Webリンク構造解析と自然言語処理による組織関係の抽出についての研究

池辺正典[†] 田中成典[††] 古田均[††]
中村健二[†] 小林建太[††]

近年のインターネットの複雑化にともない，Webの自動解析による情報取得に対する需要が高まっている．そのため，Webページをカテゴリに分類する手法やWebの関係情報を解析する手法が数多く提案されてきた．しかし，既存の研究では，Webの自動解析は，リンク関係を中心とした解析を行っており，リンク関係のないWebページを関連付けることが困難であった．このため，本論文では，リンク構造解析だけでなく，形態素解析によって任意の単語から関係情報の抽出を行うことで，リンク関係のないWebページを関連付ける．また，その結果と品質判定を行ったリンク構造解析結果を組み合わせることで，信頼性の高いWebページの関係図を作成する．さらに，アルゴリズムの評価として，Webから取得した情報を利用して，組織の関係図を作成する．そして，既存研究においての主要な方式であるリンク構造解析による結果との比較を行った．評価方式には，リンク構造解析で一般的に用いられている評価値とグラフ理論による可視化を採用し，その結果から本方式の有用性を確認した．

Research for Extracting of Organization Relationship Diagram by Web Link Structure Analysis and Natural Language Processing

Masanori Ikebe,[†] Shigenori Tanaka,[††] Hitoshi Furuta,[††]
Kenji Nakamura[†] and Kenta Kobayashi[††]

The demand for the information acquisition by the automatic analysis of the Web has been increased as the Internet recently becomes complicated. Then, a method of classifying the Web into some categories and analyzing the Web relationships with the information were suggested. In early researches, however, it was difficult to relate the Web pages without the links because the analysis was mainly done on the link relationships. Therefore, on this research, we related the Web pages without the link relationships not only by analyzing the link structure but also by extracting any kinds of topics using morphological analysis. And, we drew a high reliable relationship diagram combining our processing consequence with the quality judged result of the link analysis. In addition, we drew the organization charts using the acquired information from the Web in order to evaluate our algorithm. Then, we compared the link structure analysis which was the main method of the early researches with our result. We chose the evaluate value which is generally utilized for the link structure analysis and the visualization by graph theory as the evaluation method. We confirmed that our method was available.

1. はじめに

近年のインターネットの普及において，Webページは増加の一途をたどり，その内容は，多様化，複雑化するという傾向を見せている．Webの煩雑化にともない，ユーザが望む情報を取得するためには，検索エンジン等のWebページの自動解析ツールを利用し，対話的に情報を探す[1),2)]必要がある．現在のWebページの自動解析では，Webページを同じ目的や関心事においてグループ化し，その内容に応じて，任意のカテゴリに分類することで，Webページの各集合をWebコミュニティとして扱うという考え方が普及しつつある．

そして，Webコミュニティの発見，分類は，リンク構造解析に主眼を置いた手法が主流である．しかし，リンク構造解析から得たWebページの関係情報は，Webの構造を具現化しただけであり，分類するカテゴリの種類によっては，現実社会の関係が反映されない．このため，組織関係をWebページから抽出する

† 関西大学大学院総合情報学研究科
Graduate School of Informatics, Kansai University
†† 関西大学総合情報学部
Faculty of Informatics, Kansai University

場合，インターネットを主体として活動する企業は過大評価された状態でユーザに提示される．一方，Web ページが未公開もしくは有名でない企業は，過少評価され現実社会の規模とは異なった印象をユーザに与える．さらに，リンク関係のない組織間の情報を抽出することは困難で，現実社会では関係があったとしても，それが検出されることはない．

現在，ユーザが Web を利用するにあたり，最も身近となる Web ページの自動解析ツールが検索エンジンである．しかし，各企業が過剰な SEO（Search Engine Optimization）を行うことによって，インターネット上においての評価を過剰に上げている．このことにより，現実社会とは異なった評価をユーザに提示することが容易である．この問題は，リンク構造解析における代表的な問題点である．このため，Web ページから関係情報を取得する場合は，リンク構造以外の情報を利用し，Web ページどうしを関係付けることが必要[3] となる．

そこで，本研究では，リンク構造解析以外の手法により，Web ページの関係性を抽出することで，従来手法よりも信頼性の高い組織関係の抽出を目指す．

2. 研究の概要

Web 自動解析の既存研究は，リンク構造を中心とした研究[4] と，自然言語処理を利用した研究[5] の 2 種類に分類される．この中で，リンク構造の解析は，処理負荷が小さいために，Web ページを大まかなカテゴリに分類するといった大規模な Web コミュニティの解析[6] に適している．また，自然言語処理は，詳細な解析により，比較的高い精度の情報抽出が可能という点から小規模な Web コミュニティの解析に適している．本研究では，これらの 2 つの方式を組み合わせることで，互いの問題を解消し，処理負荷が小さく，精度の高い情報解析方式を考案した．さらに，大規模な Web コミュニティの解析において，従来手法では，カテゴリの種類によって，現実社会での関係情報を抽出することが困難であったが，本提案手法では，資本関係，業務提携関係，取引関係といった現実社会における組織間の関係情報を補完することを目的とする．

