Vol.2009-DBS-149 No.21
2009/11/21

次数制約を加えた共起語グラフに基づくキーワード間ナビゲーション

島田 諭[†1] 福原知宏[†2] 佐藤哲司[†1]

本研究では，文書空間から抽出した共起語から有向グラフを生成し，ユーザを文書空間内の関連文書に導くナビゲーション手法を提案している．提案手法では，ノードの最大出次数に制約を与え，反復度が低い語から高い語へのリンクを優先的に生成する．しかし，これらのパラメータの閾値と生成されるネットワークの特性との関連は明らかではなかった．本稿では，新聞記事を用いた実験により，最大出次数とネットワーク特性の相関を明らかにするとともに，Bow-tie 構造における SCC（強連結成分）に着目し，生成されるグラフのナビゲーションにおける有用性を考察する．

Inter Keyword Navigation based on Degree-constrained Co-Occurrence Graph

Satoshi Shimada,[†1] Tomohiro Fukuhara[†2] and Tetsuji Satoh[†1]

We had proposed a navigation method that generates directed graph based on co-occurrence words, and leads users to the related document in the document space. In the method, because of to reduce user's load, maximum outdegree of nodes is constrained. However, the correlation of the threshold and the characteristic of generated graph was not clear. In this paper, we describe the experiment that to clarify this correlation using the newspaper article. And we discuss about utility of the generated graph by considering SCC (Strong Connected Components) in the Bow-tie structure.

†1 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba
†2 東京大学人工物工学研究センター
Research into Artifacts, Center for Engineering, The University of Tokyo

1. はじめに

ある主題に関して正確な判断を下すためには，多数意見的な情報のみに偏ることなく，少数意見的な情報も可能な限り収集することが不可欠である．このためには，Wikipedia，新聞記事データベース，質問回答サイトなど，情報が網羅的に集約されていることが期待される文書集合を利用し，その文書集合内において網羅的に情報を探索することが有効である．

これまで，網羅的な検索や探索的な検索を支援する手法として，文書クラスタリング手法や意見情報の賛否を判定する手法などが提案されている．しかし，文書クラスタリングは，文書集合全体の概略の把握には適するが，文書間やクラスタ間の関連性の把握には適さない．また，分類の困難な情報や，賛否の書かれていない情報は，従来にない視点からの記述であったり，賛否の定まっていない最新のトピックであったりする可能性が高い．ある主題に関して正確な判断を下すことが可能な水準まで理解を深めるためには，単に全体を俯瞰するだけでなく，分類の困難な情報を含む多様な情報を参照することが不可欠と考えられる．

我々は，文書集合における語の共起に基づき，局所的な関連と大域的な関連をユーザに提示する，Comprehensive Web Navigation を提案している．ユーザは，システムが提示する関連項目を順番に選択していくことで，文書集合内を探索できる．このようなナビゲーションにおいては，ユーザによる項目選択にかかる心理的負荷を最小限に留めるため，一度に提示する項目数を少数に絞り込む必要がある．つまり，ノードあたりの最大出次数を制約しながらも，グラフ構造全体としては平均距離が短く平均クラスタ係数が高い Small-world グラフ[1)] を生成することが要求される．また，ユーザを多様なノードへ誘導するためには，基点となる各ノードが多様なリンクを持つこと，すなわちグラフの密度が低いことが求められる．さらに，Bow-tie 構造[2)] における SCC（強連結成分）の比率を最小限に留め，一方通行のパスを増やし，誘導性を高める必要があると考えられる．

本稿では，Comprehensive Web Navigation に適すると考えられる有向グラフを生成するためのパラメータについて検討する．新聞記事を用いた実験により，最大出次数の制約と生成されるグラフの特性の関係を明らかにするとともに，生成グラフにおいて SCC ノードとなる語の性質，および共起語グラフに対する SCC の圧縮率に着目し，生成されるグラフのナビゲーションにおける有用性を考察する．

以下，2 章で関連研究を概説し，3 章で提案手法について説明し，4 章で評価実験の概要と結果を示し，5 章で考察し，6 章でまとめる．

2. 背　　景

2.1 探索的検索におけるナビゲーションの必要性

明確な情報要求を持たない段階からの情報探索を支援する手法は，探索的検索[3] と呼ばれ，その必要性が指摘されている．探索的検索においては，検索対象の文書集合がどのような文書を含むのか，それらの文書には，どのような内容が，どのような詳しさで，どのような表現で記述されているのかをユーザが把握できるようにすることが重要である．

このような文書集合の特徴をユーザが把握しやすくするためには，文書集合全体の要約となるような内容を提示することが第一に考えられる．しかし，文書集合全体からバランスよく生成した要約では，実際の個々の文書から受ける印象とは異なる印象をユーザが受ける可能性が生じる．また，ユーザの情報要求を明確化するためには，要約ではなく個々の文書を提示してフィードバックを得ることが不可欠である．

