LF-003

構造同値に基づくリンク解析を用いたWeb文書からのキーワード抽出
Keyword extraction From the Web Using Link Analysis Based Structural Equivalence

山下　長義 †　　沼尾　正行 ‡　　栗原　聡 ‡
Nagayoshi Yamashita　　Masayuki Numao　　Satoshi Kurihara

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．このような状況において近年 Web を対象にしたデータマイニングの研究が盛んに行われている．Web ページは主にテキスト情報によって構成されているため，Web から情報を抽出するデータマイニングではテキスト処理が必要不可欠である．しかし言語はあいまい性を含むため一般的に解析には困難を伴う．そこで，これまで言語解析とリンク解析を併用するような解析方法が提案されてきた [Zamir 98, Zeng 04].

情報検索のために用いられてきたリンク解析の手法は，直接つながっている関係からページを評価する「ミクロ」な視点からページをランク付けする方法であった [Kleinberg 98, Dean 98]．そこで本論文では，社会ネットワークの分野におけるポジションとロールという「マクロ」の視点からの手法をリンク解析に導入した．そして，これらを定式化した構造同値という概念を用いることにより，直接つながっていなくても他のページとの関係が等しければ，同じグループに分類することができるようになり，有用な情報を抽出することができた．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，いくつかの例を示す．

以下，2 節では Web を対象とするデータマイニングについて簡単に述べ，3 節で本論文で用いた関連技術について説明し，4 節で提案手法を説明し，5 節で評価を行い，6 節にてまとめを述べる．

2. 関連研究（Web を対象としたネットワーク解析）

Web 上の検索エンジンによって得られる結果を分類し，それぞれの集合に対する重要語を抽出する研究が行われている．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 [Zeng 04] 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 [Zamir 98] がある．サービスとして実用化されているものには Grokker§や Clusty¶などがある．

リンク解析に関する研究は，Web ページ間を接続するリンクに着目する手法が一般的であり，PageRank や HITS[Kleinberg 98] など実際にさまざまな検索エンジンに使用されている手法も提案されている．また，関連ページを発見するアルゴリズムとしては，HITS を応用した Companion[Dean 98] や，参照共起条件を用いた Cocitation[Dean 98] というアルゴリズムが提案されている．これらは直接つながっているノードからページを評価する「ミクロな視点」でネットワークを解析する手法であり，どれだけ他のページからリンクが張られているかが重要になる．

これに対して本研究では，Web ページ間の接続するリンクに対してではなく，Web ページ群がリンクによって互いに接続されることで構成される「ネットワークの構造」そのものに着目する [Yamashita 06].

3. 研究に用いた関連技術

3.1 構造同値による類似ネットワーク構造の抽出

看護婦は，医者や患者に対する関係から勤務している病院が違っても同じ看護婦としての地位を占める．お互いに知り合いであるからでもなく，同じ医者について同じ患者を診ているからでもない．看護婦として同じ地位を占める理由は，他のグループに属する人々との関係の類似性からである．看護婦という地位は，医者というグループに属する人々に付き，患者というグループに属する人々を診るという関係の中に存在する．このようにポジションとロールとは，他のノードとのリンク関係が等しければ，ノードは同じグループに分類されるという考え方である．これら定式化する方法の一つに構造同値という概念があり，ポジションとロールをモデル化する手法の中で最もよく用いられている．

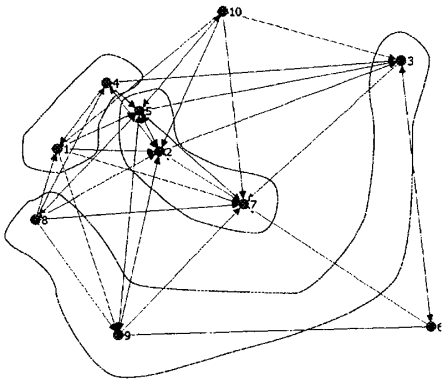

図 1: グラフの例

グラフ上のノードの接続関係を隣接行列によって表し，構造同値である度合いを求めるために隣接行列を用いて相関係数を計算する．相関係数とは 2 つのデータ列間の

†大阪大学大学院情報科学研究科情報数理学専攻
nagayosi@ai.sanken.osaka-u.ac.jp
‡大阪大学産業科学研究所　知能システム科学研究部門
§http://www.grokker.com/
¶http://clusty.jp/

