F_034

CONCORによるリンク解析を反映したTF·IDFによるWeb文書の要約
Web Documents Summarization using Link Analysis Based on CONCOR

山下 長義 † 福井 健一 ‡ 森山 甲一 ‡ 栗原 聡 ‡ 沼尾 正行 ‡
Nagayoshi Yamashita Kenichi Fukui Koichi Moriyama Satoshi Kurihara Masayuki Numao

1. はじめに

World Wide Webは双方向性を持ったメディアであり，近年オンラインで大量のデータを扱えるようになったことで，Webを対象にした研究が盛んに行われている．しかし，主な解析対象である言語表現はあいまいであるため，言語のみで正確に文書を解析するのは難しいのが現状である．一方でWebのリンク構造を解析することでコミュニティを発見する研究やキーワードに適合するサイトをランク付けする研究が行われている．

そこで，本論文ではWeb上の文書における重要単語抽出のために，リンク構造を用いる手法を提案する．まず，類似サイトを特定するためにブロックモデルを用いてマクロな視点から分割を行い，提案するアルゴリズムにより類似サイトを特定する．そして，要約する対象のサイトとその類似サイトを比較することによって，単一の文書では頻度が小さくても重要である単語を発見することを目指す．最後に，これらの重み付けをされた名詞を入力としてWeb文書の要約を行う．

2. Webを対象としたネットワーク解析

Webを対象としたネットワーク解析にはコミュニティの発見や情報検索に利用するためにサイトをランク付けする方法がある．共通していることはサイトをノード，リンクを辺と見てWebをネットワークとしてとらえることである．Web上のコミュニティを見つける手法としては，社会ネットワークの分野の中心性 [Girvan 02] を利用する手法やクリーク [Palla 05, Everett 98] を利用する手法，グラフ理論の最大流最小カット定理 [Flake 02]，2部グラフ [Kumar 99] を用いる研究などが提案されている．サイトをランク付けする手法にはPageRankやHITS[Kleinberg 98] などがある．また，Web全体のリンク構造が蝶ネクタイの形をしたものであると主張する [Border 00] など，Webを対象としたリンク解析の研究は盛んに行われている．Webのリンク解析に用いられている手法はほとんどが，どれだけ他のサイトからリンクが張られているかという直接つながっているノードからサイト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る．また，コンピュータの性能が向上し実際に存在する大規模なネットワークを扱えるようになり，WWW全体がスケールフリー性 [Watts 98] とスモールワールド性 [Barabasi 99] を有することが示されている．

本論文では社会ネットワークの分野で用いられている解析手法，ブロックモデルの一つであるCONCOR[Wasserman 94] をWebに対して適用し，マクロな視点からリンク解析を行う．どのサイトが重要かではなく，ネットワーク全体の構造から類似度を評価する．ブロックモデルではノードが同じクラスタに分割されるために中心性やクリークによる解析のように直接つながっている必要はない．他のクラスタとの関係が同じサイトが同一のサイトに分類される．このようなマクロな視点で解析することで，「類似サイト」を特定し，比較することで名詞の重み付けを変更する．

3. CONCORによる分割

CONCORは構造同値の概念を利用するネットワーク解析手法であり，これを用いてWebのリンク構造を解析し，類似するサイトを発見する．構造同値では他のすべてのノードとの結合パターンが同じであれば，ノードは同一のクラスタに分類される．たとえば，以下のような1から9までの数をラベル付けされたノードの隣接行列

	1	2	3	4	5	6	7	8	9	10
1	–	1	0	0	1	0	1	0	1	0
2	1	–	1	1	1	0	1	1	1	0
3	0	1	–	1	1	1	1	0	0	1
4	1	1	0	–	1	0	1	0	0	0
5	1	1	1	1	–	0	1	1	1	1
6	0	0	1	0	0	–	1	0	1	0
7	0	1	0	1	1	0	–	0	0	0
8	1	1	0	1	1	0	1	–	1	0
9	0	1	0	0	1	0	1	0	–	0
10	1	1	1	0	1	0	1	0	0	–