本研究では，文書集合の全体を網羅するナビゲーションを提供することにより，ユーザが文書集合内を探索的に検索できると考える．検索の開始時に検索語の入力を必須としないナビゲーションを用いることで，必ずしも明確な要求を持たないユーザも，新聞や書籍を手に取ってページをめくる感覚で文書集合内を探索できる．ユーザは，要求への適合度の低い文書も含む多様な文書を閲覧するうちに，文書集合に特有の用語や表現を把握でき，複雑に関連する一連の内容を俯瞰的に理解できると考えられる．

2.2 従来手法の問題点

現在，文書集合内のナビゲーションとして，入力された検索語や検索結果に対し，関連語や関連文書を提示する手法が広く用いられている．しかし，ユーザを多様な文書へ誘導することはほとんど考慮されておらず，類似度の高い項目を関連項目として提示するものが多い．このため，類似度の低い文書間を遷移することは困難で，文書空間内を広く探索するための支援手法にはなっていない．

また，関連語を提示する手法の多くでは，検索語の入力補完，関連記事の要約という機能が重視され，探索的な検索を支援することは必ずしも意図されていない．このため，例えばタグクラウドのように，一度に数十個以上の語を提示するものも少なくない．これでは，ユーザが直感的に語を選択することは難しいと考えられ，関連する項目間を連鎖的に遷移するためのナビゲーションとしては機能しにくいと考えられる．

服部らは，ページ中からユーザが選択した語および，その周辺語を用いる検索手法を提案している[5]．この手法では，十字キーや矢印キーなど限られた入力手段しか持たない携帯端末やセットトップボックス（STB）での利用を念頭に，文字列の入力操作を行うことなく検索を実行可能としている．しかし，検索語の選択をユーザに委ねていることから，ページ中のどの語を検索語とすればよいかを適切に判断できないユーザにとっては利用が難しいと考えられる．

酒井らは，ユーザが気づきにくいような関連情報を提示することにより，ユーザの情報要求の変化を誘発することを狙った検索インターフェースを提案している[6]．この手法では，Web 検索を支援するために，Wikipedia における参照関係を提示するが，検索対象の文書集合と Wikipedia の間では，語法や表現，概念の粒度など，さまざまな特性のミスマッチがある．この影響を回避するには，検索対象の文書集合に対して外部知識を導入するのでなく，検索対象の文書集合そのものから抽出した関連性を用いてナビゲーションを行うことが不可欠と考えられる．

また，いずれの手法においても，ユーザが文書集合内を広く探索できることを構造レベルで保証するものではなく，ユーザの行動や，導入する外部知識の質に大きく依存している．

3. Comprehensive Web Navigation

3.1 グラフ構造を用いた文書空間ナビゲーション

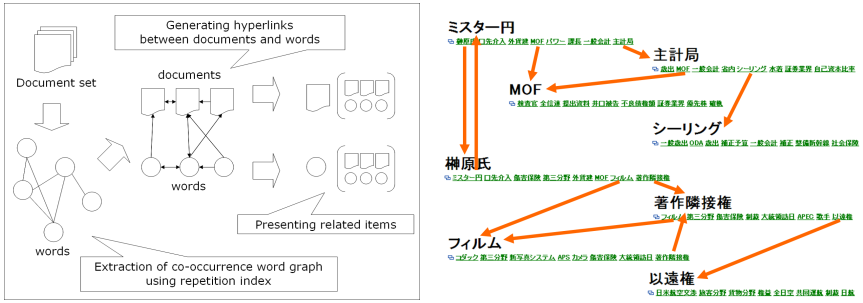
Comprehensive Web Navigation は，ユーザによる包括的な内容把握の支援を目的とする．ユーザが見落としやすい情報にも到達できるためには，明示的に入力される検索語に頼らず，提示語の選択にも前提知識を要求せず直感的に選択でき，提示語を順番に選択するだけで多様な文書に到達できるような遷移経路の生成が必要である．このためには，関連する文書間に漏れなくリンクを付与しておくことが必要である．一方で，一度に膨大な遷移経路が提示されると，ユーザによる選択が困難になる．したがって，ノードあたりの出次数を少数に留めながら，文書集合全体を網羅できることが求められる．すなわち，Comprehensive Web Navigation に適するグラフ構造としては，以下に示す要件を満たす必要がある．

