社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ICS－144 (1)
2006／7／28

CONCORによるリンク解析を用いたWeb文書からの重要語抽出

山下 長義 †, 福井 健一 ††, 森山 甲一 ††, 沼尾 正行 ††, 栗原 聡 ††

† 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

本論文ではWebにおいて他のサイトとのリンクパターンが等しいサイトは内容的にも等しいのではないかと考え，リンクパターンが等しいサイト間で比較を行うことでWebページから重要語抽出する手法を提案する．まず，CONCORでサイト間のリンク関係を表すネットワークをクラスタに分割し，提案手法によってそれぞれのサイトに対する類似サイトを特定する．そして，CONCORにより分割された同一クラスタ内のサイト間に共通して出現する名詞とそれぞれのサイトとそれらに対する類似サイト間に共通して出現する名詞の重み付けを補正する．リンク構造を言語処理に反映することで重要語抽出をおこない，従来手法よりよい結果が得られた．

Web Documents Summarization using Link Analysis Based on CONCOR

Nagayoshi Yamashita † Kenichi Fukui †† Koichi Moriyama †† Masayuki Numao †† Satoshi Kurihara ††

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

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

In this paper, we propose a framework to extracting significant words in the Web using link structure. We use the global method in social network 'CONCOR' for link analysis. This is based on the assumption that if the link patterns of two sites and links are the same, then these two sites also contain the same in contents. In the first phase, the whole network consisting of sites are divided into clusters using CONCOR. Subsequently, by using the method we propose, we identify similarity sites. Comparing a site with other sites in the same cluster and with the similarity sites for the site, we assign higher weights to nouns that exist in two sites in common. By using link analysis to language processing, we could discover significant words.

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というのはWebにおいても同様である．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，言語処理に加えてリンク構造を用いる手法を提案する．まず，あるキーワードに基づいて収集したサイト間のリンク構造を用いて，互いに類似するサイトを特定するためにブロックモデルによるリンク解析法を用いてマクロな視点からクラスタリングを行い，今回提案するアルゴリズムにより重要単語を抽出する．そして，評価のためにこれらの重み付けをされた名詞を入力としてWeb文書の要約を行った．

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をネットワークグラフ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[1]を利用する手法やクリーク[2,3]を利用する手法，グラフ理論の最大流最小カット定理[4]，2部グラフ[5]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[6]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[7]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向上とインターネットの高速化により，誰もが大規模なデータにアクセスかつ情報を発信できるようになったことが要因となり，WWWがスケールフリー性[9]とスモールワールド性[8]を有していることが明らかにされている．

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であるブロックモデルの一つのクラスタリング手法であるCONCOR[10]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の条件として，中心性やクリーク解析のようにサイト間が直接つながっている必要がない．リンクパターンが同じサイトが同一のクラスタに分類される．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は構造同値と呼ばれる概念を利用するネットワーク解析手法である．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．図1において，ノードCとノードDは直接接続していないが，ノードAとノードBに対する関係が同じであるため構造同値では同じ位置を占める．たとえば，医者や患者に対する関係の類似度から異なった病院の看護婦も看護婦としての地位を占めるが，それぞれの看護婦はお互いを知らないし，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のではないかと考えられる．

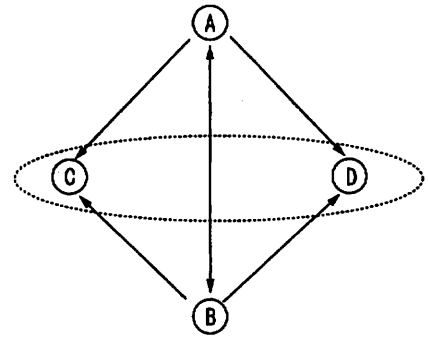

Fig. 1 構造同値の例

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2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，この点に注目し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リンク構造を用いて重要性の高い単語を抽出するアルゴリズムを提案する．