類似性の度合いを示す統計学的指標であり，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 間の相関係数は以下のように定義することができる．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．

$$r_{ij} = \frac{\sum(x_{ki}-\overline{x}_{+j})(x_{kj}-\overline{x}_{+j})+\sum(x_{ik}-\overline{x}_{i+})(x_{jk}-\overline{x}_{j+})}{\sqrt{\sum(x_{ki}-\overline{x}_{+i})^2+\sum(x_{ik}-\overline{x}_{i+})^2}\sqrt{\sum(x_{kj}-\overline{x}_{+j})^2+\sum(x_{jk}-\overline{x}_{j+})^2}} \quad (1)$$

図1において，ノード 1,4 は共にノード 2,5,7 に対してリンクを張り，ノード 3,8,9 の多くからリンクが張られている．このように，ノード 1,4 は他のノードとの関係が等しく，構造的に同値であると言える．

3.2 Web のリンク空間におけるハブ

Web のリンクネットワークにおいて，ハブと呼ばれるページが存在する [Kleinberg 98]．ハブとは，共通のトピックに関するページにリンクを張っているページのことをいう．強い関連性を示すページはひとつのトピックについて書かれているため，これらに対して同時にリンクを張っているページ，すなわちハブが存在すると考えられる．したがって，ハブによって共通のトピックに関するページ集合を抜き出すことができる．本論文ではハブになるページの条件の一つを，構造的に同値の関係にあるページ集合に同時にリンクを張っていることとする．

4. 提案手法

4.1 検索エンジンの結果から，母集団となるページ間のリンク関係を抽出する．

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間のリンク関係を母集団とする．ただし，異なるドメイン間のリンクのみを用いる．

4.2 階層的クラスタリングによって，すべてのノードが一つのクラスタに含まれるまで融合を繰り返す．

得られたリンク構造に対してページごとの相関係数を計算することで，それぞれのノード対ごとの構造同値の度合いを求める．そして，これらのページごとの相関係数を基に階層的クラスタリングを行い，ノードを部分集合に分類する．階層的クラスタリングは，連続的にノードを融合して各レベルでクラスタを形成していく．

本論文ではノードを Web ページとし，Web 空間においてリンク構造が同一のページ同士は内容的にも関連していると仮定する．

図2は，ページ A～D,1～9 を含むグラフを階層的クラスタリングを行った結果である．たとえば，ページ A は，階層的クラスタリングによって他のページ B,C,D が属するクラスタと融合されることを示している．

4.3 融合されるクラスタ間で他のページとの隣接関係の差異を注目し，ハブとなるページを発見する．

一度に融合されるクラスタ間において，他のページとの隣接関係に違いが存在する．隣接関係が等しければそれぞれのクラスタは構造的に同値であり，はじめから一つのクラスタを形成していたからである．よって一つのクラスタのみにリンクを張っていて，このクラスタを特徴付けるページが必ず存在する．

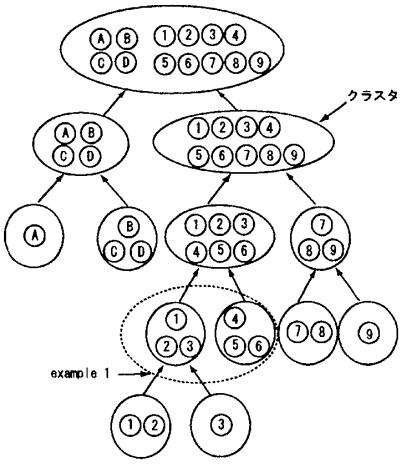

図 2: 分割プロセスの具体例

そこで融合される複数のクラスタのうち一つのクラスタに属するページに対してのみリンクを張っているページをハブとする．そしてクラスタに対してハブの関係にあるページが存在するとき，そのクラスタに属するページは互いに関連するページであると考えられ，このようなクラスタを関連クラスタとする．そして，ハブとなったページをこのクラスタに対する関連ページと呼ぶことにする．またクラスタに属する複数のページが特定の同一のページにリンクを張っている場合，クラスタに属するページは特定のページに対してハブとなるので，このようなクラスタも関連クラスタに含める．