Small-world 性

短い平均距離で多様なノードへ到達できるグラフ構造として，Small-world ネットワークの存在が知られている[1]．グラフの Small-world 性は，ノード数およびエッジ数が同数のランダムグラフと比較し，平均距離（L）が同程度で，平均クラスタ係数（C）が非常に大きい場合として定量化されている[4]．また，共起語グラフが Small-world 性を示すことが明らかになっている．

Comprehensive Web Navigation においては，共起語グラフが示す Small-world 性を維

Vol.2009-DBS-149 No.21
2009/11/21



(a) 提案システムにおける処理の流れ　　(b) ナビゲーションによる遷移イメージ

図 1　提案手法の概要

持しながら，後述する他の要件を同時に満たすようにグラフ構造を変換する必要がある．

リンクの多様性

ユーザを多様なノードへ誘導するためには，基点となる各ノードが多様なリンクを持つことが求められる．リンクの多様性とは，あるノードと別のノードが同じノードへのリンクを持つ確率が低いこと，すなわちグラフの密度が低い状態を意味する．共起語グラフは本来，無向グラフであるが，共起関係の参照方向を限定することにより有向グラフとし，ノード間のリンクを非対称にすることで密度を低下させることができると考えられる．

誘　導　性

ユーザを誘導する性質（以下，誘導性と記す）を高めるためには，みちなりに進むしかないような一方通行のパスを用意することが有効と考えられる．提示された少数の項目をユーザが適当に選択するだけで，ノード間に生成された一方通行のパスをみちなりに進むことができれば，数段の隔たりはあるものの関連はあるといった，必ずしも類似度の高くない関連項目への誘導性を高めることができると考えられる．

Broder らは，World Wide Web（WWW）のリンク構造が Bow-tie 構造となることを明らかにした[2]．Bow-tie 構造において，グラフの各ノードは，SCC（Strongly Connected Components，強連結成分），SCC へのリンクのみを持つ IN，SCC からのリンクのみを持つ OUT，IN または OUT との間にのみリンクを持つ Tendrils，IN からのリンクと OUT へのリンクを持つ Tubes，および，これらとの間にリンクを持たない DCC（Disconnected Components）のいずれかに分類される．本稿では，この分類を応用し，提案手法により生成される有向グラフの構造を分析する．

表 1　キーワード区分の閾値と関連度計算のためのスコア

区分	反復度	df	スコア
I	≥ 0.6	> 2	10
II	$\geq 0.35, < 0.6$	$> 9, < 19$	1
III	≥ 0.1	> 3	0.1
IV	（区分 I，II，III 以外）		0.01

一方通行のパスが多いことは，Bow-tie 構造における IN，OUT，Tendrils，Tubes の比率が高いことと等しい．グラフ構造の誘導性を高めるためには，SCC に流入する IN の比率を高め，SCC そのものの比率は最小限に留める必要があると考えられる．

3.2　グラフ構造の生成手法

提案手法では，図 1 に示すように，文書およびキーワードをノードとするグラフを生成してナビゲーションを行う[9]．なお，本稿では提案手法のうち，キーワードのみをノードとする関連語グラフについて議論する．また，本稿では，提案手法による関連語グラフの生成におけるキーワードあたりのリンク数の上限値を最大出次数と呼ぶ．

本手法におけるグラフ構造の生成手法の概要を以下に示す．

キーワードの抽出および重み付け

漢字とカタカナからなる文字列，および英数字と一部の記号からなる文字列を抽出してキーワードの候補とする．ただし，1 文字の漢字またはカタカナからなる文字列，2 文字以下の英数字からなる文字列，数字のみからなる文字列，および URL として解釈できる文字列を除く．

表 1 に示す反復度（df_2/df）[7]，および df，df_2 の閾値を用いて 4 段階に区分し，スコアを付与する．区分 I は，特定のトピックを表現する性質の強い語，区分 II は，区分 I の語に付随することでトピックを狭める性質を持ちつつ，異なるトピック間を橋渡しする性質を示すと考えられる語である[⋆1]．区分 I，II に含まれない語については，おもに df の範囲に応じて区分 III または区分 IV に区分する．これらの閾値は，約 300 件のブログ記事を用いた予備調査により決定した[8]．なお，$df < 3$ となる低頻度語は除外する．

⋆1 本手法では品詞推定を用いないが，区分結果を目視すると，区分 I には固有名詞が，区分 II にはサ変名詞が多く含まれる．

Vol.2009-DBS-149 No.21
2009/11/21

表 2 実験に用いたデータ

データ名	文書数	キーワード数
朝日新聞（1996 年版）経済面	7,770	20,103

表 3 実験に用いたデータから抽出された共起語グラフとランダムグラフの比較

グラフ	ノード数	リンク数	密度	C	L	最長最短距離
共起語グラフ	20,103	1,783,732	0.0044138	0.6636081	2.49359	7
ランダムグラフ	(同上)	(同上)	(同上)	0.0044116	2.66863	3

