわち、隣接行列の相関による階層的クラスタリングで出現する各クラスタについて、含まれるページの内容が互いに関連する最大のレベルが存在するものとする。ページ分類結果として出力される各クラスタをそのような最大レベルのものに近づけるために、本論文では「ハブ」ページを定義し、ハブが存在する最大のクラスタの集合を結果として出力する手法を提案する。

提案手法の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計算する(3.2節)。
- (3) これらの値をもとに階層的クラスタリングを完全連結法<sup>(注3)</sup>によって行う(3.3節)。相関係数が0から1の間に起こる各融合で形成される部分集合をクラスタとし、加えて相関係数が1であるページ同士はあらかじめ1つのクラスタに属しているものとする。
- (4) (3)の階層的クラスタリングにおける同じレベルで融合されるクラスタ集合ごとに定義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。
- (5) (4)で求めた極大関連クラスタの集合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  $C_k$  との間の密度  $\Delta_{i \rightarrow C_k}$  と  $\Delta_{C_k \rightarrow i}$  を以下のように定義する。

$$\Delta_{i \rightarrow C_k} = \frac{\sum_{j \in C_k} X_{ij}}{|C_k|} \quad (2)$$

(注3)：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

	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	0.31	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	0.50	0.29	0.28	0.30	0.07	0.33	0.20	1.00

図2 図1のグラフの隣接行列における相関

$$\Delta_{C_k \rightarrow i} = \frac{\sum_{j \in C_k} X_{ji}}{|C_k|} \quad (3)$$

ただし、

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

とする。

[定義 2] 任意のページ  $i$  とクラスタ  $C_1, C_2, \dots, C_n$  があると  
する。このとき、ある  $1 \leq l \leq n$  において

$$\Delta_{i \rightarrow C_l} \geq \alpha \quad \text{かつ} \quad \Delta_{i \rightarrow C_m} < \alpha \quad (\text{for all } m \neq l)$$

または

$$\Delta_{C_l \rightarrow i} \geq \alpha \quad \text{かつ} \quad \Delta_{C_m \rightarrow i} < \alpha \quad (\text{for all } m \neq l)$$

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。ただし、すべてのページが任意の極大関連クラスタに含まれるわけではない。

[定義 3]  $C$  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ぶ。

Web ページ上のハイパーリンクは、そのページ上の文書と内容において関連がある Web ページに対してリンクを張ることが多いことから、同じ Web ページとの間でリンクが存在するページ同士は内容において関連していると考えられる。そこで、階層的クラスタリングにおける同じレベルで融合されるクラスタ集合ごとに定義 1 から定義 3 を適用することで、極大関連クラスタを出力する。そして、極大関連クラスタに含まれるページの内容が互いに関連する最大レベルとみなす。つまり極大関連クラスタ内のページは、隣接関係において正の相関を示すページ同士であり、かつこれらのページとの間にリンクを有するページが存在する最大のページ集合である。

たとえば、図 4 においてクラスタ 1 (ページ 2, 4, 5, 7) とクラスタ 2 (ページ 1) が融合されるとき、仮にページ 10 のようにクラスタ 1 に対してのみリンクを張っていて、クラスタ 2 に対してリンクを張っていないページがあるとする。この場合定義からページ 10 をハブとし、クラスタ 1 を関連クラスタとする。ハブとなるページ 10 は、クラスタ 1, 2 間における破線のように境界を定めるものと考え、この境界よりも内側 (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. 実験と評価

### 5.1 データの収集方法と詳細

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 を得る。そして、これらのページからリンク

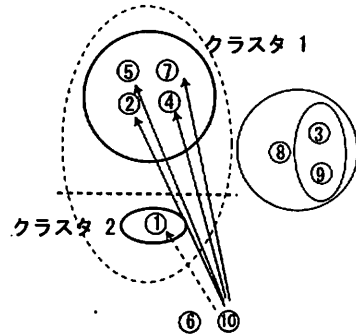


図 4 ハブの例 (この例ではページ 10 がハブとなる。)

クが張られているページと、これらのページに対してリンクを張っているページを収集する。ただし、異なるドメイン間のリンクのみを用いる。この集合内のページには、検索語と関連のないページが多数含まれているが、将来より効率的なデータ収集が機械的にできることを想定して、検索語と関連のないページは取り除いた。

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 (収集した全ページ数 1968)

このようにして得られたページの中には、S 社のトップページのような入次数が 117 ページであるような極端なページもあった。

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 1.73 ページであった。全体の約半数の 260 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。

そして階層的クラスタリングを行い、階層的クラスタリングにおいて相関係数が 0 から 1 の間に融合されるすべてのクラスタに対して、提案手法によりハブを同定した。また、クラスタに対してハブとなるために必要な密度  $\alpha$  をこの実験では 0.33 とした。