自然言語処理において，Web ページの内容を解析する場合に注意すべきことは，内容の信頼性を判定する必要があるという点である．Web は，だれもが情報を発信できるという特徴がある．そして，個人発信の情報と企業発信の情報を比較した場合，企業発信の情報の方が高い信頼性を保持している．しかし，企業発信の情報においても，すべてが信頼できるわけではない．そこで，本研究では，この問題を解消するために，Web 構造解析の対象とする情報を，企業発信の情報に限定する．さらに，企業の中でも，検索エンジンの検索結果から主要な企業を抽出し，Web 探索の基点とすることにより，取得する情報の信頼性を確保する．以上の問題点をふまえ，Web ページから組織の関係を抽出するために，本提案手法では，**図 1** に示す処理の流れで関係情報の取得を行う．

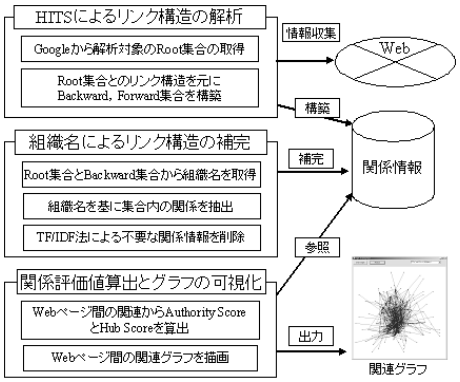


図 1 処理の流れ
Fig. 1 Flow of process.

処理の手順としては，最初に，HITS（Hyperlink-Induced Topic Search）アルゴリズムによるリンク構造解析の基準となる Root 集合として，企業の Web ページ集合を取得する．そして，Root 集合からリンク構造解析を行うことで，被リンクおよびリンクしている Web ページの集合を Backward 集合，Forward 集合として取得し，基準となる集合に追加する．次に，自然言語処理により，組織名を利用した関係情報の補完として，得られた Web ページの集合に対して，形態素解析と組織名辞書による組織名の抽出を行う．そして，抽出した組織名の重要度を個別に計算することで，関係の強度を判断し，弱い関係情報を削除することで有用な関係情報のみを追加の関係情報として補完する．最後に，関係情報の評価として，評価値の算出を行い，評価値を基にして，関係情報および Web ページをグラフとして可視化する．

本論文は，3 章において，従来手法におけるリンク構造解析とその問題点について解説し，4 章において，自然言語処理においての関係情報の抽出について解説する．そして，5 章において本提案手法の精度を検証するために，実証実験を行い，6 章で本研究により得られる成果と今後の展望についての考察を述べる．

3. リンク構造解析による組織関係の取得

3.1 HITS アルゴリズム

HITS アルゴリズム[7] は，Kleinberg によって提案された概念であり，Web ページ間の関係をリンク構造解析によって抽出する手法である．現在では，HITS アルゴリズムを改良したリンク構造解析手法が数多く提案されている．本提案手法においても，リンク構造解析は，HITS アルゴリズムを基本とした改良手法を利用する．

HITS アルゴリズムでは，Authority Score と Hub Score という 2 種類の評価値から解析ページの有用性を取得するという特徴を持つ．Authority Score は，分類するカテゴリに対して，有益な情報が多く含まれていることを示し，Hub Score は，Authority Score の高いページへのリンクが豊富なことを示す．Authority Score および Hub Score は，次式によって得ることができる．

$$Auth(p) = \sum Hub(q) \tag{1}$$

$$Hub(p) = \sum Auth(q) \tag{2}$$

式 (1) と式 (2) において，q は評価値を計算する各 Web ページを示し，p は，評価値の計算が行われた Web ページを示す．Authority Score と Hub Score は，互いに導出関係があるために，この 2 式を反復計算し，2 乗和を 1 に補正することで，評価値は一定の値に収束する．

HITS アルゴリズムの処理方式としては，最初に，任意のカテゴリに沿った Web ページ集合を Web 探索の基点となるページとして収集し，Root 集合 $R(x_1)$ として定義する．次に，$R(x_1)$ の各 Web ページへのリンクを含む Web ページを Backward 集合 $B(x_2)$ とし，$R(x_1)$ の各 Web ページからリンクをしている Web ページを Forward 集合 $F(x_3)$ とする．そして，式 (3) に示すように，これら集合をあわせたものを評価値の計算の基準となる Web ページ集合 $S(x)$ として定義する．

$$S(x) = R(x_1) \cup B(x_2) \cup F(x_3) \tag{3}$$

最後に，評価値の計算では，Authority Score と Hub Score を取得するために，$S(x)$ のリンク構造からグラフ $G(x)$ を作成する．グラフ $G(x)$ の作成では，ミラーページや同一ホストのリンクを削除し，同一ホスト内での Authority Score の過剰な上昇を抑制する．そして，$G(x)$ における Web ページ間の n 次の隣接行列 A と，その転置行列 A^T を作成し，式 (1) を適用することで反復計算を行い，各 Web ページの Authority Score と Hub Score を取得する．