キーワード間のリンクの生成

基点となるキーワード t_i が出現する文書集合 D_i において出現するキーワード集合を $T_i = \{t_1, ..., t_n\}$ とし，キーワード t_i と t_k が共起する文書集合を $D_{ik} = \{d_1, ..., d_m\}$ とする．基点となるキーワード t_i とキーワード t_k の間の関連度 $r(t_i, t_k)$ を，式 (1) により算出する．ここで，w_k は語 t_k のスコアである．

$$r(t_i, t_k) = mw_k : i \neq k \quad (1)$$

基点となるキーワード t_i と 1 件以上の文書で共起する全キーワードについて，関連度 r を算出する．起点となるキーワードと，関連度 r の上位から一定数のキーワードとの間に，リンクを生成する．

以上の方法により，文書集合に出現する $df \geq 3$ となるすべてのキーワードが，設定した最大出次数を上限とする関連語への出リンクを有する，有向グラフが生成される．このとき，キーワードの区分に応じたスコアに基づき，一部の文書でのみ繰り返し出現する語へのリンクを優先する．また，関連語の取得範囲は，起点となる語が出現する全文書であり，共起文書数の多い語へのリンクを優先する．

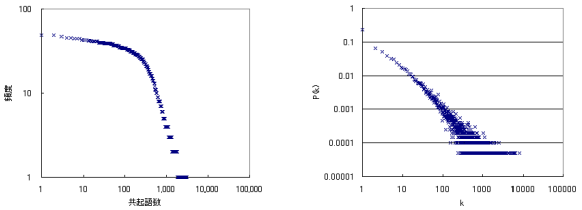
4. ネットワーク特性の分析

本研究では，関連語グラフを用いてユーザを多様な文書へ誘導することを目指しているが，提案手法におけるパラメータと生成されるグラフの特性との関係は明らかではなかった．このことについて，新聞記事を用いた実験を行ない，Bow-tie 構造における SCC（強連結成分）に着目し，最大出次数の制約と生成されるグラフの特性の関係を明らかにする．

4.1 実験に用いるデータ

本稿では,「朝日新聞記事データ集 学術研究用」(1996 年版，以下「朝日新聞」と記す）を用いた．文書数，および提案手法により抽出されるキーワード数を，表 2 に示す．

1996 年版より「1 経済」「2 経済」「3 経済」の各面に掲載された 1 年分の記事をすべて利用した．記事に付与されている見出しは，記事本文と一体で扱う．企業名や経済に関する用語が頻出する一方，用語の統制により同一の概念は常に単一の語で表記される．このような性質は，読者層やテーマを絞って書かれるブログや，新聞記事の書き方を参考にして書かれるニュースサイト風のブログ等においても見られる性質である．



(a) 共起語数　　(b) 次数分布

図 2 実験に用いたデータから抽出された共起語グラフの特性

共起語グラフ

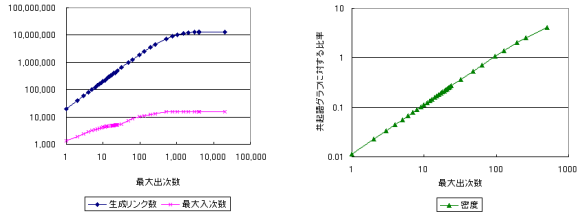
本稿では，提案手法により生成される関連語グラフとの比較対象として，共起語グラフ，およびランダムグラフを用いる．本稿における共起語グラフとは，実験対象データから関連語グラフと同様の方法で抽出したキーワードに対し，同一文書内で共起するキーワードとの間にエッジ（無向リンク）を付与したものである．共起文書数は考慮せず，エッジの重複のないグラフとなる．抽出された共起語グラフの特性を表 3 および図 2(a) に示す．ノードおよびリンクが同数のランダムグラフと比較して平均距離が同程度，平均クラスタ係数が非常に大きい，Small-world グラフである．次数分布はべき則に従っている．なお，共起語グラフのエッジ数は，表中に示したリンク数の半数である．

4.2 実験の概要

本稿における実験の概要を以下に示す．

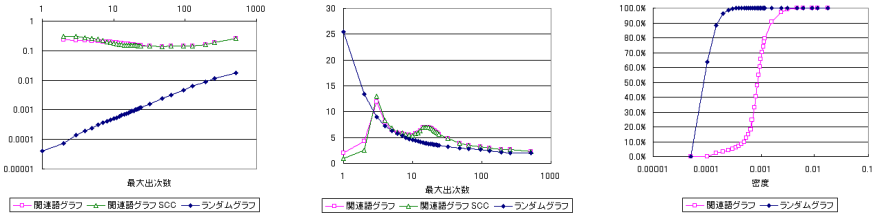
(1) 最大出次数を変えて関連語グラフを生成し，ランダムグラフと比較する．
(2) 生成された関連語グラフの次数分布を分析する．
(3) 生成された関連語グラフにおける Bow-tie 構造を抽出する．
(4) Bow-tie 構造を用いて切り出したサブグラフを可視化する．

