情報処理学会論文誌　データベース　Vol. 3　No. 2　111–122 (June 2010)

文書空間ナビゲーションのための出次数制約付き有向グラフ生成手法

島田　諭[†1]　福原知宏[†2]　佐藤哲司[†1]

探索的検索を円滑に行うには，検索対象の文書集合で特徴的に使われる用語および用語間の関連性を，探索過程でユーザが容易に把握できなければならない．本研究では，このような文書集合の特徴を反映し，多様な関連項目へ遷移でき，かつ提示項目数が抑制された文書空間ナゲーションの実現を目指す．このためには，Small-world性を示し，リンクの多様性が高く，誘導性が高い，有向グラフを生成する必要がある．本論文では，ナビゲーションにおける関連項目の提示数を意味する出次数を制約した，語の共起関係に基づく有向グラフの生成手法を提案し，上述の要件を満たす有向グラフを生成できる最大出次数の範囲を明らかにする実験を行った．その結果，最大出次数 3 以上 10 以下において，文書空間ナビゲーションに適するグラフ構造を生成できることが明らかになったので報告する．

A Method of Generating Directed Graphs with Out-degree Constraints for Document Navigation

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

In exploratory search, it is necessary that relativity between words in documents is clear to find documents effectively. The purpose of our study is to construct a navigation that reflects characteristics of document set, and that leads users to various related words and documents using reduced routes. The directed graph used for this navigation should have a small-world characteristic that guides a user, and high diversity of links. In this paper, we propose a generation method of directed graph that uses the co-occurrence relation of words, and that restricts a maximum out-degree as the number of related items to present in the navigation. We confirmed ranges of appropriate range of maximum out-degree based on an experiment. As a result, our method generates appropriate graphs having the maximum out-degree range from 3 up to 10.

1. はじめに

現在広く使われている検索語を入力する情報検索システムでは，ユーザは，検索結果を予想し，そこに出現する語の組合せで検索語を入力しなければならない．これまで情報検索が対象としてきた新聞記事，特許，法律などの専門的な文書集合においては，検索に適した統制語彙をユーザが事前に学習することで効率的な検索が可能だった．しかし，不特定多数のユーザが投稿する電子掲示板やブログなど，語彙が統制されない文書集合を対象とした検索では，実際に使われる用語をユーザが事前に把握することがきわめて困難になる．この場合，まず適当な文書を閲覧し，実際に使われる用語を把握しない限り，適切な検索語を決定することはできない．

このようなプロセスは探索的検索（exploratory search）[1)] と呼ばれるが，ユーザの入力に基づく試行錯誤では無駄が多く，探索の成否もユーザのスキルに依存する．探索的検索を支援するためには，検索対象の文書集合の特徴，すなわち文書集合において特徴的に使われる用語および用語間の関連性をユーザに提示し，ユーザは提示された項目を次々に選択していくだけで多様な文書に到達できるナビゲーション手法が有効と考えられる．

本論文では，検索語の入力なしで，ユーザが文書集合内を遷移可能とする文書空間ナビゲーションの実現を目的として，文書空間ナビゲーションの基盤となる有向グラフの生成手法を提案する．提案手法では，文書集合から特徴的な語を抽出し，それらの語の共起関係に基づいて，出次数を制約しながら，関連語間，関連文書間，および関連する語–文書間にリンクを生成する．本論文では，提案手法における語の抽出と重み付け，および有向グラフの生成方法について説明するとともに，提案手法を用いて文書空間ナビゲーションに適する有向グラフが生成できる最大出次数の範囲を，実験により明らかにする．

以下，2 章で本研究で実現すべき文書空間ナビゲーションの要件をあげ，関連研究との比較により本研究の位置付けを示す．3 章でナビゲーションに適する有向グラフの要件を定義し，提案する有向グラフ生成手法の詳細を述べる．4 章で，提案手法を用いてナビゲーションに適する有向グラフを生成できる最大出次数の範囲を明らかにするための評価実験について述べ，5 章で実験結果のナビゲーションにおける有用性について考察し，6 章でまとめる．

†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 産業技術総合研究所サービス工学研究センター
Center for Service Research, National Institute of Advanced Industrial Science and Technology

2. 背　　景

2.1 探索的検索におけるナビゲーション

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，探索的検索と呼ばれ，近年注目が高まっている[1)]．探索的検索においては，検索対象の文書集合の特徴，すなわち用語および用語間の関連性をユーザが容易に把握できることが重要である．

本研究では，探索的検索を支援する文書空間ナビゲーションの実現を目指す．このためには，文書集合の特徴を反映し，多様な関連項目へ遷移でき，かつ提示項目数が抑制されたナビゲーションが必要であると筆者らは考える．これらの要件について，以下に詳述する．

文書集合の特徴の反映

ユーザによる文書集合の特徴の把握においては，語の共起関係や出現頻度など，検索対象の文書集合から直接抽出した特徴量のみを用いることが有用と筆者らは考える．これは，たとえば検索対象とは別の文書集合から生成されたシソーラスやオントロジなどを外部知識として用いる場合，検索対象の文書集合と外部知識の間で，内容や粒度に差が生じやすいためである．

多様な関連項目への遷移

ユーザが文書集合の特徴を把握するまで探索を続行できるためには，文書集合中の任意の文書や語（以下，項目と総称する）を基点とし，検索語の入力なしで多様な関連項目へ遷移可能とする必要がある．多様な関連項目とは，基点となる文書との類似度が比較的低い文書，およびそのような文書中に出現する語である．類似度の高い関連項目と低い項目の混在提示が，項目間の関連性把握に有用であると筆者らは考える．また，検索語の入力を不要とすることは，ユーザが適切な検索語を想起する負担を軽減し，特に探索の初期段階での失敗を防止するために不可欠である．

提示項目数の抑制

ユーザが直感的に遷移先を選択できるためには，提示項目数の抑制が必要である．これは，多様な項目へ遷移できる経路を生成するには，文書集合全体を網羅するようにリンクを付与する必要があるが，1 度に提示する関連項目数が増えるにつれ視認性が低下し，ユーザによる選択が困難になるためである．

2.2 関 連 研 究

現在，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，いずれの手法においても，類似度の高い項目を提示する場合が多く，類似度の低い項目への遷移は困難である．このため，探索的検索において重要となる，項目間の関連性の把握という目的に対して，必ずしも有用ではない．また，ユーザによる探索可能性（findability）[2)] を空間構造によって保証することを意図した手法は提案されていない．