3.2 HITS アルゴリズムの問題点

HITS アルゴリズムでは，Web ページの探索時に，Backward 集合と Forward 集合を取得した．しかし，Forward 集合は，特定の主要な Web ページからのリンクが含まれる可能性が高く，Forward 集合の一部に対して，Authority Score が局所的に集中する Topic Drift という問題がある．本提案手法では，この問題を解消するために，Forward 集合を解析対象から排除する．このため，Authority Score と Hub Score 算出の基準となる Web ページの集合 $S'(x)$ を次式のように考える．

$$S'(x) = R(x_1) \cup B(x_2) \tag{4}$$

式 (4) の問題点としては，基準となる Web ページが減少するために，関係情報が不足し，グラフの作成が困難になる可能性があげられる．この問題を解決する手法としては，Backward 集合の探索範囲を広くするといったリンク構造解析の範囲を広げる方法があげられる．しかし，リンク関係の過剰な取得には，つねに Topic Drift がつきまとうという問題がある．このため，基準となる Web ページの抽出には，リンク関係以外の関係情報の抽出手法が必要である．本提案手法においては，関係情報の不足を解消するために，形態素解析と組織名のマッチングを利用して，リンク関係のない組織の Web ページ間の関係情報を抽出する．

4. 自然言語処理による組織関係の取得

4.1 Web ページの関係情報の補完

本提案手法では，HITS アルゴリズムの改良として，Topic Drift の抑制のために，Forward 集合の取得を行わないことは前述のとおりであるが，これでは，関係情報が不足し，満足な関係図が作成できないという問題がある．このため，Root 集合および Backward 集合から組織名を抽出し，各 Web ページのタイトルとマッチングを行うことで関係情報の補完を行う．関係情報の抽出は，以下の手順で行う．

(1) 式 (4) の集合 $S'(x)$ の各 Web ページに対して，HTML ソースからタグ箇所を排除し，タイトル，リンク文字列と文書部分を抽出する．

(2) 上記の文書部分に対して，形態素解析を行い，抽出した名詞を組織名辞書と比較し，一致した名詞を組織名として抽出する．そして，抽出した組織名の集合を $T_1(xy_n)$ その個数を $C(xy_n)$ とする．また，y は抽出した組織名を示す．

(3) 組織名と関連付ける Web ページの候補集合と

して，新たに組織の Web ページ集合を取得し，これと $S'(x)$ の和を $S_{add}(x)$ とする．

(4) $S_{add}(x)$ の各 Web ページのタイトルに対して，形態素解析を行い，抽出した名詞を組織名辞書と比較し，一致した名詞を組織名として抽出する．そして，抽出した組織名の集合を $T_2(xy_m)$ とする．

(5) $T_1(xy_n)$ と $T_2(xy_m)$ のマッチングを行い，一致したものを関係情報 $RI_{nl}(xy_i)$ とする．

(6) $RI_{nl}(xy_i)$ において，弱い関係を排除するために，文書に含まれる単語の重要度算出方式として一般的に利用される TF/IDF（Term Frequency-Inverse Document Frequency）法により各組織名の重み係数を算出する．

(7) 各組織名の重み係数において，閾値による切捨てを行い，優良な関係情報を抽出する．

本提案手法の組織名の抽出においては，形態素解析器として，茶筌[8] を利用した．茶筌では，入力文の単語列に対して，品詞の出現確率の対数をリスクと定義し，実現可能な単語列，品詞列のうちリスクの和が最小となる組合せを形態素解析結果として採用する．また，組織名の抽出前に形態素解析を行った理由としては，Web ページの内容を単純に組織名辞書とマッチングする場合では，名詞以外の単語が組織と誤判定される問題を回避するためである．また，組織名の抽出で利用する組織名辞書は，茶筌に含まれる組織名辞書を利用した．組織名辞書に収録されている語数は，16,610 件である．そして，各組織名の重要度算出は，重要語句算出において，一般的に用いられている手法である TF/IDF 法を採用した．TF/IDF 法による，各組織名の重要度算出は，次式のとおりである．

$$TF(xy_n) = \frac{C(xy_n)}{\sum_{k=1}^{n'} C(xy_k)} \quad (5)$$

$$IDF(xy_n) = \log \frac{\sum_{k=1}^{x'} k}{df(xy_n)} \quad (6)$$

$$TFIDF(xy_n) = TF(xy_n) * IDF(xy_n) \quad (7)$$

最初に，式 (5) において，Web ページ x における任意の組織名 y_n を Web ページ内部の組織名の総数で割ることで，その Web ページ内部の任意の組織名の重み係数を算出する．次に，式 (6) において，対数計算を行い，任意の単語について，$S_{add}(x)$ の全体個数と任意の組織名が含まれる個数から，組織名の重みを調整する．ここで，df は，任意の組織名が含まれる Web ページの合計個数を示す．この処理により，知名度の高い組織は，組織名 1 件あたりの重みが低下する．最後に，式 (7) において，重み係数の積によって，抽出を行った組織名のその Web ページにおいての重要度を算出する．

そして，各組織間の関係情報の切捨てを行うために閾値を算出する．閾値の算出方法としては，まず，集合 $S_{add}(x)$ について，任意の個数の Web ページについて，手作業で組織間の関係を抽出し正解データを作成する．次に，正解データと本提案手法で取得した関係情報を比較することで F 値を算出し，これを評価値として最急降下法により最適な閾値を算出する．F 値とは，特定語句の抽出等で頻繁に利用される手法で，抽出データと正解データを比較し，抽出データの正解率を適合率，正解データの補完率を再現率として，両者の調和平均が F 値となる．また，閾値の算出時に局所解を回避するために，初期値をランダム抽出し，複数回の試行を行う．これにより，弱い関係情報を排除し，関係情報の品質を向上させる．

