3A2-2

ブロックモデルによるリンク解析を用いた複数文書の要約

Multiple Documents Summarization using Link Analysis Based on Blockmodel

山下　長義 *1　森山　甲一 *2　栗原　聡 *2　沼尾　正行*2
Nagayoshi Yamashita　Koichi Moriyama　Satoshi Kurihara　Masayuki Numao

*1大阪大学大学院情報科学研究科情報数理学専攻
Department of Information and Physical Science, Graduate School of Information Science and Technology, Osaka University

*2大阪大学産業科学研究科
The Institute of Science and Industrial Research, Osaka University

WWW is an interactive media that utilizes large amount of information. Consequently, many researcher related to Web mining have been emerging. Although there are documents structured by HTML, natural language is so vague that it is difficult to analyze correctly. This is why our reserch proposes the methodology of using link structures to summarize documents on the Web. By verifying the correlation between contents and link structures, we discuss the merits of our method.

1. はじめに

WWW 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，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，本研究では Web 上の文書を要約すためにテキスト情報と Web 空間のリンク構造の両方を利用する手法を提案する．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類を行い，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，その結果を文書個々の要約に反映させることで Web 文書の要約を行う．最後に Web 上の内容とリンク構造の相関性を検証し，本研究の手法の有用性を考察する．

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．スモールワールドネットワークはクラスタ係数が高く，ランダムグラフのように頂点間の距離の平均が小さいという性質がある．また，スケールフリーネットワークは WWW から科学論文の共著関係まで多岐にわたる複雑ネットワークにおいて，ネットワーク内のあるひとつの頂点が他の k 個の頂点とつながっている確率 P(k) がべき法則 $P(k) \sim k^{-\tau}$ に従い減少する性質を持つ．スケールフリー性は，Web ページの参照パターンが全体から部分まで同じ構造であることを意味し，スモールワールド性は任意の知識が少数の知識を媒介として関係付けられていることを意味する．

本研究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手法がリンク構造の解析に有効であるかどうかを検証し，解析結果を Web 文書の要約に応用する．

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，互いに類似するサイトを抽出する．たとえば，以下のような 1 か

連絡先: 山下長義
大阪大学産業科学研究所，大阪府茨木市美穂ヶ丘 8-1
電話番号:06-6879-8426，Fax 番号 06-6879-8428
nagayosi@ai.sanken.osaka-u.ac.jp

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
1	−	0	0	1	0	0	0	0	1
2	0	−	0	0	1	0	1	0	0
3	0	1	−	0	1	1	1	1	0
4	1	0	0	−	0	0	0	0	1
5	0	1	0	0	−	0	1	0	0
6	0	1	1	0	1	−	1	1	0
7	0	1	0	0	1	0	−	0	0
8	0	1	1	0	1	1	1	−	0
9	1	0	0	1	0	0	0	0	−

を構造同値の概念を基に分割すると，

	6	3	8	4	1	9	2	5	7
6	−	1	1	0	0	0	1	1	1
3	1	−	1	0	0	0	1	1	1
8	1	1	−	0	0	0	1	1	1
4	0	0	0	−	1	1	0	0	0
1	0	0	0	−	1	1	0	0	0
9	0	0	0	1	1	−	0	0	0
2	0	0	0	0	0	0	−	1	1
5	0	0	0	0	0	0	−	1	1
7	0	0	0	0	0	0	1	1	−

となり，部分行列間の接続パターンによって分類されることが分かる．構造同値では任意の二つのノードが他のすべてのノードとの結合パターンが同じであれば 2 つノードが同一のクラスタに分類される．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，同じ医者について，同じ患者を診ているわけではない．類似する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同様ならば，サイトの内容も類似しているのではないかということである．

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，CONCOR でできるだけ分割を繰り返し，クラスタ間で高い関連性を示すクラスタ群を見つけるアルゴリズムを提案する．

はじめに，CONCOR により分割を任意の回数繰り返す．次にクラスタに属するサイトのリンク関係に注目する．一つのクラスタが 2 つの部分集合に分割されるためには，そのクラスタに属するサイトの結合パターンが違うことが必要である．そこで，ぞれぞれの分割において分割を特徴付けたリンクに注目し，そのリンクを有するサイトはクラスタに属するサイトと関連が高い「類似サイト」であると考える．つまり，分割によってできた 2 つのクラスタを比べ，一方のみにリンクを有するサイトもしくは，一方のクラスタにのみリンクを有するサイトを探す．そのサイトはクラスタを形成するにあたって直接影響を与えている可能性が高いと推測できる．クラスタ間の接続関係が同じならば，それ以上分割されることはないからである．