Google Suggest [⋆1] に代表される関連語提示手法では，検索結果の絞り込みに主眼が置かれている．ユーザが入力した検索語よりも具体的な語が提示されることが多いため，項目間の関連性の把握には必ずしも適さない．また，検索履歴や入力された検索語に基づいて推薦されるため，ユーザが予想しないような内容が提示されることは少ない．

若木らは，検索結果に含まれるトピックをクラスタリングし，検索語の曖昧性を解消する手法を提案している[3)]．この手法では，入力された検索語が多義性を持つ場合に，検索結果をトピックに分類し，個々のトピックを表現する語を提示する．ユーザは，検索結果を絞り込むだけでなく，提示された別のトピックに着目して再検索することもできる．

酒井らは，ユーザが気付きにくい関連情報の提示により，情報要求の変化の誘発を狙う検索インタフェースを提案している[4)]．この手法では，検索語の関連情報を Wikipedia [⋆2] から抽出し提示する．ただし，検索対象の Web ページと Wikipedia との間で内容や粒度に差があることから，必ずしも有効な提示ができるとは限らない．

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型端末やリモコンでの利用を念頭に，文字列の入力操作なしに検索を実行可能としている．ユーザは手軽に探索が続行できるが，適切な検索語をページ中からユーザ自身が選択する必要がある．

ユーザ・インタフェースの設計において，良好な操作性が得られる項目数については議論がある．Miller は，人間の短期記憶可能な項目数が 7 ± 2 であることを明らかにした[6)]．この知見に基づき，Web ページの設計において，ナビゲーションの項目数を 7 ± 2 にするデザイナは少なくない．一方，Larson らは，階層型ナビゲーションにおける 1 階層あたりの項目数を，32 にまで増やしても操作性が損なわれないことを報告している[7)]．

タグクラウド（tag cloud）[⋆3] に代表される関連語提示手法では，関連文書の要約を提示することに主眼が置かれている．このため，提示語数は数十程度と多く，各語間の類似度も高いため，ユーザによる選択は困難である場合が多い．

⋆1 http://www.google.com/
⋆2 http://ja.wikipedia.org/
⋆3 http://en.wikipedia.org/wiki/Tag_cloud

3. 文書空間ナビゲーションのための出次数制約付き有向グラフ生成手法

本論文では，探索的検索を支援するための文書空間ナビゲーションの要件をグラフ構造の特性により定義し，この要件を満たす有向グラフの生成手法を提案する．

3.1 節で，本研究で実現する文書空間ナビゲーションを概説し，3.2 節で，文書空間ナビゲーションに適するグラフ構造を定義し，3.3 節で，提案手法である出次数制約付き有向グラフ生成手法について詳説する．

3.1 探索的検索を支援するための文書空間ナビゲーション

本研究で実現する文書空間ナビゲーションの概要を**図 1** に示す．文書集合から語の共起関係を抽出し，関連語間，関連文書間，関連する語–文書間にリンクを生成し，任意の語または文書に対し，関連する語および文書を提示する．ここで生成される，語をノードとするグラフを関連語グラフ，文書をノードとするグラフを関連文書グラフ，語および文書をノードとし語–文書間のリンクを持つグラフを語–文書グラフと呼ぶ．いずれのグラフも，リンクの向きを考慮する有向グラフである．また，語–文書グラフは，語間と文書間にはリンクを持たない，2 部グラフ（bipartite graph）である．

文書集合の特徴をナビゲーションに反映するには，文書集合における語の共起関係が有用である．ここで，語の共起関係をリンクとする共起語グラフは無向グラフであるが，ナビゲーションにおいてはユーザを特定の方向へ誘導するために有向グラフを用いる必要がある．このため，共起語グラフの特性を可能な限り保持しながら，ナビゲーションに適する有向グラフに変換できる手法が必要となる．

多様な関連項目への遷移は，たとえばランダムな項目の提示でも実現できるが，これでは項目間の関連性の把握に有用ではない．語の共起関係においては，低頻度な共起であっても何らかの関連性があると見なせるため，共起語のうち文書頻度が低い語や，そのような語を含む文書へのリンクを優先して生成する方法が妥当である．

同時に提示する関連項目数，すなわち有向グラフの各ノードが持つ出次数は，次節で定義するナビゲーションに適する有向グラフの要件を満たす範囲において，可能な限り抑制する必要がある．このためには，リンク生成の優先度を決定する手法および適切な出次数制約（以下，値に着目する場合は最大出次数と記す）の付与が必要となる．

3.2 文書空間ナビゲーションに適するグラフ構造

本論文では，探索的検索を支援するための文書空間ナビゲーションの要件を，以下に示すグラフ構造の特性により定義する．すなわち，本論文の課題は，文書集合から抽出する語の共起関係に基づき，(1) Small-world 性を示し，(2) リンクの多様性が高く，(3) 誘導性が高い有向グラフを生成することである．以下，文書集合における文書および語に関して，グラフ構造に着目して言及する場合にはノードと記す．

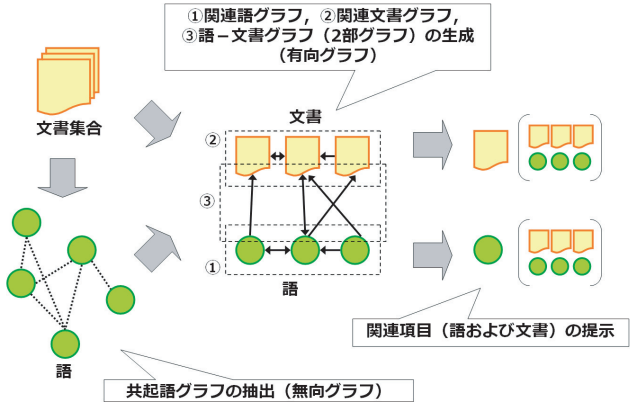


図 1　提案手法におけるグラフ生成の手順
Fig. 1　The procedures of generating graphs in proposal method.