4.2 関係評価値の算出

HITS アルゴリズムにおける Authority Score と Hub Score の概念は，現在も多数の検索エンジンや既存研究で利用されており，非常に信頼性の高い評価方式であると考えられる．このため，本提案手法においても，関係評価値の算出は，HITS アルゴリズムと同様に，Authority Score と Hub Score による算出を行う．実際の評価値算出の手順を以下に示す．

(1) $S'(x)$ のリンク構造を被リンクのみ解析することで，関係情報の集合 $RI_{back}(x)$ を取得する．

(2) $RI_{back}(x)$ と $RI_{nl}(xy_i)$ の和により，関係情報の全体集合となる $RI_{all}(xy_i)$ を取得する．

(3) $S_{add}(x)$ と $RI_{all}(xy_i)$ から式 (1)，(2) による反復計算を行い，評価値の収束値を算出する．

以上の手順により得られた情報を利用することで，組織の関係グラフ $G(x)$ を作成する．

4.3 関係情報の可視化

本提案手法では，Authority Score および Hub Score の算出を行った後に，$S_{add}(x)$ から Authority Score が 0 に収束したものを削除し，残った Web ページの集合を $S_{auth}(x)$ とする．また，$S_{auth}(x)$ の算出において，削除された Web ページとの関係情報を $RI_{all}(xy_i)$ から削除する．そして，$S_{auth}(x)$ を頂点，$RI_{all}(xy_i)$ を辺としてグラフ $G(x)$ を作成する．グラフ $G(x)$ の描画では，最初に，$S_{auth}(x)$ をランダムに等間隔で格子状に配置し，各頂点において，関係情報にて結び付くすべての頂点との重心計算を行うことで，位置の補正を行う．その後に，関係情報に従って，頂点間に辺を描画する．

5. 評価実験

5.1 関係情報の品質評価

本節では，各手法により取得した関係情報を評価するために，組織名を基準として Web ページ群を収集し，Web ページ間の関係を解析した．Web ページ間の関係解析は，HITS アルゴリズムを利用した従来手法と本提案手法の 2 つの方法を用いて行い，Web ページ間の関係情報の品質を評価する．関係情報の品質評価には，組織間の関係において，特に重要な関係と思われる資本関係，提携関係と取引関係を収集し，取得データとの個別比較を行う．

実験で利用する Web ページ集合は，まず，Root 集合として，Google[9] から「co.jp」のキーワードで検索し，検索結果の上位 100 件の Web ページを取得した．次に，Root 集合に対して，リンク関係を解析し，Backward 集合 225 件と Forward 集合 770 件を取得し，合計 1,095 件の Web ページの集合を形成した．そして，1,095 件の Web ページの集合に対してリンク構造解析を行った結果，従来手法においては Backward 関係から 367 件，Forward 関係から 1,259 件の合計 1,626 件のリンク関係が得られた．また，本提案手法においては，Root 集合と Backward 集合の合計 325 件の Web ページから組織名を抽出し，新たに Google から同様の検索方法で取得した 101〜600 位の 500 件を追加した 825 件の Web ページを候補集合として，関係情報の補完を行った．その結果，1,352 件の関係を取得することができた．取得した関係の内訳は，資本関係が 745 件，取引関係が 139 件，業務提携関係が 468 件である．

これに対して，本提案手法によって取得した Web ページを目視にて確認を行い，資本，取引，業務提携の関係が Web ページで明示されているものを正解データとして 1,615 件の関係情報を取得した．正解データの内訳は，資本関係が 56.03%，取引関係が 12.69%，業務提携関係が 31.27%である．そして，この正解データを利用して HITS アルゴリズムおよび本提案手法に対して F 値を算出した結果を**表 1** に示す．また，本提案手法で用いた閾値は，初期値のランダム取得と最急降下法により算出した 0.31662 を採用した．表 1 の結果を確認すると，F 値においては，本提案手法が 61.5 と精度良く関係情報を取得していることが分かる．Forward 関係の精度が低い要因としては，組織以外の Web ページへのリンクが多かったためと考えられる．そして，Backward 関係の再現率が低い要因としては，関係情報の総数が少ないためと考えられる．また，本提案手法において関係情報が取得できなかった Web ページは，組織名辞書に収録されていない組織名が存在することが主要因であり，関係情報取得失敗のうち 7 割以上がこれに該当した．このため，本提案手法は，辞書の拡張および他の固有表現抽出技術を利用することで，さらなる精度向上が期待できる．

表 1 F 値による関係情報の評価
Table 1 Evaluations of information related by F-measure.