極大関連クラスタは 85 個形成され、極大関連クラスタ内にある平均ページ数は 4.55 であった。また、すべての極大関連クラスタ内の総ページ数は 387 ページで、クラスタ内の文書数の出現頻度分布はべき分布に従った (図 5)。今回は極大関連クラスタの有効性を検証するため、極大関連クラスタ内に含まれるページのみを評価の対象とする。

極大関連クラスタに対するハブの数の出現頻度はべき分布を示した (図 6)。ハブの数が 1 つの極大関連クラスタは 62%、ハブ数が 3 つ以内の極大関連クラスタが全体の 87% を占めた。

### 5.2 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

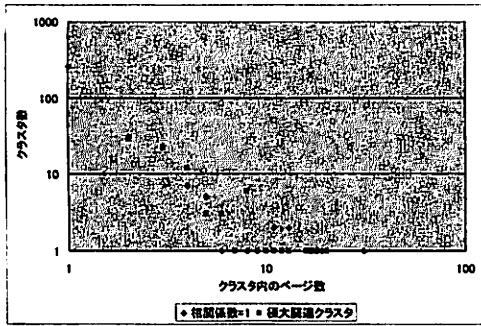


図5 クラスタ内のページ数の出現頻度分布

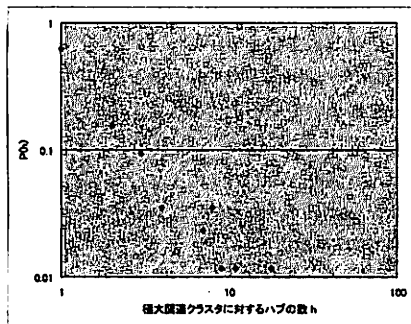


図6 極大関連クラスに対するハブの数の出現頻度分布

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は内容が関連しているとみなした。たとえば、以下のような言葉は共通の概念を持つ言葉であるとみなした。

- 「商品名」と「製品名」
- 「ナビ」と「カーナビ」
- 「充電機」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」
- 「半導体」と「電子デバイス」
- 「太陽電池」と「電池」

85個の極大関連クラスのうち内容が関連があるクラスは63個あり、85個の極大関連クラス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74%であった。また、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ーの精度は73%であった。このような結果から、完全に構造同値であるページで構成されるクラスターを、極大クラスターに融合しても、精度は下がらなかった。

図7は、相関係数を横軸に精度を縦軸にとったものである。たとえば、横軸において0.4とは、0.4以上の相関係数を持つ極大関連クラスターの精度を示している。相関係数が0.8以上の極大関連クラスターにおける精度は0.8を超えていたが、相関係

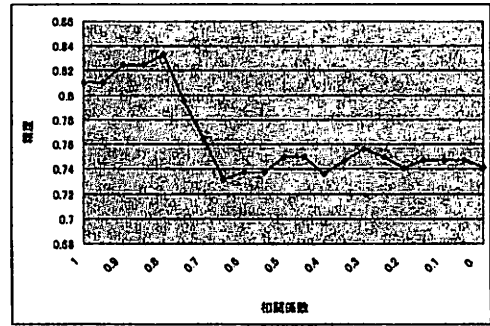


図7 相関係数に対する精度（横軸において0.4とは、0.4以上の相関係数を持つ極大関連クラスターの精度を示している。）

数が0.6以上の極大関連クラスターにおける精度は0.74付近まで下がり、以下精度は0.74辺りで推移した。

また図8は極大関連クラスターに対するハブの数を横軸に、精度を縦軸にとったものである。たとえば横軸で5とは、ハブを5個以上持つ極大関連クラスターの精度を示している。図8より、極大関連クラスターに対するハブの数とクラスター内のページの内容の関連度に相関が見られなかった。

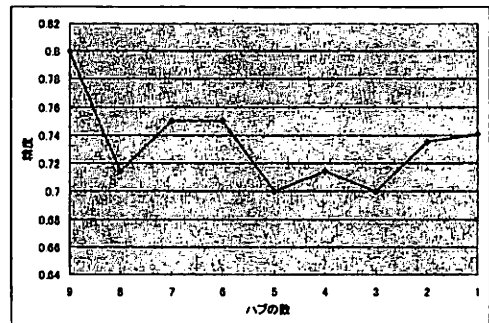


図8 ハブの数に対する精度（横軸で5とは、ハブを5個以上持つ極大関連クラスターの精度を示している。）

## 6. 考 察

S社のトップページのような次数の大きなページがハブとなった場合、被験者実験によって、その極大関連クラスターは内容は関連していないと評価されるものが多いことがわかった。たとえば、出次数が2のページが2つのページからなるクラスターのハブになった場合、このハブのすべてのリンクは1つのクラスター内のページに対してリンクを張っているが、出次数が20のページがこのクラスターのハブになった場合、少なくとも18のリンクは、他のページへもリンクを張っている。このような2つのハブを区別し、次数が大きなハブを除外する指標が必要であると考えられる。