次に，それぞれの融合についてハブとなるページの基準を示す．クラスタ k を C_k，i から j へのリンクを $X_{ij}=1$，クラスタ k に属するノード数を g_k，ページ i とクラスタ k 間の密度を Δ_{ik} とすると，この密度を以下のように求めることができる．

$$\Delta_{ik} = \frac{\sum_{j \in C_k} X_{ij}}{g_k} \quad (2)$$

そして，ハブとなるページの基準を以下のように定義する．融合される複数のクラスタのうち，1つのクラスタに対してのみ α 以上の密度 Δ を有するページを，そのクラスタに対するハブとする．以下では，図2上の example.1 の部分を図3に抜粋し，クラスタ 3,4 について α が 0.5 の場合を例に説明する．

- ページ A は，クラスタ 3 に含まれるすべてのページに対してリンクを張り (Δ_{A3}=1)，クラスタ 4 に含まれるすべてのページにリンクを持たない (Δ_{A4}=0)．よってページ A は，クラスタ 3 に対してハブとなる．
- ページ B は，クラスタ 3 に含まれる 3 ページのうち 1 ページに対してリンクを張り (Δ_{B3}=0.33)，クラスタ 4 に含まれる 3 ページのうち 2 ページに対し

てリンクを張っている (Δ_{B4}=0.67)．よってページ B は，クラスタ 4 に対してハブとはならない．

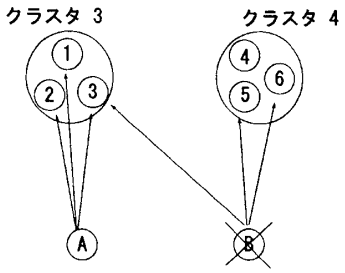

図 3: 関連ページの基準

そして，すべての融合において以上の基準でハブを同定する．図 2 では，5ヶ所の融合すべてにおいて同様にしてハブを同定する．

このように連続的に融合して形成されるクラスタにおいて，どのレベルまでに形成されたクラスタが互いに関連するページであるかどうかを，Web のリンク構造の特徴であるハブとなるページを同定することで行う．

4.4 すべての文書について名詞の重み付けを計算する

それぞれのページの文書に対して形態素解析を行って名詞のみを取り出す．TFIDF によって名詞の重みを求め，名詞の重みベクトルに変換する．そして，同じ関連クラスタ内に含まれていたページの重みベクトルを用いて，関連クラスタごとの名詞に対する重みベクトルを式 (3) によって計算する．

クラスタ i に属するページの重みベクトル Q_i は，

$$Q_i = \frac{1}{N_c}\sum_{i=1}^{N_c} D_c \qquad (3)$$

のように求める．N_c は関連クラスタに属するページとそのクラスタに対してハブとなるページの合計，D_c は関連クラスタ i に属するページと，そのクラスタに対してハブとなるページそれぞれの重みベクトルである．

5. 実験

5.1 データ収集

データに関する詳細は以下の通りである．

- 検索語　三洋電機
- ページ数　2701

この母集団内のリンク構造に対して相関係数を求め，階層的クラスタリングを行った．そして，提案手法によりリンク解析の結果を反映した名詞に対する重み付けと，TFIDF を用いた重み付けの比較を行った．提案手法によるそれぞれのページに対する名詞の重み付けとは，ページが属するクラスタの名詞の重みベクトルに，そのページに出現する名詞が要素の単位ベクトルを乗じたベクトルにより示される．また，クラスタに対してハブとなるために必要な密度 Δ をこの実験では 0.33 とした．

5.2 得られたクラスタ

はじめに，三洋電機製のデジタルムービーカメラ XactiDMX-CA6 の公式ページ‖を取り上げ考察を行った．DMX-CA6 とは，手の中サイズでムービーも写真も 1 台でこなすメモリーカード搭載で生活防水タイプのムービーデジタルカメラである．このページは，三洋電機製のデジタルムービーカメラの DMX-HD1 の公式ページともに同じクラスタに属していた．そして，このクラスタに対してハブとなったページは以下の通りであった．他社の製品も同一のページに扱っているビデオカメラの通販のページが 5 ページ，DMX-CA6 を取り上げた記事，DMX-HD1 を取り上げた記事，kakaku.com 上の DMX-CA6 のページ，はてなブックマーク上の「Xacti」を含む注目記事のページ，「tobuy」を含む注目記事のページ，三洋電機のイベント告知のページである．