をCONCORによって分割すると，

	1	4	5	2	7	8	3	9	6	10
1	–	0	1	1	1	0	0	1	0	0
4	1	–	1	1	1	0	0	0	0	0
5	1	1	–	1	1	1	1	1	0	1
2	1	1	1	–	1	1	1	1	0	0
7	0	1	1	1	–	0	0	0	0	0
8	1	1	1	1	1	–	0	1	0	0
3	0	1	1	1	1	0	–	0	1	1
9	0	0	1	1	1	0	0	–	0	0
6	0	0	0	0	1	0	1	1	–	0
10	1	0	1	1	1	0	1	0	0	–

となり，部分行列間の接続パターンによって分類される．それぞれの部分行列においてノードの過半数が1ならば1，それ以外を0に置き換えると，表1のように隣接行列を縮約して表すことができる．この縮約された行列はクラスタ間の関係を示している．クラスタ1に分類されたノードはクラスタ2に対してリンクを持ち，またクラスタ2,3からリンクを指されているという特徴を持つノードの集合である．つまり，ノードは直接的な接続関係からではなく，ネットワーク全体における他のクラス

†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

	1	2	3	4
1	0	1	0	0
2	1	1	1	0
3	1	1	0	0
4	0	1	1	0

表 1: 縮約された行列

タとの関係の類似度によってクラスタに分類される．以上からWebにおいても同様に，サイト間に直接リンクがなくても，その他のサイトとの関係において結合パターンが等しければ，サイトの内容も類似していると考えられる．

CONCORは隣接行列の行ごとの相関をピアソンの積率母関数よって計算する．次に，隣接行列の相関を入力として同様に相関の相関を求める．このプロセスを繰り返すと行列のすべての成分が $+1$ と -1 に収束する．以上から，全体を相関値が $+1$ と -1 の部分集合の二つに分割することができる．前の分割によってできた部分集合に対して繰り返し適用することで，より細かく分割することができる．分割プロセスは図1のように木構造になる．

4. 提案手法

CONCORは2分割を繰り返す性質上，一度別のクラスタに分割されると再び同じクラスタになることはない．そこで，分割によって形成されたクラスタ間の関係を分析することで，それぞれのサイトに対して高い関連性を示すクラスタを見つけるアルゴリズムを提案する．

まず，2つのクラスタとその他すべてのクラスタとの関係の差異に注目する．1つのクラスタが2つの部分集合に分割されるためには，分割後の2つのクラスタ間で他のすべてのクラスタとの隣接関係に違いが必ず存在する．隣接関係がすべて同じ等しければ，CONCORではそれ以上分割されることはないからである．そこで，分割によって形成された2つのクラスタと他のすべてのクラスタとの隣接関係において，一方のクラスタのみとリンクを有するクラスタを探し出す．この一方のみにリンクを張っているクラスタこそがこの分割，そしてそれぞれのクラスタを特徴付けている．なぜなら，このリンクが存在しなければこの分割は行われず，1つのクラスタのままであったが，一方のみにリンクを張るクラスタが存在することで隣接関係に違いが生じ分割が行われる．このクラスタが他のすべてのクラスタと違う固有の存在であることは，このリンクの存在によってのみ保証されるからである．そこで，ひとつのクラスタに対して，その位置を特徴付けるリンクを有するクラスタはネットワーク上の位置と同様にWeb上の文書の内容においても，そのクラスタの文書の内容に対して特徴付ける存在であると考え，「類似サイト」と呼ぶ．