そこで、任意のハブ*i*の極大関連クラスター*C<sub>j</sub>*に対する集中度*S<sub>i</sub>(C<sub>j</sub>)*を以下のように定義した。ハブ*i*の次数を*H<sub>i</sub>*、次数*H<sub>i</sub>*のうち1つの極大関連クラスター*C<sub>j</sub>*間にあるリンク数を*L<sub>i</sub>*とする。

$$S_i(C_j) = \frac{L_i}{H_i} \quad (5)$$

図9はすべての極大関連クラスタと集中度が0.5以上のハブを持つ極大関連クラスタを、精度と再現率によって比較したものである。ここで再現率は、すべての極大関連クラスタの中で内容が関連していると判断されたクラスタのうち、集中度0.5以上のハブを持つ極大関連クラスタの中で内容が関連していると判断されたクラスタの比を表している。集中度が0.5以上のハブを持つ極大関連クラスタは、すべての極大関連クラスタと比べ再現率は0.71であったが、精度は常に上回っていた。

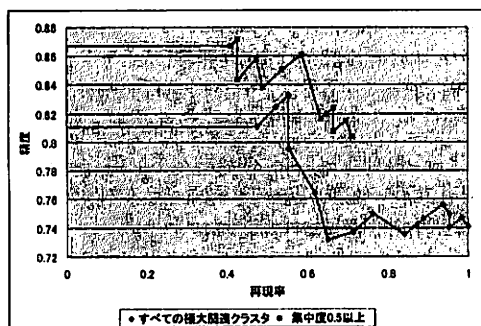


図9 集中度を考慮した再現率に対する精度

また Clusty に検索語として S 社を入力すると、検索結果がいくつかのフォルダに分類され表示される。その中に「事業」、「イオン」、「NTT」という名のフォルダがあった。「事業」という名前のフォルダの中には、メディカル・ヘルス事業、液晶パネル事業、岐阜事業所、携帯電話事業についてのページが分類され、「イオン」という名前のフォルダには、マイナスイオンドライヤーについてのページと S 社とイオン(株)との共同商品開発についてのページが分類されていた。また「NTT」というフォルダには、携帯電話のページと、S 社と NTT 東日本のパドミントンの試合結果のページが分類された。このように、言語処理による分類は特定の単語に影響されることが多く見られる。

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。言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すると、29個のクラスタに分類された。これら29個のクラスタには138個のページが含まれていた。被参照関係のみの相関で分類することの有意性を見出せなかった。これは解析対象のページが583と少なかったため、共参照関係にあるページが少なく全体の一部しか分類できなかったと考えられ、より多くのデータが必要であると考えられる。

## 7. まとめとこれからの課題

構造同値の度合いを相関係数によって求め、相関係数の値を基に階層的クラスタリングによってノードを分類する手法を

Web ページのリンク解析に用いた結果、リンク構造の類似度とページの内容における関連性に相関性があることがわかった。そこで、階層的クラスタリングにおいてどの部分までが関連しているかを判断するためにハブを定義し、その効果の検証を行った。その結果、分類精度が下がらずに、より少ない数のクラスタに分類できた。また、集中度という指標を導入することでハブを選別し、分類の精度を向上させることができた。

今後の課題としては、本手法における各種パラメータを変えたときの変化を検証する。そして、より客観的な被験者実験の方法を考えるとともに、他の手法との比較を行う。本論文では分類された結果に内容における関連性があるかどうかを検証したが、さらに個々の極大関連クラスタの範囲が適切であるかを検証し、高い分類精度でかつ少ない数のクラスタへの分類のために提案手法を改善する必要がある。

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ために検索語に関するページを効率的かつ自動的に収集する方法を前処理として行うことが必要である。また、現在 Web サービスとして多数公開されている API を用いた商品のネットワークデータの解析にも応用することを検討中である。

## 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] S. Wasserman and K. Faust: "Social Network Analysis", Cambridge University Press (1999).
- [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] 榎, 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] M. E. J. Newman and E. A. Leicht: "Mixture models and exploratory analysis in networks", Proc. Natl. Acad. Sci. USA 104, pp. 9564-9569 (2007).
- [8] R. R. Larson: "Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace", Proceedings of the ASIS Annual Meeting, 33, pp. 71-78 (1996).
- [9] 安田: "ネットワーク分析", 新曜社 (1997).
- [10] 安田: "実践ネットワーク分析", 新曜社 (2001).