関係種別		適合率	再現率	F 値
HITS	Backward 関係	51.8	11.8	19.2
	Forward 関係	46.8	36.5	41.0
本提案手法	閾値切捨てなし	39.6	67.0	49.8
	閾値切捨てあり	67.5	56.5	61.5

また，各手法が取得した関係情報の傾向を確認すると，Forward 関係は，資本関係が多く，Backward 関係は，資本関係と取引関係が多く取得できていた．これに対して，本提案手法では，資本関係，業務提携関係を多く取得できており，資本関係は，Forward 関係と重複した関係情報を取得する傾向が見られた．このことから，組織関係の Web コミュニティを作成する場合は，適合率が高く，関係情報に重複の少ない Backward 関係と本提案手法を組み合わせて利用する方式が有効であると考えられる．

5.2 Authority Score による比較

先の実験において，関係情報の品質を評価することができた．しかし，組織間の Web コミュニティを生成する場合には，各 Web ページを評価し，主要な Web コミュニティを抽出する必要がある．このため，HITS アルゴリズムで利用される Authority Score および Hub Score を利用して，既存手法の Forward 関係，Backward 関係と本提案手法を比較することで評価を行う．

具体的な実験方法は，先の実験で得られた Root 集合，Backward 集合，Forward 集合の合計 1,095 件から Backward 関係と Forward 関係を利用し，個別に Authority Score を算出した．そして，本提案手法では，825 件の Web ページに対して，Authority Score の算出を行った．各手法の結果について，Forward 関係を**表 2**，Backward 関係を**表 3**，本提案手法を**表 4** に示す．

表 2 の結果を確認すると，上位 3〜26 の Web ページは，同一の Authority Score を示している．それは，これらの Web ページが，影響力の強い複数の Hub ページ群からのリンクに影響を受けて Authority Score が出力されていることを示す．そのため，一部の Web

表 2 Forward 関係のリンクに関しての処理結果
Table 2 Result on links of Forward relationship.

順位	ドメイン	値	組織
1	www5.hokkaido-np.co.jp	0.001821	×
2	www.pressnet.or.jp	0.001821	○
3	ad.hokkaido-np.co.jp	0.001821	×
4	jyoho.hokkaido-np.co.jp	0.001821	×
5	sumai.hokkaido-np.co.jp	0.001821	○
6	www.aurora-net.or.jp	0.001821	○
7	photokaido.hokkaido-np.co.jp	0.001821	○
8	kumasanda.com	0.001821	×
9	hplist.hokkaido-np.co.jp	0.001821	×
10	motto.hokkaido-np.co.jp	0.001821	○
11	www.uhb.co.jp	0.001821	○
12	www.air-g.co.jp	0.001821	○
13	www.tri-b.co.jp	0.001821	○
14	www.atkyushu.com	0.001821	○
15	ichioshi.info	0.001821	○
16	www.jrk-hotels.com	0.001821	○
17	www.mjr-sasabaru.com	0.001821	○
18	www.sjr.jp	0.001821	○
19	www.oita-kagoshima.jp	0.001821	×
20	www.mjr-tosu.com	0.001821	○
21	www13.jrkyushu.co.jp	0.001821	○
22	www.yoyaku.jrkyushu.co.jp	0.001821	○
23	www.jrkyushu-timetable.jp	0.001821	○
24	www.jrbeetle.co.jp	0.001821	○
25	www.jrkbus.co.jp	0.001821	○
26	www.jrsumai.co.jp	0.001821	○
27	www.kotorikyo.org	0.001787	○
28	shop.knt.co.jp	0.001787	○
29	holiday.knt.co.jp	0.001787	○
30	www.etabi-c.com	0.001787	○

表 3 Backward 関係のリンクに関しての処理結果
Table 3 Result on links of Backward relationship.

順位	ドメイン	値	組織
1	www.jterc.or.jp	0.022221	×
2	hokkaido.yomiuri.co.jp	0.021991	×
3	www.nikkei.co.jp	0.021868	×
4	ekikara.jp	0.021725	×
5	crocro.com	0.021263	×
6	news.kyodo.co.jp	0.021230	×
7	www.jr-central.co.jp	0.020964	○
8	www.kochinews.co.jp	0.020952	×
9	www.shinmai.co.jp	0.020923	×
10	melody.poke1.jp	0.020909	×
11	www.kahoku.co.jp	0.020900	×
12	www.iwate-np.co.jp	0.020795	×
13	www.nagasaki-np.co.jp	0.020703	×
14	www.hon-michi.net	0.020687	○
15	www3.ocn.ne.jp	0.020631	○
16	www.sanyo.oni.co.jp	0.020565	×
17	www3.coara.or.jp	0.020488	×
18	www.minyu.co.jp	0.020453	○
19	pub.bookmark.ne.jp	0.020409	×
20	expresscard.jp	0.020323	○
21	www.izu.co.jp	0.020263	○
22	www.jrkyushu.co.jp	0.020204	○
23	www.cyberstation.ne.jp	0.020190	○
24	www.pref.mie.jp	0.020179	○
25	www.jrkbus.co.jp	0.020122	○
26	www.jrkyushu-hospital.jp	0.020122	○
27	www.city.miyakonojo.miyazaki.jp	0.020122	○
28	artist.on.arena.ne.jp	0.020071	×
29	www.so-net.ne.jp	0.020069	○
30	www.net.pref.aomori.jp	0.020029	×

ページ集合の Authority Score が必要以上に上昇する Topic Drift 問題が発生していることが分かる．これにより，Web コミュニティの生成に Topic Drift の影響が出ることが容易に推測できる．