次に，TFIDF の値と提案手法を用いて得られた名詞に対する重み付けの値を比較した．図 4 の横軸は個々の名詞を表し，縦軸はこれら名詞に対する重み付けの大きさを表している．図 4 に示すように提案手法により「デジカメ」，「デジタル」，「写真」，「カメラ」，「カード」などのキーワードとなるべき語を抽出することができた．

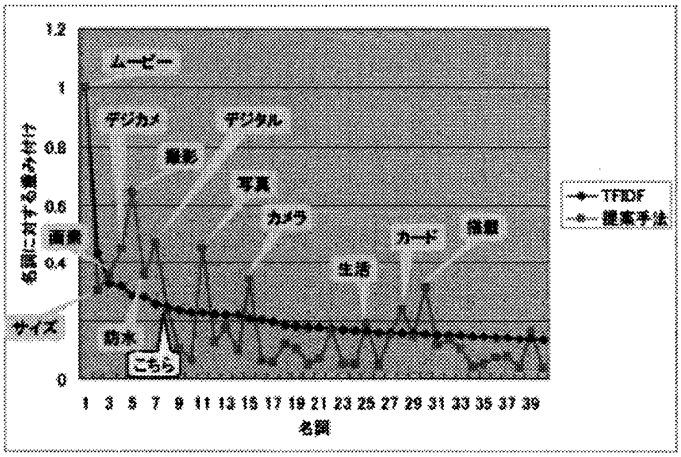

図 4: Xacti に関する文書に対する名詞の重み付けの値の比較

もう一つの例として，三洋電機の社員の画像と氏名がネット上に流出し，mixi から個人情報が漏洩したことに関する記事**に対して，同様に提案手法を用いた結果の分析をした．提案手法を用いた結果，同一クラスタに含まれているページと，このクラスタに対してハブとなったページに隣接しているページを全体から抜粋し，図 5 に示す．図 5 において，このページはノード 1594 にあたり，このページと同一のクラスタに属するページは，三洋電機の社員の画像流出に関するページ (ノード 17)，三洋電機の社員の事件を契機に，運営事務が mixi のアカウントを強制削除する事例が急増しているというニュース記事 (ノード 2348)，SNS 上での個人情報の扱い方を論じているページ (ノード 1769)，三洋電機の株価につい

‖http://blog.livedoor.jp/paintbox77/archives/50662639.html
**http://cortro.net/m/2006/10/miximixi_1.html

てのページ (ノード 132) と三洋電機の一般の話題を扱ったページ (ノード 177) である．また，このクラスタに対してハブになったページは以下の通りであった．三洋電機社員の画像流出事件のページ (ノード 1453)，mixi のログインのページ (ノード 2063)，はてなブックマークのキーワード「個人情報流出」を含む注目記事のページ (ノード 2118)，はてなのログインのページ (ノード 2006) である．