(1) 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．Watts らは，グラフの Small-world 性を，ノード数およびリンク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している[8]．また，文書集合における語の共起関係をグラフ化した共起語グラフが一般に Small-world 性を示すことが明らかになっている[9),10)]．

共起語グラフでは，多数の文書に出現し共起語数が多い一般語がハブとなり，多数の項目間を短絡する．文書空間ナビゲーションにおいては，提示項目数の抑制が必要であることから，共起語グラフをそのまま利用することはできず，リンクを削減する必要がある．しかし，Small-world 性は，ユーザを多様なノードへ短い距離で遷移させるために有用な特性である．本論文では，共起語グラフが示す Small-world 性を，生成したグラフが維持していることを必要条件とする．本論文では，Watts らの定義[8] に基づき，平均距離がランダムグラフの 0.5 倍から 2 倍の範囲内にあり，平均クラスタ係数がランダムグラフより 10 倍以上大きく，かつ 0.1 以上である場合に Small-world 性を示すと見なす．

(2) リンクの多様性

リンクの多様性とは，任意の 2 ノードが同一ノードへのリンクを有する確率の低さを意味する．文書空間ナビゲーションにおいて，提示項目数を抑制しながら多様な項目へ遷移可能とするには，リンクの多様性が高いことが必要である．

本論文では，グラフの次数分布および密度に着目し，提案手法により生成される有向グラフにおけるリンクの多様性を評価する．次数分布が対数正規分布に近い，すなわち次数の高いノードが多いグラフでは，各ノードが同じノードへのリンクを持つ確率が高くなり，リンクの多様性は低くなる．一方，次数分布がベキ分布に近い，すなわち次数の低いノードが多いグラフでは，各ノードが同じノードへのリンクを持つ確率が低くなり，リンクの多様性は高くなる．また，密度が高いグラフでは，各ノードが同じノードへのリンクを持つ確率が高くなり，リンクの多様性は低くなる．本論文では，変換前のグラフと比較し，密度が低下し，ベキ分布に近い次数分布となる場合，リンクの多様性が高まったと見なす．

(3) 誘導性

誘導性とは，グラフの非循環性（acyclicity）によって生じる，ユーザを特定の方向へ誘導する性質である．複数のノードを経由し起点ノードに戻る経路をサイクル（cycle）と呼び，基点に戻らない経路をウォーク（walk）と呼ぶ．文書空間ナビゲーションにおいて，ユーザを類似度の低い関連項目へ誘導するためには，ユーザの遷移経路がサイクルとならずウォークとなるノードへ誘導することが必要である．ここで，ユーザによる経路選択確率に偏りがないと仮定すると，サイクルを形成するノードの比率を抑制すれば，ユーザの実際の遷移経路がサイクルになる確率も低下し，すなわち誘導性が高まると考えられる．

Broder らは，World Wide Web（WWW）のグラフを分析し，Bow-tie 構造と呼ばれる構造になることを明らかにした[11]．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．

本論文では，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誘導性を評価する．Bow-tie 構造における SCC はサイクルを形成し，IN，OUT，Tendrils，Tubes はウォークを形成する．OUT，Tendrils はウォークの終端，Tubes は OUT への短絡路であり，探索を続行させるという目的に合致しないため，可能な限り削減する必要がある．SCC は，グラフ全体の Small-world 性に寄与するため，Small-world 性が損なわれない程度に保持する必要がある．すなわち，変換前のグラフと比較し，SCC サブグラフの密度が同程度で，IN の比率が SCC の比率よりも高い場合，グラフの誘導性が高まったと見なす．

表 1　語の重み付けに用いる反復度および文書頻度の閾値および重み w
Table 1　Thresholds of df_2/df and df for weighting the words and w.

区分	反復度	文書頻度	重み w
I	$df_2/df \geq 0.6$	$df \geq 5$	10
II	$df_2/df \geq 0.35$，$df_2/df < 0.6$	$df \geq 10$，$df < 50$	1
III	$df_2/df \geq 0.1$	$df \geq 4$	0.1
	（ただし区分 I，II 以外）		
IV	すべて（ただし区分 I，II，III 以外）		0.01

$df < 3$ となる語は，事前に除く．

3.3　語の共起関係を用いた出次数制約付き有向グラフ生成手法

提案手法では，3.3.1 項に示す方法でリンク生成の優先度を決定し，3.3.2 項に示す方法で，3 種類の有向グラフ，すなわち関連語グラフ，関連文書グラフ，語–文書グラフを生成する．

本論文では，提案手法の核となる関連語グラフについて，詳細な評価を行う．

3.3.1　反復度および文書頻度を用いた語の重み付け

文書空間ナビゲーションにおいては，ユーザを多様な関連項目へ誘導する必要がある一方，提示項目数は抑制する必要がある．提示項目数すなわち探索の基点となる項目から関連項目へのリンク数を制約するためには，リンク生成の優先度を決定する必要がある．本手法では，語の反復度（df_2/df）および文書頻度 df を用いて語の重み付けを行い，リンク生成の優先度決定に用いる．df_2 は，その語が 2 回以上出現する文書数を意味する．

反復度および文書頻度は，キーワードや専門用語の抽出[12),13)]，自動要約手法[14)] において用いられる，語の特徴量を表す指標である．文書頻度（document frequency）は，文書の要点を表す語や，ある分野で共通して用いられる語において高くなることが知られている[14)]．反復度（adaptation）は，内容語において高くなることが知られており[15)]，日本語テキストを用いた実験でも同様の結果になることが報告されている[16)]．

本手法において，抽出された語に与える重み w，および重み付けの区分に用いる反復度および文書頻度の閾値を，**表 1** に示す．まず，2 文字以上の漢字またはカタカナからなる文字列，3 文字以上の英数字からなる文字列を抽出し，1 語とする．次に，文書集合における各語の反復度および文書頻度を求め，一致する区分の重み w を語の重みとする．区分との一

致は，区分 I，II，III，IV の順に評価する．

これらの閾値は，308 件のブログ記事を用いた予備実験により決定した[17]．反復度が高い領域には，単独でトピックを表現可能な具体的な語が含まれ，反復度が低い領域には，単独ではトピックを表現しにくい抽象的な語が多い．反復度に比例する重み付けも検討したが，小規模な文書集合では反復度の変動が大きいため，本手法では 4 区分の重み w を付与することとした．一方，文書頻度が高い領域には抽象的な語が多く，文書頻度が低い領域には具体的な語が多い．反復度が高く，かつ文書頻度が高い領域には，どのような文書にも出現する一般的な語が多く，このような語の重みを下げるため，文書頻度の閾値を組み合わせることとした．

3.3.2　語の共起関係を用いた有向グラフの生成

提案手法では，3 種類の有向グラフ，すなわち関連語グラフ，関連文書グラフ，語–文書グラフを生成する．前節で決定した各語の重み w を用いて，語の共起関係を持つ語間，文書間，語–文書間の関連度を算出し，これをリンク生成の優先度とする．この優先度の上位より一定数の項目に限定してリンクを生成することにより，探索の基点となる項目あたりの出次数を制約する．