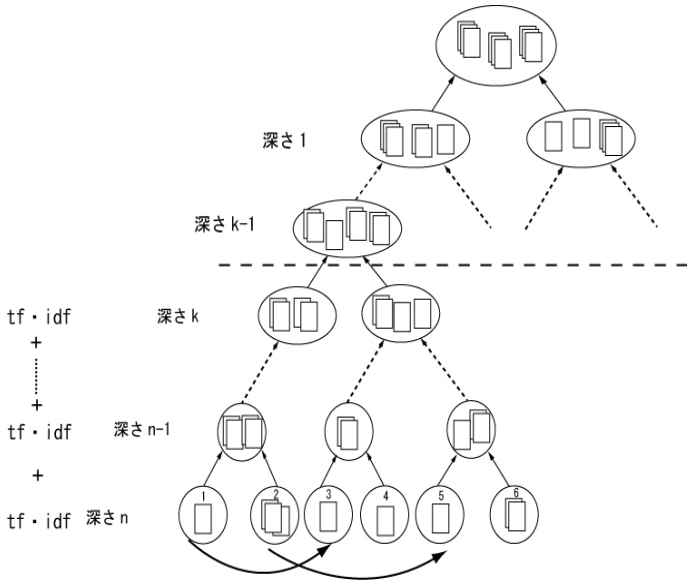

図 1: CONCOR による分割プロセス

たとえば，図 1 のように CONCOR により分割を n 回繰り返しクラスタに分割する．そこで，分割プロセスの木構造の図の葉の部分から順にクラスタ間に存在する実際のリンク関係の差異に注目する．図 1 の下部の矢印はクラスタ内のサイトから他のクラスタ内のサイトへ実際に存在しているリンクを表している．クラスタ 1 はクラスタ 3 に対してリンクを張っていて，クラスタ 2 はクラスタ 5 に対してリンクを張っている．つまり，クラスタ 1 とクラスタ 3，そして 2 と 5 は CONCOR では異なるクラスタに分類されるものの，内容的に関連している可能性があるということである．さらに，分割プロセスを表す図 1 の木構造で，深さ k から n までのすべての分割についてリンク関係の差を調べ,「類似サイト」を特定する．そして，個々のサイトの文書とそれに対応する「類似サイト」の文書との間で共通に出現する名詞を探し，その名詞の重み付けを大きくする．そして，MMR という要約アルゴリズム [Carbonell 98] を用いて要約を行う．

5. 実験

5.1 データ収集方法

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 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．

- 収集日時 2005 年 12 月 31 日
- 検索語　郵政＆民営化
- サイト数　452

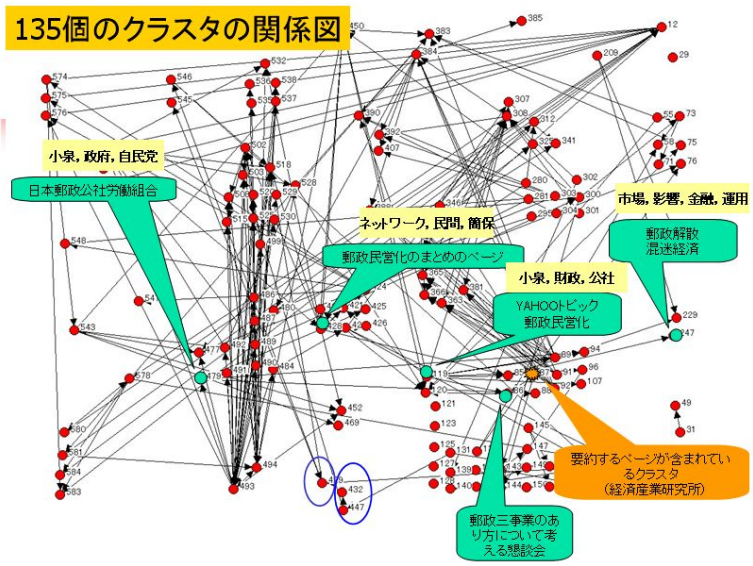

図 2: 結果の例
CONCOR によって 15 回分割を行った結果できた 135 のクラスタの関係を示している.

5.2 得られたクラスタ

要約される対象となる文書 RIETI 経済産業研究所の「郵政民営化の論点」[*1]を例に考察を行った.

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] という要約アルゴリズム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,

*1 http://www.rieti.go.jp/jp/columns/a01_0126.html
*2 http://www.mm-m.ne.jp/dave/declaration/qanda.htm

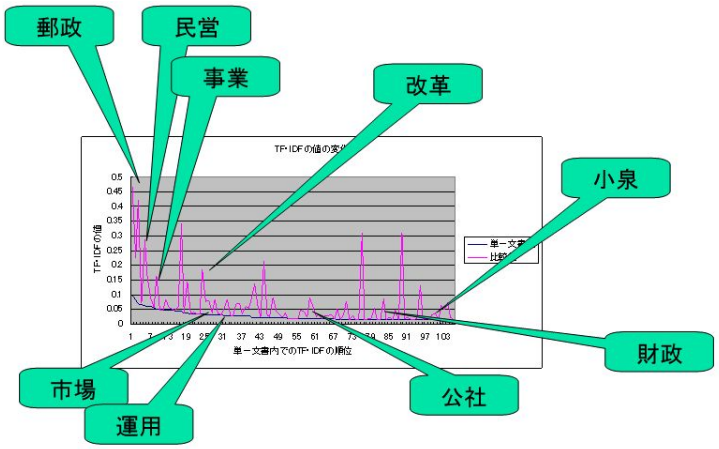

図 3: 名詞の重み付けの変化

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.「次に，2）は...」と後に続く文書の話題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．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．このように本システムによって変化した名詞の重み付けが精度のよい要約文作成に重要な役割を果たしていることが分かった.

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な単語を抽出し，これにより Web ページを要約すると従来の方法と比べ精度の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．またページ間の比較するとき単純に共通している単語の重み付けを変えたが，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.* 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. *Stanley Wasserman, Katherine Faust.* 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.