Vol.2009-DBS-149 No.21
2009/11/21



(a) 生成リンク数と最大入次数　　(b) 密度

図 3　最大出次数の制約による生成グラフの変化



(a) 平均クラスタ係数 (C)　　(b) 平均距離 (L)　　(c) 到達可能ペア

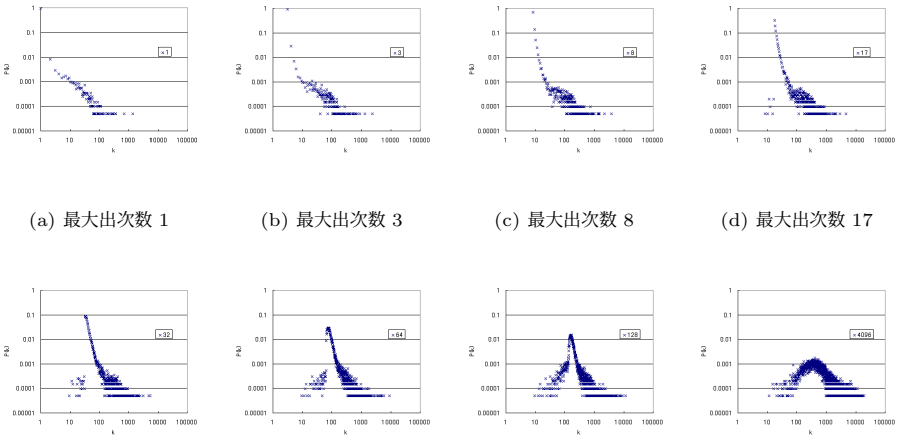
図 4　生成された関連語グラフとランダムグラフの比較

ネットワーク指標の算出，ランダムグラフの生成，およびグラフの可視化には Pajek[⋆1] を用いた．

4.3 実験結果

4.3.1 関連語グラフの生成およびランダムグラフとの比較

ナビゲーションにおいて，提示項目を選択するユーザの負荷に直結する最大出次数と，グラフ全体の特性の関連を明らかにするため，ノード数を一定とし，最大出次数を変えて関連語グラフを生成した[⋆2]．設定した最大出次数と生成されたリンク総数および最大入次数の関係を図 3(a) に示す．また，共起語グラフとの密度を比較した結果を図 3(b) に示す．最大出次数ごとの密度を，共起語グラフの値を 1 とした比率で示す．最大出次数 64 以下において，共起語グラフの密度を下回っている．提案手法では共起語間のリンク生成を関連度スコアの上位 20%に限っているため，極端に多数のリンクが生成されることや，関連の極めて薄い語間にリンクが生成されることは少ない．このため，最大入次数には制約を加えていないものの，実質的には上限が存在する．なお，関連語グラフ，共起語グラフともにリンクの重みは考慮していない．



(a) 最大出次数 1　(b) 最大出次数 3　(c) 最大出次数 8　(d) 最大出次数 17

(e) 最大出次数 32　(f) 最大出次数 64　(g) 最大出次数 128　(h) 最大出次数 4,096

図 5　次数分布

生成された関連語グラフが Small-world グラフであるか調べるため，同数のリンクを持つランダムグラフを生成して比較した．平均クラスタ係数を図 4(a) に，平均距離を図 4(b)

⋆1 http://pajek.imfm.si/doku.php

⋆2 最大出次数は，実験結果を対数グラフ上にプロットすることを考慮し，1 から 24，および，32，48，64，96，128，192，256，512，768，1024，1536，2048，3072，4096 の 38 通りに設定した．

に示す．関連語グラフでは，全域にわたり平均クラスタ係数がランダムグラフよりも非常に大きくなった．平均距離は，最大出次数が 1 および 2 ではランダムグラフを大幅に下回ったが，3 以上の領域ではほぼ同程度になった．このことから，最大出次数 3 以上において，関連語グラフは Small-world 性を示した．平均距離は，最大出次数が 17 前後の領域においては，ランダムグラフよりも長めになる傾向が見られた．

なお，ランダムグラフでは最大出次数の制約がなく，リンク数が少なくても出次数の大きいノードが出現する．このため関連語グラフでは，図 4(c) に示すように，グラフの密度に対する到達可能ペアの比率の変化がランダムグラフとは異なる．