関連語グラフ

式 (1) を用いて，基点となる語 t_i とその他の語 t_k の間の関連度 $r(t_i, t_k)$ を算出し，関連度 r の上位より一定数の語へのリンクを生成する．

$$r(t_i, t_j) = mw_j : i \neq j \tag{1}$$

ここで，基点となる語 t_i が出現する文書集合 D_i において出現する語を共起語 t_j とし，語 t_i と t_j が共起する文書集合を $D_{ij} = \{d_1, ..., d_m\}$ とする．語 t_j の重み w_j に，文書集合 D_{ij} の文書数 m，すなわち語 t_i と t_j の共起文書数を乗算する．基点となる語 t_i の重みは考慮しないことで，2 語間の関連度を非対称とする．

関連文書グラフ

式 (2) を用いて，基点となる文書 d_i とその他の文書 d_j 間の関連度 $r(d_i, d_j)$ を算出し，関連度 r の上位より一定数の文書へのリンクを生成する．

$$r(d_i, d_j) = \sum_{k=1}^{n} w_k : i \neq j \tag{2}$$

ここで，基点となる文書 d_i と，文書 d_j との間で共起する語集合を $T_{ij} = \{t_1, ..., t_n\}$ とする．語集合 T_{ij} に含まれる語 t_k の重み w_k の総和を，T_{ij} の共起により生じる 2 文書間の関連度と見なす．

語–文書グラフ

語–文書グラフは，文書から語へのリンクを有するグラフと，語から文書へのリンクを有するグラフで構成する．

まず，式 (3) を用いて，基点となる文書 d_i と，文書 d_i に出現する語 t_j 間の関連度 $r(d_i, t_j)$ を算出し，関連度 r の上位より一定数の語へのリンクを生成する．

$$r(d_i, t_j) = \sum_{k=1}^{n} m_k w_k \tag{3}$$

ここで，基点となる文書 d_i において出現する語集合 $T_i = \{t_1, ..., t_n\}$ とする．語集合 T_i に含まれる語 t_k の重み w_k に，語 t_k の文書頻度 m_k を乗算した値の総和を，T_i の出現により生じる文書–語間の関連度と見なす．

次に，式 (4) を用いて，基点となる語 t_i と，語 t_i が出現する文書 d_j 間の関連度 $r(t_i, d_j)$ を算出し，関連度 r の上位より一定数の文書へのリンクを生成する．

$$r(t_i, d_j) = \sum_{k=1}^{n} w_k \tag{4}$$

ここで，t_i が出現する文書集合 $D_i = \{d_1, ..., d_m\}$ に含まれる文書 d_j において出現する語集合を $T_j = \{t_1, ..., t_n\}$ とする．語集合 T_j に含まれる語 t_k の重み w_k の総和を，T_j の出現により生じる語–文書間の関連度と見なす．基点となる語 t_i の重みは考慮しない．

4. 評　　価

4.1　実験の概要

本論文では，文書空間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，提案手法により生成されるグラフの特性との関連を明らかにし，最適な最大出次数の範囲を明らかにする．このため，3.2 節であげた，文書空間ナビゲーションに適するグラフ構造の 3 要件に対応し，以下に示す 3 項目の実験を行う．また，共起語グラフの次数分布が異なる 2 種類の文書集合を用い，その影響の有無を調べる．

(1) Small-world 性

本実験の目的は，Small-world 性を示すグラフを生成できる最大出次数の範囲を明らかにすることである．最大出次数を変えて関連語グラフを生成し，ノード数およびリンク数が同

数のランダムグラフと，平均距離 L，平均クラスタ係数 C を比較する．変換前のグラフである共起語グラフについても同様に示す．参考として，関連文書グラフ，語–文書グラフ，および 3 種類の生成グラフの結合グラフについて，最大出次数を 8 とした場合の結果を示す．

(2) リンクの多様性

本実験の目的は，変換前のグラフと比較し，密度が低下し，べキ分布に近い次数分布となるグラフを生成できる最大出次数の範囲を明らかにすることである．最大出次数を変えて生成した関連語グラフを，変換前の共起語グラフと比較し，最大出次数による次数分布および密度の変化を分析する．

(3) 誘導性

本実験の目的は，変換前のグラフと比較し，SCC サブグラフの密度が同程度で，IN の比率が SCC の比率よりも高くなるグラフを生成できる最大出次数の範囲を明らかにすることである．最大出次数を変えて生成した関連語グラフにおける Bow-tie 構造を抽出し，変換前の共起語グラフと比較し，最大出次数による変化を分析する．

ネットワーク分析指標の定義

本論文で用いるネットワーク分析指標の定義を以下に示す[18]．

平均距離 L（average path length）は，式 (5) を用いて算出する．

$$L = \frac{1}{N(N-1)/2} \sum_{i>j} L_{ij} \tag{5}$$

ここで，N はグラフが持つノードの総数，L_{ij} は，任意のノード v_i，v_j 間における最短距離，すなわち最短経路上に存在するリンク数である．なお，L_{ij} の最大値を最長最短距離と呼ぶ．

平均クラスタ係数 C（average clustering coefficient）は，式 (6) を用いて算出する．

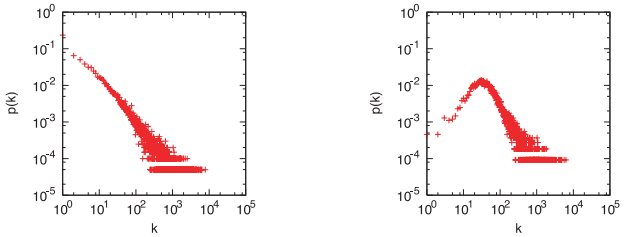
$$C = \frac{1}{N} \sum_{i=1}^{N} \frac{2E_i}{k_i(k_i-1)} \tag{6}$$

ここで，k_i は，ノード v_i が持つリンク数，E_i は，ノード v_i が属するクラスタ数，N はグラフが持つノードの総数である．クラスタとは，3 個のノードがそれぞれ 1 本のリンクによりつながっていることをいう．有向グラフにおいては，リンクの向きを無視して算出する．

有向グラフの密度 D（density）は，式 (7) を用いて算出する．

表 2　実験に用いるデータの概要
Table 2　Outline of data set used for evaluation.