たとえば，クラスタ1とクラスタ2について比較を行う．表1よりクラスタ1はクラスタ2,3からリンクが張られ，またクラスタ2はクラスタ1,2,3,4からリンクが張られていることが分かる．クラスタ1,2ともにクラスタ2,3からリンクが張られていることは共通している．一方で，クラスタ1,4はクラスタ2に対してのみリンクを張っている．クラスタ2にノード2,5,7が属する原因は，ノード2,5,7がクリークを形成しているからではなく，他のノードからの被リンク数が等しいからでもない．クラスタ1との関係において違いが存在するからである．つまり，クラスタ2を特徴付けているのはクラスタ1との差異であるクラスタ1,4からのリンクである．逆に，クラスタ1,4からのリンクが存在しないことがクラスタ1を特徴付けている．また，クラスタ2,3はクラスタ1,2の両方に対してリンクを張っている．これらのリンクはクラスタを形成するにあたって差異を生み出さず分割の原因とならないため，クラスタを特徴付けるリンクとは考えない．そして，クラスタ2を特徴付けるリンクに注目し，そのリンクを有するクラスタ1,4はクラスタ2と関連性が高い「類似サイト」であると考える．また，クラスタ1,2が他のクラスタに対して張っているリンクに注目すると，クラスタ2に対してクラスタ1,3が「類似サイト」であることも分かる．

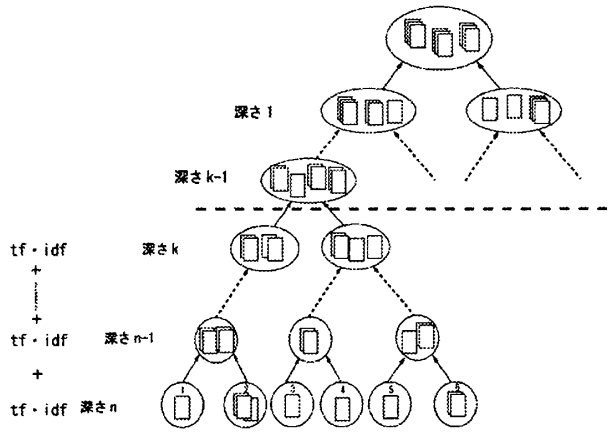

図 1: CONCORによる分割プロセス

そして，個々のサイトの文書と同一クラスタに分類されたサイトの文書間で共通する名詞のTF·IDFの値を大きくすることに加え，そのクラスタに対応する「類似サイト」の文書との間で共通に出現する名詞を探し，その名詞のTF·IDFの値を大きくする．「類似サイト」と比較することによって，単一文書内での出現頻度の偏りにとらわれない名詞の重み付けを実現し，精度の向上を図る．比較対象をグラフから得た結果，重み付けを補正する式は以下の通りである．

$tf \cdot idf$[要約するサイト] $= tf \cdot idf + \alpha(tf \cdot idf$[同一クラスタ内のサイト] $+ tf \cdot idf$[類似サイト]) $(0 < \alpha < 1)$

5. 実験

5.1 データ収集

検索エンジンにキーワードを入力し検索結果上位100までのサイトのURLを得る．これらのURLを入力としてプログラムを実行することで100サイト間のリンク構造とこれらの100サイトから3回以上リンクを張られているサイトのリンク関係を得る．得られたリンク構造をUCINET[Borgatti 02]を用いてCONCORを実行す

る．そして，その出力に対して提案手法を実行するプログラムを用いてすべてのサイトについて「類似サイト」を特定する．データに関する詳細は以下の通りである．

- 検索語　郵政＆民営化
- サイト数　452

以上のデータに対してCONCORにより15回分割を行い，135のクラスタが得られた．

5.2　得られたクラスタ

以下ではRIETI経済産業研究所の「郵政民営化の論点」§の要約例について考察を行った．

このサイトは郵政民営化の論点をまとめている．本手法によってこの文書と関連があると特定したサイトは15あり，図2に代表的なサイトを示す．この15サイトのうち，要約対象となる文書と内容において最も関連性が高い郵政民営化の論点をまとめているサイトが7，法案に反対の立場を取る労働組合のサイトが1つ含まれていた．その他，要約される原文章が属しているサイトに含まれているページが2，読売新聞関連のサイトが3，楽天市場のメインページ，2chのメインページが含まれていた．読売新聞，楽天市場，2chのサイトはリンク構造を示したグラフにおいて葉の位置を占めているため，他のサイトと多くのリンクを持つサイトと比べて，他との関係を反映されていないことが原因の一つであると考えられる．