表 3 の結果を確認すると，取得した関係情報の上位のデータには，リンク集や新聞記事との関係が多い．これは，Authority Score の高い Web ページは互いに存在を承認し難いという性質[10] に起因するものである．そして，本研究の目的は，企業間の資本関係，業務提携関係，取引関係等の現実社会における組織関係情報を取得することであるが，これらの関係情報は，Authority Score の高い Web ページ間で形成されることが多いために，Backward 関係のみでは，Authority Score の高い Web ページ間の関係情報が十分に取得できていないことが分かる．

表 4 の結果を確認すると，グループ関係にある会社が上位に表示されていることが分かる．これは，企業のホームページの会社概要ページにおいて，資本関係や業務提携等の情報が多く公開されていることに起因する．また，表 2，表 3，表 4 それぞれの上位 50 件に表示されているページの内訳を**表 5** に示す．

表 5 を確認すると，本提案手法と Forward 関係とでは，組織の Web ページ間の関係情報が多く取得できていることが分かる．しかし，Forward 関係に関しては，Topic Drift が発生していると予想されるために，表 5 で取得できている組織名の割合に疑問が残る．また，Backward 関係においては，組織以外にニュース，新聞やリンク集の Web ページとの関係が多い．これは前述の Authority Score の高い Web ページの競合問題によるものと考えられるが，これらの Web ページは，組織名の関係情報が多く存在するため，本提案手法と組み合わせることで，Web 上を効率良く探索できると考えられる．また，本提案手法を利用して取得した関係情報は，組織の Web ページが上位に多く存在することから，Backward 関係を基点に含むことによる精度低下は発生しなかったと考えられる．

5.3 グラフの可視化による比較

本節では，Forward 関係と Backward 関係，そして本提案手法により取得した関係情報から構成する Web コミュニティを確認するために，グラフによる

表 4 本手法で取得したリンクに関しての処理結果
Table 4 Result on links acquired with original method.

順位	ドメイン	値	組織
1	www.jrkyushu.co.jp	0.054430	○
2	www.jrkyushu-hospital.jp	0.054430	○
3	www.jreast.co.jp	0.054128	○
4	jreast.eki-net.com	0.054094	○
5	voice.jreast.co.jp	0.054094	○
6	www.jrhokkaido.co.jp	0.053592	○
7	www3.jrhokkaido.co.jp	0.053592	○
8	mobile.jrhokkaido.co.jp	0.053592	○
9	expresscard.jp	0.053329	○
10	www.jr-central.co.jp	0.053104	○
11	www.cyberstation.ne.jp	0.051139	○
12	www.jrk-hotels.com	0.051139	○
13	www.jrkyushu-timetable.jp	0.051139	×
14	www.jreast-timetable.jp	0.051139	×
15	www.calc.eki-net.com	0.051139	○
16	news.kyodo.co.jp	0.003664	×
17	bb.i-seven.ne.jp	0.002799	○
18	www.sanyo.oni.co.jp	0.002758	×
19	www.ebookjapan.jp	0.001418	×
20	www.panasonic-europe.com	0.000910	×
21	www.excite.de	0.000910	○
22	ritz-photo-images.com	0.000910	×
23	hokkaido.yomiuri.co.jp	0.000906	×
24	www.sanplatec.co.jp	0.000886	○
25	210.150.210.59	0.000886	×
26	www.mitsui.co.jp	0.000885	×
27	www.mbfutures.com	0.000885	○
28	www.btm.co.jp	0.000880	○
29	www.japanfs.org	0.000510	○
30	www.ibm.com	0.000510	○

表 5 本手法で取得したリンクに関しての処理結果
Table 5 Result on links acquired with original method.

項目	従来手法		本提案手法
	Forward	Backward	組織名抽出
リンク集	2	16	0
ニュース・新聞	2	10	6
組織	26	15	39
その他	20	9	5

可視化[11),12)] を行う．具体的な可視化方式については，Authority Score の比較実験と同じく，Forward 関係と Backward 関係については，基準となる Web ページとして 1,095 件を取得し，目視による判定で組織の Web ページ 678 件を抽出した．また，本提案手法では，基準となる Web ページ 825 件から組織の Web ページ 627 件を抽出した．さらに，これらの Web ページから，先の実験で Authority Score が 0 に収束した Web ページを排除し，主要な Web コミュニティのみを抽出する．そして，Web ページを点としてランダムに配置し，これらの Web ページ集合に関する関係情報のみを線として描画した．また，各点において，重心計算による点の再配置を行った．本実験において，優良なグラフは，線の多いグラフであり，さらに，グラフを判定する評価値としては，先の関係情報の品質評価での適合率とグラフの本数の積が有用なグラフの本数を示す指標となるため，これを比較し，グラフの評価を行う．各手法の結果について，Forward 関係を**図 2**，Backward 関係を**図 3**，本提案手法を**図 4** に示す．

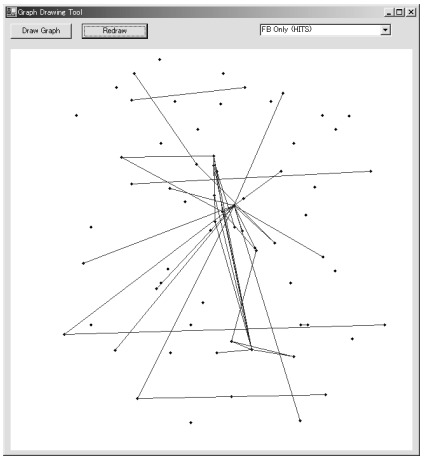


図 2 Forward 関係のリンク構造解析による関係情報の可視化
Fig. 2 Visualization of Forward information related by link structure analysis.