データ名	文書数	抽出語数
朝日新聞	7,770	20,103
Yahoo!知恵袋	39,914	10,899



(a) 朝日新聞　(b) Yahoo!知恵袋

図 2　共起語グラフの次数分布
Fig. 2　Degree distribution in co-occurrence word graph.

$$D = \frac{k}{N(N-1)} \tag{7}$$

ここで，k はグラフが持つリンクの総数，N はノード数である．無向グラフの密度は，リンク数を $2k$ として算出する．

各指標の算出，ランダムグラフの生成には，ネットワーク分析ツール Pajek [⋆1] を用いた．

4.2　実験に用いるデータ

本論文では，以下に示す 2 種類のデータを用いた．各データの文書数，および提案手法による抽出語数を，**表 2** に示す．「朝日新聞」は，「朝日新聞記事データ集 学術研究用」1996 年版に含まれる，「1 経済」「2 経済」「3 経済」の各面に掲載された全記事である．「Yahoo! 知恵袋」は，ヤフー株式会社が国立情報学研究所に提供したデータに含まれる，2005 年 9 月分の「パソコン、周辺機器」カテゴリの質問および回答の全記事である．

各データから抽出した共起語グラフの次数分布を，**図 2** に示す．k はノードの次数，$p(k)$ は次数が k のノードの正規化頻度である．「朝日新聞」ではベキ分布に従うが，「Yahoo!知

⋆1 http://pajek.imfm.si/doku.php

表 3　「朝日新聞」における語の区分結果および df 上位語
Table 3　Result of division of words and a list of high-ranked word on df.

区分 I（477 語）			区分 II（433 語）			区分 III および IV（19,193 語）		
語	df	df_2/df	語	df	df_2/df	語	df	df_2/df
米ドル	254	0.98	法案	49	0.37	日本	1,510	0.47
NTT	224	0.65	発効日	48	0.35	発表	1,428	0.13
APEC	125	0.63	ゲーム	48	0.35	情報ファイル	1,201	0
監査役	112	0.69	日本石油	46	0.37	米国	1,021	0.40
ソニー	107	0.62	制裁	46	0.37	必要	938	0.27

表 4　「Yahoo!知恵袋」における語の区分結果および df 上位語
Table 4　Result of division of words and a list of high-ranked word on df.

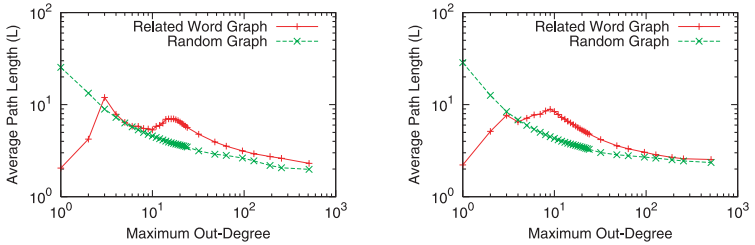
区分 I（2,740 語）			区分 II（1,208 語）			区分 III および IV（6,951 語）		
語	df	df_2/df	語	df	df_2/df	語	df	df_2/df
quot	394	0.91	Java	49	0.45	パソコン	3,944	0.20
Sub	25	0.96	ヘッド	37	0.43	場合	2,927	0.21
ttfCache	24	0.83	ハイパーリンク	37	0.38	ソフト	2,526	0.18
Select	15	0.60	LAME	36	0.44	出来	2,332	0.14
入力	12	0.83	台目	35	0.37	表示	2,306	0.26

恵袋」では対数正規分布に従う．共起語グラフの次数分布は，記事の内容のばらつきや語彙統制の有無など，文書集合の特徴を反映していると考えられる．このため，上述の 2 種類の文書集合を用いて実験し，文書集合の特性が提案手法に与える影響を検討する．

提案手法による語の区分結果，df 上位語と，その df，df_2/df の値を**表 3**，**表 4** に示す．「朝日新聞」では，区分 I には，固有名詞や役職名など，単独で主題を表現可能な具体的な語が含まれる．区分 II には，「法案」「増税」など，他の語と組み合わせると主題を表現可能な語が含まれる．区分 III，IV には，「日本」「米国」などの高頻度語が含まれる．「Yahoo!知恵袋」では，関数名やファイル名などが区分 I に含まれる．区分 II に含まれる「台目」は，「PC の購入が 1 台目か 2 台目かで選択基準が変わる」という文脈で出現する語である．

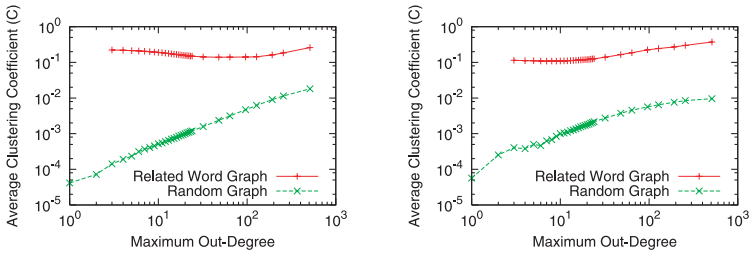
4.3　実 験 結 果

本節では，3.2 節であげた，文書空間ナビゲーションに適するグラフ構造の 3 要件に対応し，実験結果を以下に詳述する．



(a) 朝日新聞　(b) Yahoo!知恵袋

図 3　最大出次数の制約による平均距離 L の変化（関連語グラフとランダムグラフの比較）
Fig. 3　Transition of average path length (L) by restriction of maximum out-degree (Comparison between related word graphs and random graphs).



(a) 朝日新聞　(b) Yahoo!知恵袋

図 4　最大出次数の制約による平均クラスタ係数 C の変化（関連語グラフとランダムグラフの比較）
Fig. 4　Transition of average clustering coefficient (C) by restriction of maximum out-degree (Comparison between related word graphs and random graphs).

(1) Small-world 性

提案手法により生成されるグラフが Small-world 性を示すことを確認するため，生成グラフと，ノード数およびリンク数が同数のランダムグラフとの間で，平均距離 L および平均クラスタ係数 C を比較した．

生成グラフの L および C を，設定した最大出次数[⋆1]ごとに，生成グラフとノード数およびリンク数を同数としたランダムグラフと比較した結果を**図 3**，**図 4** に示す．横軸は最大出

⋆1 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，32，48，64，96，128，192，256，512，768，1,024，1,536，2,048，3,072，4,096 の 38 通りに設定した．

表 5　共起語グラフとランダムグラフの比較
Table 5　Comparison between co-occurrence word graphs and random graphs.