もう一つの例として「郵政民営化監視市民ネット」¶に対してもリンク解析による結果の分析を行った．このサイトは郵政民営化法案に対して否定的な意見を述べているサイトである．このサイトのと関連付けられたサイトは4つあり，そのうち3つは同様に民営化に対して反対の立場のサイトであった．

5.3　得られた要約

分割アルゴリズムの結果の検証でも取り上げたRIETI経済産業研究所の「郵政民営化の論点」の文書を要約率25%で要約を行った．本手法により変化させた名詞の重み付けと単一文書のみの名詞の頻度を用いた名詞の重み付けとをそれぞれ入力としてMMR[Carbonell 98]によって作成された文書を比較した．

本手法によって関連するページと共通する単語の重み付けを増加させた結果，「郵便」「民営」「事業」「改革」「公社」「市場」などの名詞の重みが増加した．サイトごとの重み付けの変化の関与は図2に個々のサイトごとに示している．またそれぞれの名詞の重みの変化を図3に示す．横軸は単一文書内での単語の重み付けの結果重みが大きな順に並べたものであり，縦軸は重みの値である．単一文書内では頻度が低く重み付けが小さい単語でも，類似サイトと比べ共通する名詞の重み付けを大きくすることでキーワードとなるべき単語の重みを大きくすることができている．

次に要約文についての比較を行う．要約対象となる原文章は前半が郵政民営化の4つの論点を取り上げ，後半は小泉首相の私的懇談会「郵政三事業の在り方について

§http://www.rieti.go.jp/jp/columns/a01_0126.html
¶http://www.mm-m.ne.jp/dave/declaration/qanda.htm

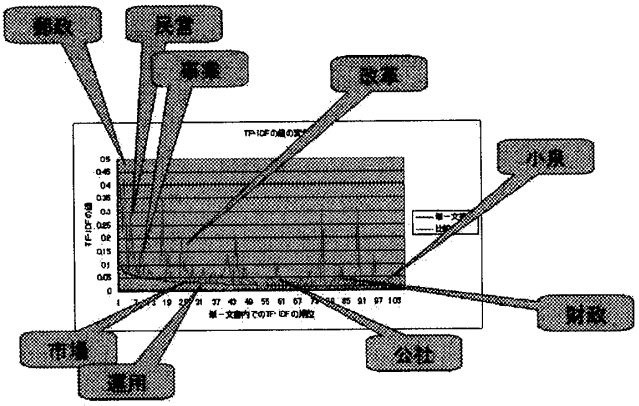

図3: 名詞の重み付けの変化

考える懇談会」がまとめた3つの案について説明を行っている．二つの異なる名詞の重み付けを入力とし作成された要約文の大きな違いは後半冒頭，

- 平成14年9月に小泉首相の私的懇談会「郵政三事業の在り方について考える懇親会」（首相官邸）が，3つの民営化案をまとめた．
- まず，1)は，郵政三事業を一体として特殊会社とし，その会社の株を政府が保有する，というものだ．

という文章が単一文章の名詞の頻度を基に名詞を重み付けている場合，要約文には含まれず，本手法では要約文に残された点である．二つの文章において本手法を用いた重み付けの変化によって値が大きくなった名詞を強調した．「次に，2）は...」と後に続く文書の話題の転換点であり，この文章が要約文になければ，前半の郵政民営化の4つの論点と私的懇談会がまとめた3つの案の違いが要約文を読んだだけでは分からず，重要な文である．このようにWebのリンク構造を利用することが有効であることが分かった．

6.　まとめ

Webのリンク構造をCONCORにより分析し，クラスタ化された関係と実際のリンク関係の差を利用して「類似サイト」を特定し，文書間で共通に出現する名詞のTF·IDFの値を増加させることで，重要な単語が抽出できた．また，これによりWebページを要約すると従来の方法と比べ，よりよい要約文を作成することができた．

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．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，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．

参考文献

[Girvan 02] Girvan, M. and Newman, M. E. J. *Community structure in social and biological networks.*