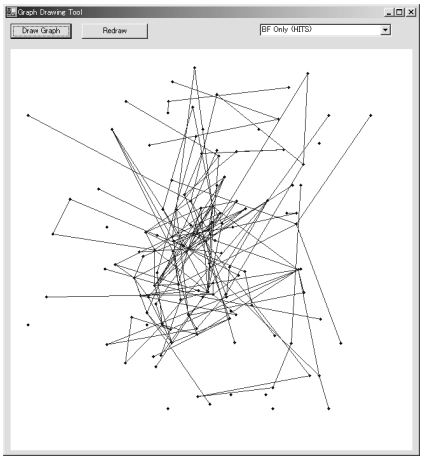


図 3 Backward 関係のリンク構造解析による関係情報の可視化
Fig. 3 Visualization of Backward information related by link structure analysis.

図 2 の結果から，Forward 関係では，Topic Drift の影響で Authority Score が特定の Web ページに集中しているため，Authority Score を基準とした Web コミュニティの構築では，Topic Drift の影響を受けた Web ページ群しか抽出できていないことが分かる．また，グラフの評価値は，21.5 となり，他の 2 手法と

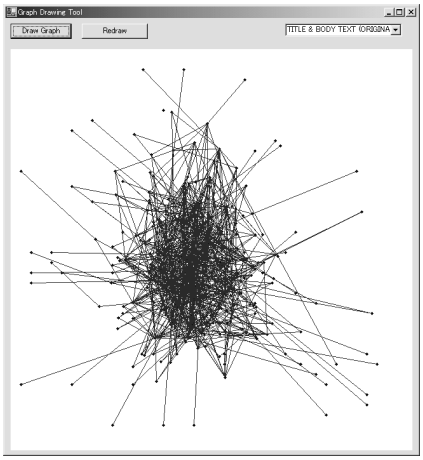


図 4 自然言語処理による関係情報の可視化
Fig. 4 Visualization of information related by natural language processing.

比較すると非常に低い結果となった．

図 3 の結果では，Backward 関係は，Forward 関係と比較すると，組織の Web ページの Authority Score が 0 に収束する件数が少なかったために，Forward 関係よりも多くの線が描画されていることが分かる．また，グラフの評価値は，149.2 となり，Forward 関係によるグラフ描画よりも品質が良いことが確認できた．しかし，実際の Web コミュニティの構築では，1 つの Web ページに対して複数の関係情報が接続されている状態が理想的[13),14)] であるために，Backward 関係のみでは，抽出情報が少ないと考えられる．

これに対して，図 4 の結果では，図 3 の結果よりも，多くの線が描画されていることから，組織の Web ページの Authority Score が 0 に収束する件数がさらに少なかったことを示す．そして，評価値が，764.1 と他の 2 手法と比較すると非常に高く，複数の線で結合された点が多いことから，Web コミュニティ判定においても有用[15),16)] であると考えられる．以上の結果から，本提案手法は，リンク構造解析において取得した関係情報より Web コミュニティ生成に有効であると判断できる．

6. おわりに

本研究では，自然言語処理を利用することで，現実社会における資本関係，業務提携関係，取引関係といった組織間の関係を従来手法より多く取得することができた．また，取得した関係情報は，本提案手法の実験結果より，品質の高い関係情報が取得できたことが分かる．このため，本提案手法は，信頼度の高い組織間の関係グラフを作成することが可能といえる．また，本提案手法は，組織名の関係抽出においての実験を行ったが，同様に，人名や地名等の固有表現は，Web ページにおいて重要度の高い単語となるため，本提案手法による関係情報の抽出が有用であると考えられる．

本提案手法の今後の発展としては，Semantic Web における RDF（Resource Description Framework）や FOAF（Friend of a Friend）をリンク構造や Web ページ内容との関連付けに活用することでより詳細な情報を取得し，信頼性の高い Web の自動解析の研究を行う予定である．