データ名	ノード数	リンク数	C	L	最長最短距離
朝日新聞	20,103	891,866	0.6636081	2.494	7
（ランダムグラフ）	（同上）	（同上）	0.0022152	2.920	4
Yahoo!知恵袋	10,899	810,559	0.5701085	2.094	5
（ランダムグラフ）	（同上）	（同上）	0.0068396	2.591	3

次数，縦軸が L および C である．なお，最大出次数 1 ではクラスタが形成されないため，C は 0 になる．実験に用いた「朝日新聞」と「Yahoo!知恵袋」のいずれのデータにおいても，L は，最大出次数 3 以上においてランダムグラフと同程度となった．C は，最大出次数 2 以上において，ランダムグラフに比べて 10 倍以上大きく，かつ 0.1 以上である．このことから，最大出次数 3 以上において，生成グラフは Small-world 性を示した．

提案手法では，文書集合における共起語グラフが Small-world 性を示すことを前提としているため，対象データにおいて共起語グラフが Small-world 性を示すかを調べた．

各データについて，共起語グラフと，ノード数およびエッジ数が同数のランダムグラフを生成し，比較した結果を**表 5** に示す．いずれのデータにおいても，ランダムグラフと比較し，L が同程度で，C が 10 倍以上大きく，かつ 0.1 以上であり，いずれのグラフも Small-world 性を示した．

本研究における文書空間ナビゲーションでは，関連語グラフとともに，関連文書グラフ，語–文書グラフも生成する．参考として，最大出次数を 8 として生成した各グラフの特性を**表 6**，**表 7**，**表 8** に示す．また，関連語グラフと関連文書グラフを語–文書グラフを用いて結合した結合グラフの特性を**表 9** に示す．語–文書グラフはクラスタを形成しない 2 部グラフであるため，Small-world 性の評価は行わない．最大出次数を 8 として生成した関連語グラフおよび関連文書グラフは，いずれのデータにおいても，ランダムグラフと比較し，L が同程度で，C が 10 倍以上大きく，Small-world 性を示した．

平均距離を共起語グラフと比較すると，関連語グラフおよび関連文書グラフでは同程度であるのに対し，結合グラフでは 3 倍程度延びている．これは，語–文書グラフによって，関連語グラフおよび関連文書グラフではリンクのなかったノード間にもリンクが追加されるためである．結合グラフで最長最短距離が延びているのも，このためである．

(2) リンクの多様性

最大出次数とリンクの多様性の関係を明らかにするため，提案手法により生成された関連語グラフの次数分布および密度 D を調べた．

表 6　最大出次数 8 における関連語グラフとランダムグラフの比較
Table 6　Comparison between related word graphs and random graphs when out-degree is 8.

データ名	ノード数	リンク数	C	L	最長最短距離
朝日新聞	20,103	160,824	0.1990513	5.521	18
（ランダムグラフ）	（同上）	（同上）	0.0003832	5.001	9
Yahoo!知恵袋	10,889	86,836	0.1087435	8.550	26
（ランダムグラフ）	（同上）	（同上）	0.0007179	4.711	8

表 7　最大出次数 8 における関連文書グラフとランダムグラフの比較
Table 7　Comparison between related document graphs and random graphs when out-degree is 8.

データ名	ノード数	リンク数	C	L	最長最短距離
朝日新聞	7,770	62,136	0.6016123	2.033	6
（ランダムグラフ）	（同上）	（同上）	0.0009685	4.548	8
Yahoo!知恵袋	39,914	302,000	0.4076854	2.698	12
（ランダムグラフ）	（同上）	（同上）	0.0001806	5.468	10

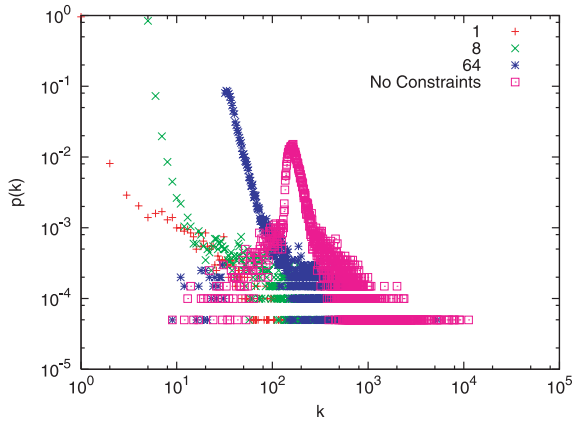
表 8　最大出次数 8 における語–文書グラフの特性
Table 8　Characteristic of word-document bipartite-graphs when out-degree is 8.

データ名	ノード数	リンク数		C	L	最長最短距離
		語→文書	文書→語			
朝日新聞	27,873	113,543	61,091	0	7.532	18
Yahoo!知恵袋	50,803	61,993	216,324	0	6.730	18

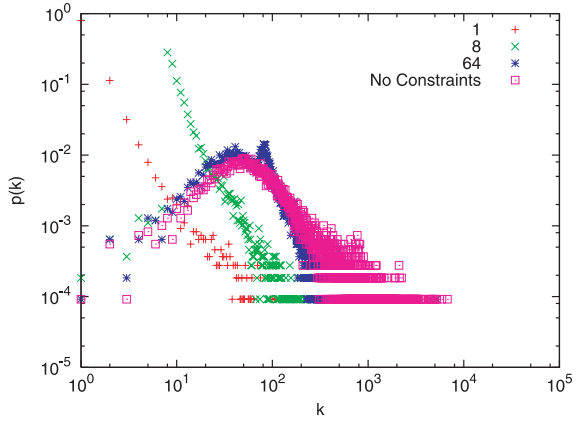
表 9　最大出次数 8 における結合グラフとランダムグラフの比較
Table 9　Comparison between united graphs and random graphs when out-degree is 8.

データ名	ノード数	リンク数	C	L	最長最短距離
朝日新聞	27,873	397,594	0.1520279	6.056	15
（ランダムグラフ）	（同上）	（同上）	0.0004968	4.120	7
Yahoo!知恵袋	50,803	667,149	0.1674282	5.554	13
（ランダムグラフ）	（同上）	（同上）	0.0002597	4.503	7

まず，主な最大出次数における次数分布を**図 5** に示す．横軸の次数 k は出次数と入次数の合計，縦軸の $P(k)$ は，次数が k であるノードの比率である．最大出次数が小さいほどベキ分布に近く，大きいほど対数正規分布に近くなった．



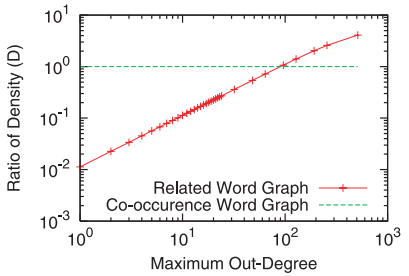
(a) 朝日新聞



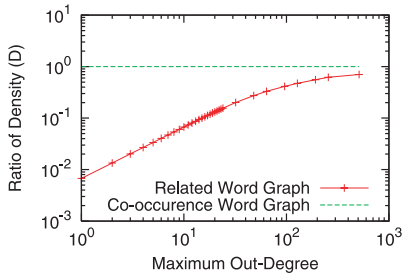
(b) Yahoo!知恵袋

図 5　最大出次数の制約による次数分布の変化（関連語グラフ）
Fig. 5　Transition of degree distribution by restriction of maximum out-degree (Related word graphs are shown).