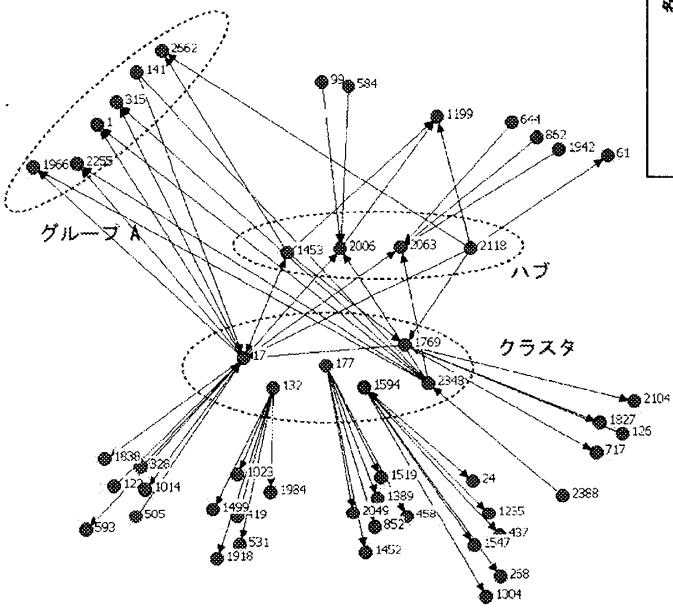

図 5: 画像流出事件に関するページ周辺のリンク関係

図 5 においてグループ A に属するページは，クラスタに属する複数ページとリンクを有しているが，他のページにもリンクを張っているため，ハブではないと判断されたページ集合である．これらページは，提案手法で定義したハブのようなネットワーク構造が似ているページ集合に対してだけリンクを有しているのではなく，より広い範囲のページとリンクを有している．したがって，クラスタに属するページと内容の上でもあまり関連性がないか，またはより広いトピックを網羅するページではないかと考えられる．これらページの内訳は，三洋電機のトップページ，はてなブックマークのタグ「三洋電機」を含むページのリンク集，J-CAST というニュースサイトの英語で書かれたホンダについての記事のページ，J-CAST というニュースサイトの中国語記事のページ，J-CAST の日本語のトップページ，はてなブックマークのヘルプページであった．

図 6 に示しているように，提案手法によって「流出」，「社員」，「写真」，「画像」，「事件」，「炎上」などの言葉を抽出することができた．

6. まとめ

Web のリンク構造における類似度を利用して関連サイトを同定し，文書間の名詞の重みベクトルを用いることより重要な単語が抽出できた．

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．またクラスタ間を比較するとき単純に共通している名詞の TF·IDF を変えたが，単語の重み付けの比較方法の更なる検討が必要である．さらに，このアルゴリズムを適用する範囲を広げ情報検索の分野に応用することを検討中である．

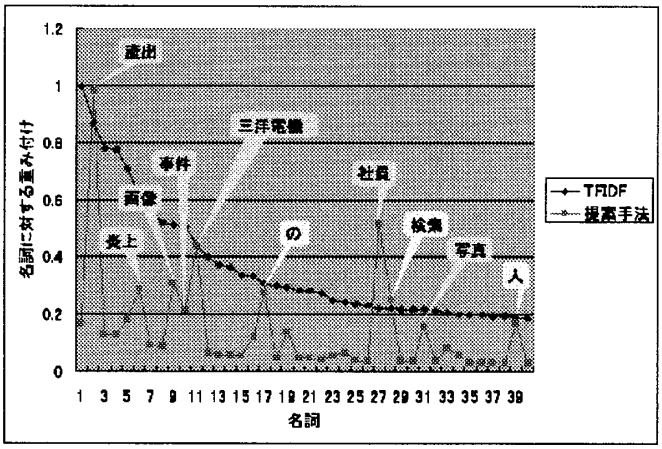

図 6: mixi による画像流出事件に関する文書に対する名詞の重み付けの値の比較

参考文献

[Zeng 04] Hua-Jun Zeng Qi-Cai He Zheng Chen Wei-Ying Ma Jinwen Ma.*Learning to Cluster Web Search Results* Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004.

[Zamir 98] Oren Zamir, Oren Etzioni.*Web Document Clustering: A Feasibility Demonstration* Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval , 1998.

[Dean 98] Jeffrey Dean, Monika R. Henzinger.*Finding Related Pages in the World Wide Web* In Proceedings of the 8th International World Wide Web Conference 1998, pages 389-401. 7, 1998.

[Kleinberg 98] Kleinberg, J. *Authoritative sources in a hyperlinked environment.* Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[Wasserman 94] Stanley Wasserman, Katherine Faust. *Social Network Analysis* Cambridge university press,1994

[Yamashita 06] Nagayoshi Yamashita, Masayuki Numao, and Satoshi Kurihara. *Salient Word Extraction Using Link Analysis Based on Structual Equivalence* Proccedings of the 5th 21st Century COE "Towards Creating New Industry Based on Inter-Nanosicence" International Symposium. Awaji, Japan. Dec 2006. p. 161.