参 考 文 献

1) 松生泰典，小山　聡，田中克己，是津耕司：Web 検索結果とその周辺情報の近似的内包表現とその視覚化，情報処理学会データベースシステム研究会研究報告，Vol.104, No.176, pp.151–156 (2004).
2) 丸山謙志，王　冠超，徳山　豪：Web 検索結果におけるクラスタリングアルゴリズムの研究，情報処理学会アルゴリズム研究会研究報告，Vol.2005, No.26, pp.17–24 (2005).
3) 友部博教，松尾　豊，武田英明，安田　雪，橋田浩一，石塚　満：Semantic Web のための人の社会ネットワークの抽出と利用，情報処理学会論文誌，Vol.46, No.6, pp.1470–1479 (2005).
4) 加藤一民，松尾啓志：Markov Cluster Algorithm を用いた Web コミュニティ群の発見手法，情報処理学会自然言語処理研究会研究報告，Vol.2005, No.22, pp.87–93 (2005).
5) 山本仁志，太田敏澄，石田和成，岡田　勇：リンク構造と共起関係を用いた Web 空間の視覚化，情報処理学会デジタルドキュメント研究会研究報告，Vol.2004, No.36, pp.95–101 (2004).
6) 豊田正史，吉田　聡，喜連川優：ウェブコミュニティチャート―膨大なウェブページを関連する話題を通して閲覧可能にするツール―，電子情報通信学会論文誌，Vol.J87-D-1, No.2, pp.256–265 (2004).
7) Kleinberg, J.M.: Authoritative Sources in A Hyperlinked Environment, *J. ACM*, Vol.46, No.5, pp.604–632 (1999).
8) 茶筌．http://chasen.naist.jp/hiki/ChaSen/
9) Google. http://www.google.co.jp/
10) 野村早恵子，小山　聡，早水哲雄，石田　亭：Web コミュニティ発見のための HITS アルゴリズムの分析と改善，電子情報通信学会論文誌，Vol.J85-D-1, No.8, pp.741–750 (2002).
11) 田地　晶，宮寺庸造，櫨山淳雄，横山節雄：ユーザ思考に基づく学術論文関係図の可視化手法の提案，電子情報通信学会教育工学研究会研究報告，Vol.100, No.420, pp.37–44 (2000).
12) 土橋　喜，山内平行，立花隆輝：キータームの関連性の視覚化による知識連鎖の発見支援―

TermLinker システムの可視化機能，情報処理学会知能と複雑系研究会研究報告，Vol.103, No.304, pp.41–46 (2003).
13) Anderberg, M.R.，西田英郎，佐藤嗣二，江藤香，寺尾 裕，宮井正彌：クラスター分析とその応用，内田老鶴圃 (1988).
14) 一森哲男：グラフ理論，共立出版 (2002).
15) 秋山 仁：グラフ理論最前線，朝倉書店 (1998).
16) 立花俊一，奈良知恵，田澤新成：グラフ理論への入門，共立出版 (1991).

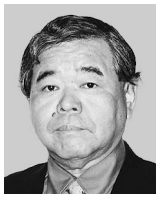
(平成 17 年 10 月 18 日受付)
(平成 18 年 4 月 4 日採録)

池辺 正典（学生会員）

1977 年生．2002 年関西大学総合情報学部卒業．2004 年関西大学大学院総合情報学研究科知識情報学専攻博士前期課程修了．現在，関西大学大学院総合情報学研究科総合情報学専攻博士後期課程在学中．修士（情報学）．文書処理，自然言語処理，データマイニング等の研究に従事．2000 年（株）関西総合情報研究所入社，現在に至る．Web アプリケーション，データモデル設計等の研究開発に従事．土木学会学生会員．

田中 成典（正会員）

1963 年生．1986 年関西大学工学部土木工学科卒業．1988 年関西大学大学院工学研究科土木工学専攻博士前期課程修了．同年（株）東洋情報システム（現在，TIS）に入社，知識情報処理システムに関する研究受託開発業務に従事．1994 年関西大学総合情報学部専任講師．1997 年助教授．2003 年教授，現在に至る．博士（工学）．2002 年 8 月から 1 年間カナダの UBC にて客員助教授．専門は知識工学と土木情報学．2000 年（株）関西総合情報研究所を起業．土木学会，GIS 学会，IABSE，人工知能学会，日本知能情報ファジィ学会と情報知識学会各会員．現在，土木学会土木情報システム委員会幹事長，国土交通省建設情報標準化委員会委員，ISO/TC184/SC4 委員．



古田 均

1948 年生．1971 年京都大学工学部卒業．1973 年京都大学大学院工学研究科修士課程修了．1976 年同大学院工学研究科博士課程修了．同年京都大学工学部助手．その後講師，助教授を経て，1994 年関西大学総合情報学部教授，現在に至る．その間，米国パディー大学客員助教授，米国プリンストン大学客員研究員，2004〜2005 年米国コロラド大学客員教授．構造物の信頼性解析，最適設計，ライフサイクルコスト解析，ソフトコンピューティングの構造設計・維持管理への応用に関する研究に従事．著書に『ファジィ理論の土木工学への応用』，『建築土木技術者のためのファジィ理論入門』，『遺伝的アルゴリズムの構造工学への応用』，『Life-Cycle Cost Analysis and Design of Civil Infrastructure Systems』等．日本知能情報ファジィ学会，計測自動制御学会，システム制御情報学会，土木学会，日本建築学会，日本材料学会，日本鋼構造協会，ASCE 各会員．

中村 健二（学生会員）

1981 年生．2004 年関西大学総合情報学部卒業．2006 年関西大学大学院総合情報学研究科知識情報学専攻博士前期課程修了．現在，関西大学大学院総合情報学研究科総合情報学専攻博士後期課程在学中．修士（情報学）．システム設計手法，自然言語処理，データモデル等の研究に従事．2002 年（株）関西総合情報研究所入社．現在に至る．システム設計，データモデル設計等の研究開発に従事．土木学会学生会員．

小林 建太（学生会員）

1983 年生．現在，関西大学総合情報学部在学中．自然言語処理，データマイニングの研究に従事．2005 年（株）関西総合情報研究所入社，現在に至る．Web アプリケーションの研究開発に従事．土木学会学生会員．