ただし，用いたデータによって，最大出次数を大きくした場合の次数分布に差が現れた．最大出次数 8 では，いずれのデータでも出次数 8 未満のノードはほとんど存在せず，提案手法により各ノード（語）が最大出次数個の関連語へのリンクを付与された状態になってい



(a) 朝日新聞



(b) Yahoo!知恵袋

図 6　最大出次数の制約による密度 D の変化（関連語グラフと共起語グラフの比較）
Fig. 6　Transition of density (D) by restriction of maximum out-degree (Comparison between related word graphs and co-occurence word graphs).

る．最大出次数 64 では，「朝日新聞」では次数 k が 64 未満のノードは少ないが，「Yahoo!知恵袋」では多いという差がある．これは，「Yahoo!知恵袋」では共起語数が少ない語が多く，最大出次数に満たない少数のリンクのみが生成されたノードが多いことを意味する．

次に，密度 D を共起語グラフと比較した結果を**図 6** に示す．横軸に最大出次数，縦軸に最大出次数ごとの D を，共起語グラフの値を 1 とした比率で示す．「朝日新聞」では最大出次数 64 以下において，「Yahoo!知恵袋」では設定したすべての最大出次数において，共起語グラフを下回った．「朝日新聞」では最大出次数 96 以上において，密度が共起語グラフの密度を上回っているが，これは，最大出次数を大きくするほど，多くの語と共起する一般語が多くのリンクを持つためである．同様のことが「Yahoo!知恵袋」では起きないのは，1 文書あたりの語数が少なく，共起語の組合せ数が「朝日新聞」と比べ多くないためである．

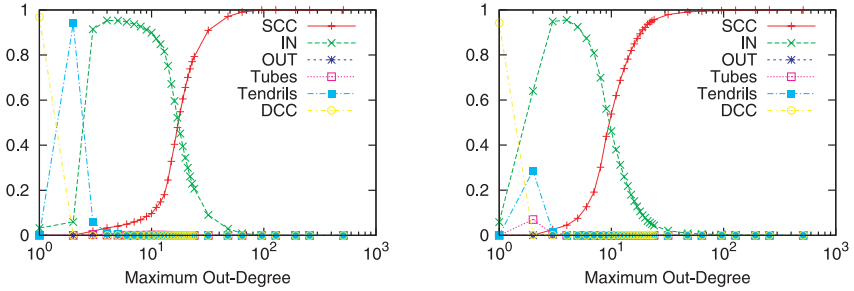
これらの結果から，データの特性に影響されず，生成される関連語グラフの次数分布がベキ分布に近くなり，かつ共起語グラフの密度に対する密度 D の比率が 0.1 以下となる最大出次数の範囲は，1 から 10 程度となった．

(3) 誘導性

最大出次数とグラフの誘導性の関係を明らかにするため，提案手法により生成された関連語グラフの Bow-tie 構造を調べた．

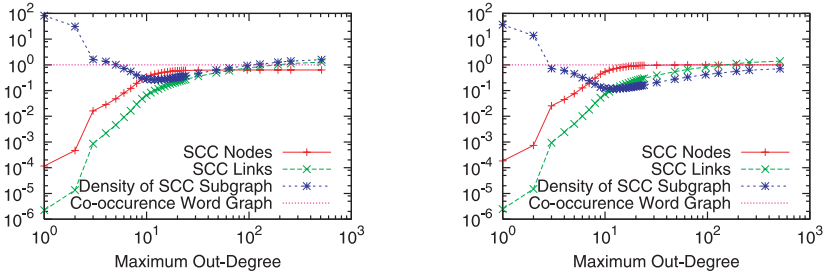
まず，最大出次数の制約による Bow-tie 構造の要素（ノード）の比率の変化を，**図 7** に示す．横軸は最大出次数，縦軸はノード総数に対する各要素の比率である．

いずれのデータにおいても，最大出次数 1 では DCC，最大出次数 2 では IN もしくは



(a) 朝日新聞　(b) Yahoo!知恵袋

図 7　最大出次数の制約による Bow-tie 構造の変化（関連語グラフ）
Fig. 7　Transition of bow-tie structure by restriction of maximum out-degree (Related word graphs are shown).



(a) 朝日新聞　(b) Yahoo!知恵袋

図 8　最大出次数の制約による SCC サブグラフの変化（関連語グラフと共起語グラフの比較）
Fig. 8　Transition of SCC subgraphs by restriction of maximum out-degree (Comparison between related word graphs and co-occurence word graphs).

Tendrils が大部分を占めた．最大出次数 3 において SCC が増加しはじめ，SCC を除くほとんどのノードが IN になる点が共通する．これは，各ノードが最低 2 本のリンクを持てばクラスタを形成でき，3 本以上になると複数のクラスタに属すことが可能となるためである．一方，IN の比率は最大出次数 4 をピークに減少し，「朝日新聞」においては最大出次数 18 で，「Yahoo!知恵袋」においては最大出次数 10 において，SCC と比率の大小が逆転した．これは，各ノードの次数が高くなるほど，他の SCC ノードとの間でクラスタを形成し，自ノードも SCC ノードになる確率が高まるためである．

次に，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 8** に示す．SCC サブグラフとは，SCC ノードのみからなる部分を切り出した部分グラフである．横軸は最大出次数，縦軸に，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 とした比率を示す．

SCC サブグラフのノード数およびリンク数は，「朝日新聞」では最大出次数 24 以下，「Yahoo!知恵袋」では最大出次数 128 以下において，共起語グラフを下回る．SCC サブグラフの密度は，いずれのデータでも最大出次数 2 以下では極端に大きくなっている．これは，上述のように，最大出次数が 3 以上でないと SCC ノードが安定して形成されないためである．最大出次数 3 以上の範囲に着目すると，「朝日新聞」では最大出次数 10 以下で SCC ノード数の増加にかかわらず一定であり，最大出次数 32 以上では，SCC ノード数およびリンク数の増加に応じて上昇する．「Yahoo!知恵袋」では最大出次数 12 までは低下し，13 から上昇に転じるが，最大出次数を増加させても共起語グラフを上回ることはなかった．

一例として，最大出次数 8 では，「朝日新聞」では密度は 94.6%，ノード数は 8.1%，リンク数は 1.2%になった．「Yahoo!知恵袋」では，密度は 17.8%，ノード数は 30.1%，リンク数は 3.2%になった．