4.3.2 次数分布

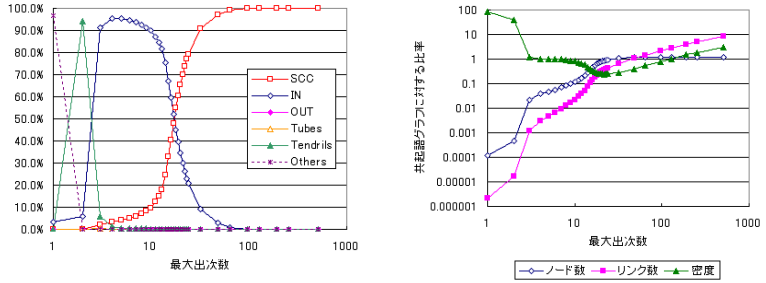
Small-world グラフにおいては，次数分布の違いによって特性が大きく異なる可能性が指摘されている．主な最大出次数における次数分布を図 5 に示す．次数 k は出次数と入次数の合計，$P(k)$ は，次数が k であるノードの数である．最大出次数が小さいほどべき分布に近く，大きいほど対数正規分布に近い．

なお，最大出次数 1 から 8 までは，最小出次数が最大出次数と一致した．最大出次数 9 以上では，最小出次数は 8 となり，最大出次数を大きくしても一定になった．これは，実験に用いた新聞記事において，共起語数が極端に少ない語のみが出現する記事や，出現語数が極端に少ない記事が存在しないためと考えられる．最大出次数 4,096 においては，生成グラフにおける実際の最大出次数は 3,879 であり，全語が出次数の制約を受けずに，生成可能な全リンクを生成した状態になった．

4.3.3 Bow-tie 構造の抽出

生成された関連語グラフから Bow-tie 構造を抽出した．密接なクラスタを形成する SCC を核に，SCC に流入するリンクを持つ IN，SCC から流出するリンクを持つ OUT，SCC を経由せずに IN と OUT をつなぐ Tubes，IN または OUT とはつながるが SCC とはつながらない Tendrils，および，これらと分離した DCC に区分される．これらの要素（ノード）の比率と最大出次数の関係を図 6(a) に示す．

最大出次数 1 では 96.8%が DCC となり，ほとんどクラスタを形成していない．最大出次数 2 では，94.0%が Tendrils となり，IN も 5.9%に増加する．最大出次数 3 において，IN が 91.4%になる．SCC ノード数は 8 から 378 に急増する．これはグラフ全体のわずか 1.9%に過ぎないが，SCC が十分に大きくなることで，他の大多数のノードが Tendrils ではなく IN になったと考えられる．IN は最大出次数 4 をピークに減少し，SCC の比率が増加していく．IN と SCC の比率の大小は，最大出次数 18 で入れ替わる．実験で設定した最大出次数においては，128 で全ノードが SCC になった．



(a) Bow-tie 構造の要素の比率　　(b) 共起語グラフとの SCC 部分の比較

図 6 最大出次数の制約による生成グラフの変化

表 4 共起語グラフにおける Bow-tie 構造

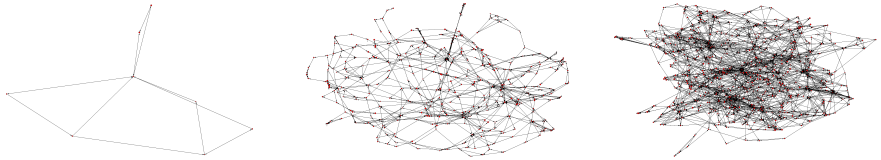
サブグラフ	ノード数	リンク数	密度	C	L	最長最短距離
SCC	17,166	891,855	0.0060532	0.7771475	2.49359	7
DCC	2,937	11	0.0000026	0.0000000	1.00000	1

共起語グラフの Bow-tie 構造を分析した結果および各サブグラフの特性を表 4 に示す．生成された関連語グラフの SCC サブグラフを，共起語グラフの SCC サブグラフと比較した結果を図 6(b) に示す．最大出次数ごとの SCC サブグラフのノード数，リンク数，密度を，共起語グラフの SCC サブグラフの値を 1 とした比率で示す．ノード数，リンク数ともに，最大出次数 24 以下において共起語グラフを下回る．密度は，最大出次数 10 以下ではノード数の増加にかかわらず一定である．最大出次数 32 以上では，ノード数，リンク数の増加に応じて密度も上昇する．SCC の圧縮という観点では，最大出次数 6 において，共起語グラフに対して密度は 100.2%を保ちつつ，ノード数は 5.7%，リンク数は 0.7%になった．