まず，CONCORによって一度の分割で形成される2つのクラスタについて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たからである．

たとえば，図3において，ノードAとノードBはクラスタ3とクラスタ4との関係は同じであるが，ノードAはクラスタ1からリンクを張られ，一方でノードBはクラスタ2に対してリンクを張っている．図3における点線で示したリンクだけが存在していれば，ノードAとノードBは同一のクラスタに分類されるが，実線で示されたリンクが存在するため別々のクラスタに分類される．よって，この実線で示されたリンクがノードAとノードBが属するクラスタの性質を決定していると考えられる．

そして，そのクラスタを特徴付けているリンクを有するクラスタはWeb上の文書内容においても同様にクラスタを特徴付けているものであると考え，「類似サイト」と呼ぶ．つまり，このリンクによって結合されたクラスタ内と共通する単語は重要性が高いと判断される．そこで，同一クラスタ内のノード同士とこのリンク先のクラスタと共通する名詞を見つけ，それらの名詞の重み付けを大きくする．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]$)$ $(0 < \alpha < 1)$

また，提案手法のアルゴリズムは図4のフローチャートの通りである．

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[11]を用いてCONCORを実行する．そして，その出力に

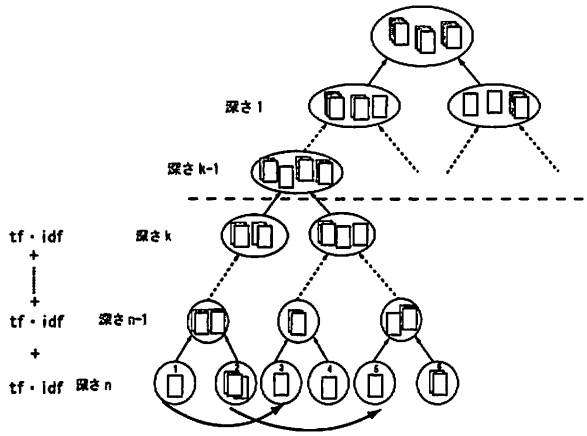

Fig. 2 CONCORによる分割プロセス

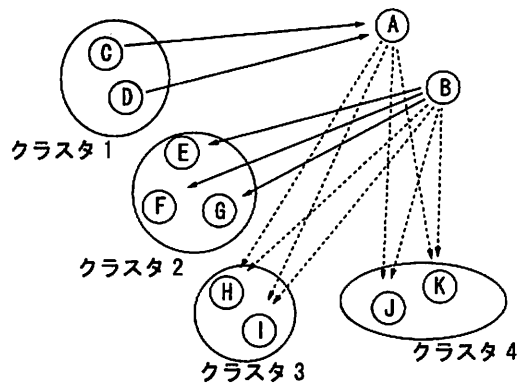

Fig. 3 提案手法の例

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重要単語の抽出を行う．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