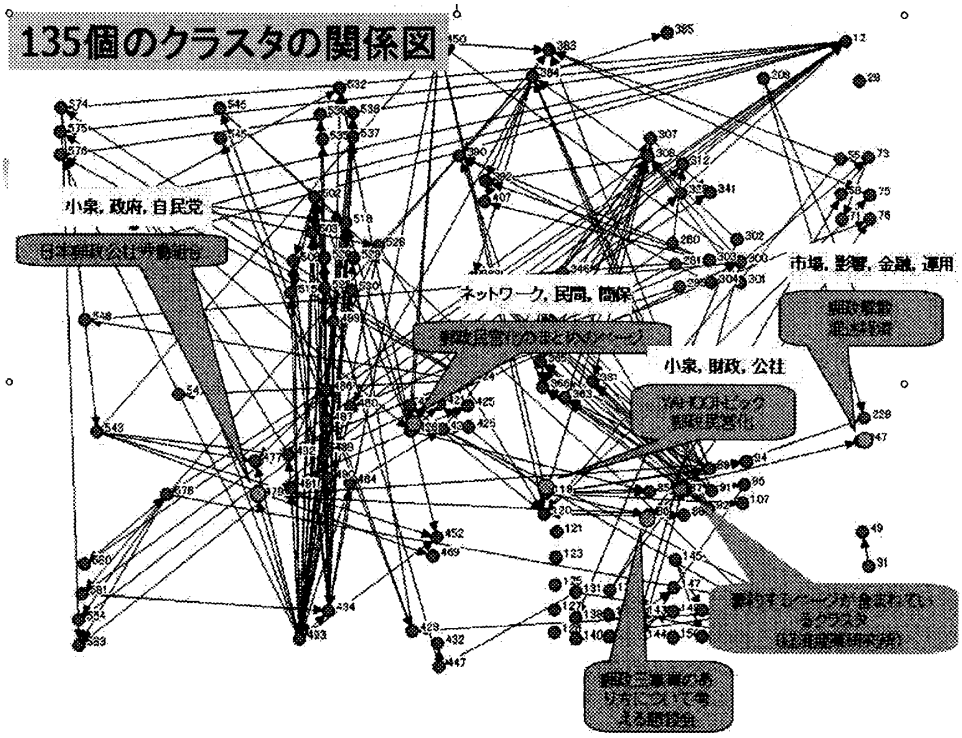

図 2: 結果の例
CONCOR によって 15 回分割を行った結果できた 135 のクラスタの関係を示している.

Proceedings of the National Academy of Sciences of the United States of America (PNAS), 99(12):7821–7826. 2002.

[Palla 05] Gergely Palla, Imre Derenyi, Illes Farkas,Tamas Vicsek. *Uncovering the overlapping community structure of complex networks in nature and society.* Nature 435, 814-818,2005.

[Everett 98] Everett, M. G. ,Borgatti.S. P. *Analyzing clique overlap.* CONNECTION 21(1):49-61,1998.

[Flake 02] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. *Self-organization of the web and identification of communities* . IEEE Computer, 35(3):66–71, 2002.

[Kumar 99] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan and Andrew Tomkins. *Trawling the web for emerging cyber-communities.* WWW8 / Computer Networks, Vol 31, p1481-1493, 1999.

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

[Border 00] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J. *Graph structure in the web* Proc. of the WWW9 Conference (2000) 309–320.

[Barabasi 99] Albert-Laszlo Barabasi and Reka Albert. *Emergence of Scaling in Random Networks.* Science, 8, October 1999.

[Watts 98] D. J. Watts and S. H. Strogatz. *Collective dynamics of 'smallworld ' networks* In Nature, vol. 393, pp. 440-442, 1998.

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

[Borgatti 02] Borgatti, Everett, and Freeman. *UCINET* Analytic Technologies,Inc 2002.

[Carbonell 98] Jaime Carbonell,Jade Goldstein. *The use of MMR, diversity-based reranking for reordering documents and producing summaries.* In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.