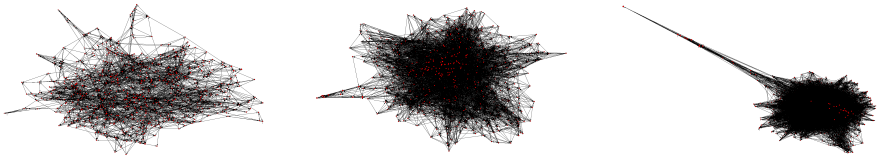
4.3.4 Bow-tie 構造を用いたグラフ分割と可視化

主な最大出次数における，Bow-tie 構造の SCC に属するノードのみからなるサブグラフを切り出して可視化した結果を図 7 に示す．可視化には，Pajek に実装されている 3 次元 FR アルゴリズムを用いた．最大出次数 11 では，SCC の拡大により，SCC 内部において

Vol.2009-DBS-149 No.21
2009/11/21



(a) 最大出次数 2　(b) 最大出次数 3　(c) 最大出次数 4



(d) 最大出次数 5　(e) 最大出次数 8　(f) 最大出次数 11

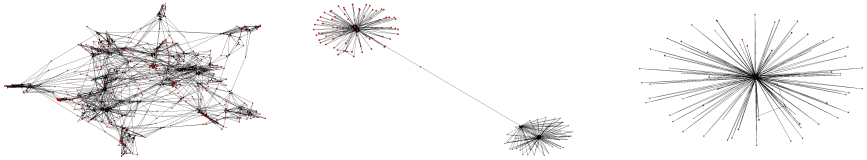
図 7　SCC サブグラフの可視化結果

距離感のあるノードが含まれている．これは，SCC が巨大になり，SCC の内部にもフラクタルに Bow-tie 構造が生じていることを意味する．

SCC および IN を除去したグラフを可視化した結果を図 9 に示す．最大出次数 3 では，OUT，Tubes，Tendrils のみからなるサブグラフにおいても，複数のクラスタが形成されている．外周部には，ビール，石油，航空，衛星放送に関連する語のクラスタが形成され，内周部には，不良債権，介護，保険のクラスタが形成された．中心部には，製紙業界再編に関連する語のクラスタが形成された．最大出次数が大きくなると，OUT，Tubes，Tendrils の部分におけるクラスタは消滅し，SCC に取り込まれていく．

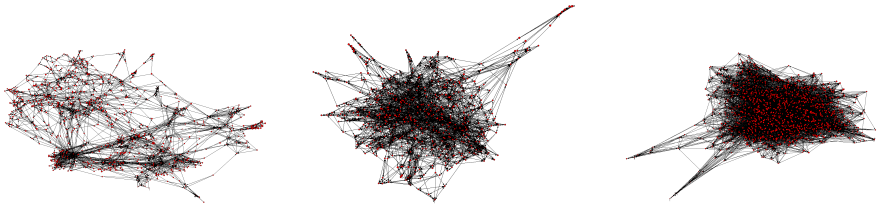
5. 考　　察

包括的な内容把握を支援するためのキーワード間ナビゲーションにおいては，関連性の高い語を精度よく提示することよりも，提示項目を次々に選択することで，文書空間内を広く飛び回れることが重要になる．提示内容の妥当性も検証する必要があるが，それ以前の課題として，グラフ構造そのものがナビゲーションに適した構造を持つことが要求される．



(a) 最大出次数 3　(b) 最大出次数 5　(c) 最大出次数 8

図 8　SCC および IN を除去したグラフの可視化結果



(a) 最大出次数 3　(b) 最大出次数 5　(c) 最大出次数 8

図 9　IN を除去したグラフの可視化結果

本節では，3 節で挙げたナビゲーションに適するグラフ構造の要件に対応して考察する．

5.1 Small-world 性

短い平均距離で多様なノードへ到達できるためには，グラフ全体が Small-world 性を示すことが大前提となる．提案手法において，UI 設計の観点から暫定的に設定した最大出次数 8 の場合に Small-world 性を示すだけでなく，最大出次数を変化させても同様に Small-world 性を示すことが確認できた．ただし，平均距離は到達可能ペアのみを用いて算出する．最大出次数が小さすぎると到達可能ペアが極めて少なくなるため，平均距離が短いとはいえ多様なノードへ到達できる構造にはならない．

また，同じ Small-world ネットワークにおいても，次数分布によって特性が大きく異な

る可能性が指摘されている．最大出次数を変化させたところ，次数分布はべき分布から対数正規分布へと連続的に変化した．どのような語の次数が，どのように変化するのか，今後さらに詳細に検討する必要がある．

5.2 リンクの多様性

多様なノードへ到達できるためには，リンクが多様である必要がある．リンクの多様性は，各ノードが持つリンクの類似性の低さ，すなわち密度の低さによって近似できると考えられる．しかし，ナビゲーションにおけるユーザビリティの観点から，最大出次数を制約する必要があるため，必然的に密度は大幅に低下する．密度だけでリンクの多様性を議論することはできない．