RIETI経済産業研究所の「郵政民営化の論点」[1]の要約例について考察を行った．

[1] http://www.rieti.go.jp/jp/columns/a01_0126.html

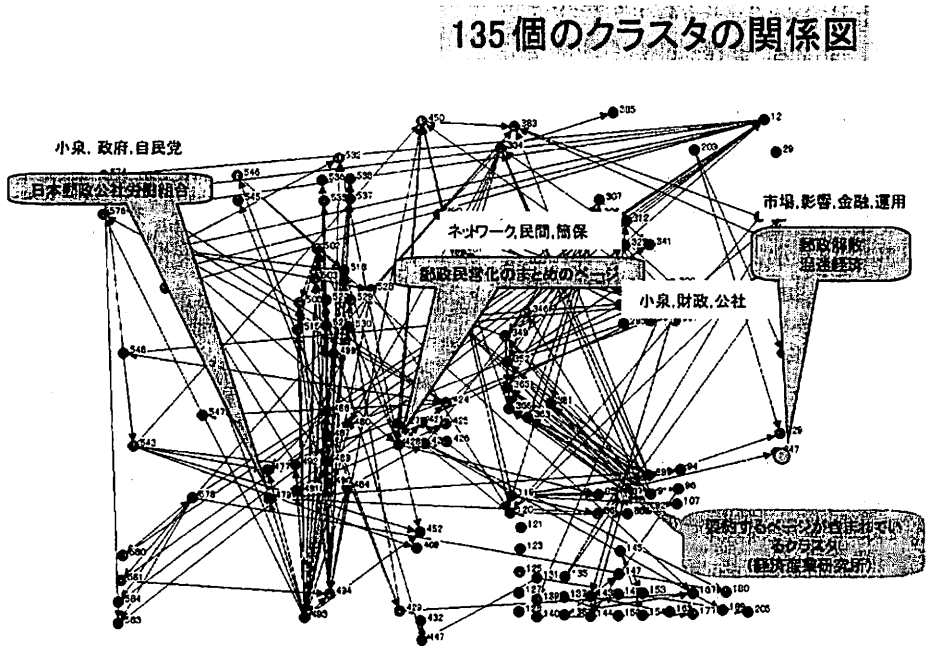

Fig. 5　結果の例

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 5 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」[2]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．本手法により抽出した重要単語と単一文書のみの名詞の頻度による重み付けそれぞれの入力として用い，MMR[12] によって要約を行い比較した．

本手法によって関連するページと共通する重要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 5 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 6 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

[2] http://www.mm-m.ne.jp/dave/declaration/qanda.htm

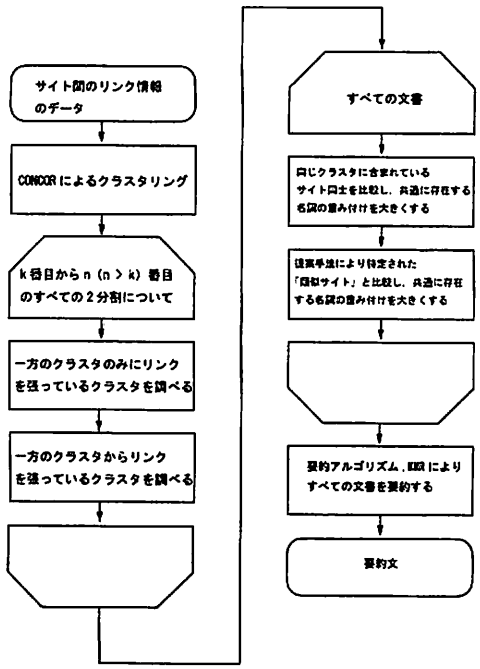

Fig. 4　提案手法のフローチャート

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．

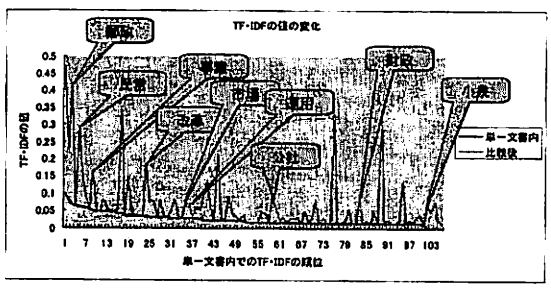

Fig. 6　名詞の重み付けの変化

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重要単語を抽出した，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，単語の重み付けの比較方法の更なる検討と他のブロックモデルとの比較が必要である．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

1) Girvan, M. and Newman, M. E. J. *Community structure in social and biological networks.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), 99(12):7821–7826. 2002.

2) 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.

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

4) 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.

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

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

7) 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.

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

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

10) Stanley Wasserman, Katherine Faust. *Social Network Analysis* Cambridge university press,1994

11) Borgatti, Everett, and Freeman.*UCINET* Analytic Technologies,Inc 2002.

12) 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.