生成された関連語グラフの全体と SCC サブグラフにおいて，平均クラスタ係数および平均距離がほぼ一致しており，グラフ全体の Small-world 性が主に SCC によって生じていると推測できる．つまり，関連語グラフの SCC 部分において，共起語グラフにおける SCC 部分と同程度の密度を維持しながら，いかにノード数を削減するかという問題に置き換えることができると考えられる．

5.3 誘 導 性

ユーザを誘導する性質を高めるためには，どこにでも行けるグラフでなく，適切に行き止まりや一方通行のあるグラフが適すると考えられる．このためには，最大出次数を小さくすることとあわせ，ノード間に双方向のリンクを持つノードのペアを少なくすることが必要と考えられる．Bow-tie 構造における IN，OUT は，SCC との間で一方通行であることを意味する．また，Tendrils は行き止まりである．すなわち，SCC の比率を少なくし，SCC 以外の部分を多くすれば，どのような語を入口としても SCC に流入しやすい構造となる．

実験の結果，最大出次数 3 以上において，SCC と IN によって大部分が占められていた．最大出次数 24 以下において，SCC のノード数およびリンク数が共起語グラフよりも少なくなった．さらに，最大出次数 10 以下においては，SCC 部分の密度が共起語グラフの SCC 部分と同程度になった．データに依存していることも考えられるが，最大出次数 3 から 10 において，期待するグラフ構造が概ね得られているのではないかと考えられる．

6. おわりに

本稿では，文書空間から抽出した共起語から有向グラフを生成し，ユーザを文書空間内の関連文書に導く Comprehensive Web Navigation の一部分を構成する，関連語グラフについて，新聞記事を用いた実験により，生成されるグラフの Small-world 性，リンクの多様性，誘導性を評価した．最大出次数を変化させたところ，その値によらず Small-world 性を示すグラフが生成されることがわかった．リンクの多様性および誘導性については，Bow-tie 構造における SCC サブグラフの特性および全体に対する比率を調べた．その結果，最大出次数 24 以下において，変換前の共起語グラフよりも SCC ノードを削減しながら，SCC サブグラフの特性は共起語グラフと同等であることがわかった．特に，最大出次数 3 以上 10 以下の範囲においては，Comprehensive Web Navigation に適すると考えられるグラフ構造の要件を満たすグラフが生成できることがわかった．

今後は，関連語グラフの生成に用いる関連度スコアの妥当性およびグラフ特性への影響を精査するとともに，ユーザによる提示項目の選択しやすさとの関連も明らかにする必要がある．また，新聞記事よりも多様な語が出現すると考えられる CGM テキストに対しても本手法を適用する予定である．

参 考 文 献

1) Watts,D. and Strogatz,S.: Collective dynamics of 'small-world' networks, *Nature*, Vol. 393, No. 6684, pp. 440–442 (1998).
2) Broder, A. et al.: Graph structure in the Web, In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications networking,*, pp. 309–320 (2000).
3) White, R., Kales, B., Ducker, S. and Schraefel, M.: Supporting exploratory search, *Communications of the ACM*, Vol.49, No.4, pp.36–39 (2006).
4) 松尾豊, 大澤幸生, 石塚満：Small World 構造に基づく文書からのキーワード抽出, 情報処理学会論文誌, Vol.43, No.6, pp. 1825–1833 (2002).
5) 服部元, 原隆浩, 菅谷史昭, 西尾章治郎：クリック型 Web 検索のための重要語推定方式, データベースと Web 情報システムに関するシンポジウム (DBWeb Forum) 2007, 1A-3 (2007).
6) 酒井哲也, 小山田浩史, 野上謙一, 北村仁美, 梶浦正浩, 東美奈子, 野中由美子, 小野雅也, 菊池豊：クリックスルーに基づく探検型検索サイトの設計と開発, 第 7 回情報科学技術フォーラム (FIT 2008) 講演論文集, 第 2 分冊, pp. 1–4 (2008).
7) 武田善行, 梅村恭司：キーワード抽出を実現する文書頻度分析, 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 2001, No. 112, pp. 27–32 (2001).
8) 島田諭, 佐藤哲司：単語の反復度と共起頻度に基づく関連記事の提示方法, 情報処理学会 第 70 回全国大会, 講演論文集, 5S-1 (2008).
9) 島田諭, 福原知宏, 佐藤哲司：社会ネットワーク分析を用いた包括的 Web ナビゲーションの評価, Web とデータベースに関するフォーラム (WebDB Forum) 2008, 5A-2 (2008).