これらの結果から，データの特性に影響されず，生成される関連語グラフにおける SCC サブグラフの密度が共起語グラフと同程度になり，生成される関連語グラフにおける IN の比率が SCC の比率よりも高くなる最大出次数の範囲は，3 から 10 となった．

5. 考　　察

本論文における評価実験の結果について，文書空間ナビゲーションにおける有用性の観点から考察する．

4.3 節で詳述したように，提案手法により生成された関連語グラフは，(1) Small-world 性は最大出次数 3 以上の全域で満たし，(2) リンクの多様性は最大出次数 1 以上 10 以下の範囲で満たし，(3) 誘導性は最大出次数 3 以上 10 以下の範囲で満たしていた．すなわち，最大出次数 3 から 10 程度において，3.2 節であげた，ナビゲーションに適するグラフ構造の要件を満たす有向グラフが生成できることが分かった．

このことから，ユーザ・インタフェース設計のうえで重要な要件である，ユーザへの関連項目の提示数を，3 から 10 程度に固定できるという示唆が得られた．筆者らは，提案手法を実装したプロトタイプシステムを構築し，ユーザ実験を進めているが，このシステムにお

いては，項目の見やすさを考慮して最大出次数を 8 としている[19]．本論文における実験により，プロトタイプシステムにおいて設定した最大出次数が，グラフ構造の上でも妥当な値の範囲にあることが確認できた．

文書集合の特性の影響については，共起語グラフの次数分布が異なる 2 種類のデータを用いて調べたが，最大出次数が小さい範囲においてはほとんど影響を受けないことが分かった．上述のプロトタイプシステムでは，本論文で用いた文書集合に加え，文書集合の規模が大きく異なる文書集合も用いているが，共起語グラフの次数分布に表れるような特性の異なりよりも，文書集合の規模の影響が大きいという示唆が得られている．

本論文における Small-world 性の評価では，最長最短距離や到達可能ノード対の比率については考慮していない．ユーザが多様な関連項目へアクセスできるためには，到達可能ノード対を増やす必要があると考えられる．しかし，到達可能ノード対を増やすためには，各ノードがより多くのクラスタに属する必要があり，これは SCC の比率が高くなることを意味する．同時に最長最短距離も長くなり，ユーザによる経路選択が困難になる恐れがある．関連語グラフおよび関連文書グラフを，2 部グラフである語–文書グラフを用いて結合した結合グラフにおいては，表 9 に示したように，結合前のグラフ単体と比べ，平均距離および最長最短距離が長くなっている．実際の文書空間ナビゲーションにおいては，ユーザは語と文書を区別して経路選択ができることから，距離が長くなっても適切な経路選択ができる可能性がある．

6. おわりに

本論文では，多様な文書へのアクセスを支援するナビゲーションの基盤となる，語の共起関係を用いた有向グラフの生成手法を提案した．提案手法は，文書集合から抽出される共起語グラフが示す Small-world 性を応用し，文書間，キーワード間，文書–キーワード間にリンクを生成し，文書空間ナビゲーションのための有向グラフを生成する手法である．

探索的検索の支援を目的とする文書空間ナビゲーションにおいては，ユーザの負担を軽減するため，同時に提示する関連項目数の抑制，すなわちグラフ構造におけるノードあたりの最大出次数の制約が必要である．本論文では，文書空間ナビゲーションに適するグラフ構造の要件として，(1) Small-world 性，(2) リンクの多様性，(3) 誘導性を定義した．最大出次数の制約を変化させて生成したグラフの特性を分析し，最大出次数 3 から 10 程度において，上述の要件を満たす有向グラフを生成できることを確認した．また，共起語グラフにおいて異なる次数分布を示す 2 種類の文書集合を用いて実験し，最大出次数が小さい範囲では，文書集合の特性の違いが提案手法に与える影響が小さいことを確認した．

今後，本論文では詳細な評価は行わなかった関連文書グラフおよび語–文書グラフの評価，およびユーザ実験による評価を行う予定である．

謝辞　本研究の一部は科研費（21500091）の助成を受けたものである．本研究の実装・評価に際し，大学共同利用機関法人国立情報学研究所から提供を受けた，Yahoo!知恵袋のデータを利用している．ここに記して謝意を示す．

参考文献

1) White, R., Kales, B., Ducker, S. and Schraefel, M.: Supporting exploratory search, *Comm. ACM*, Vol.49, No.4, pp.36–39 (2006).
2) Morville, P.: *Ambient Findability: What We Find Changes Who We Become*, O'Reilly Media, Inc. (2005).
3) 若木裕美，正田備也，高須淳宏，安達　淳：検索語の曖昧性解消のためのトピック指向単語抽出および単語クラスタリング，情報処理学会論文誌　データベース，Vol.47, No.19, pp.72–85 (2006).
4) 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).
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，No.1A-3 (2007).
6) Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information., *Psychological review*, Vol.63, No.2, pp.81–97 (1956).
7) Larson, K. and Czerwinski, M.: Web page design: implications of memory, structure and scent for information retrieval, *CHI '98: Proc. SIGCHI conference on Human factors in computing systems*, pp.25–32 (1998).
8) Watts, D. and Strogatz, S.: Collective dynamics of 'small-world' networks, *Nature*, Vol.393, No.6684, pp.440–442 (1998).
9) Ferrer, R. and Sole, R.V.: The small world of human language, *Proc. The Royal Society of London. Series B, Biological Sciences*, Vol.268, pp.2261–2265 (2001).
10) 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).
11) Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J.: Graph structure in the Web, *Proc. 9th international World Wide Web conference on Computer networks: the international journal of*

computer and telecommunications networking, pp.309–320 (2000).
12) 中川裕志，湯本紘彰，森　辰則：出現頻度と連接頻度に基づく専門用語抽出，自然言語処理，Vol.10, No.1, pp.27–45 (2003).
13) 白井　智，鳥井　修，金井達徳：反復文字列階層グラフによる文書からのキーワード自動抽出，日本データベース学会 letters，Vol.4, No.1, pp.77–80 (2005).
14) 小峰　恒，山田剛一，絹川博之，中川裕志：文書頻度と節長を利用した図書概要縮約方式，*NII journal*, Vol.8, pp.23–33 (2004).
15) Umemura, K. and Church, K.W.: Empirical term weighting and expansion frequency, *Proc. 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pp.117–123 (2000).
16) 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).
17) 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).
18) 安田　雪：実践ネットワーク分析，新曜社 (2001).
19) 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).

(平成 21 年 12 月 20 日受付)
(平成 22 年　2 月 10 日採録)

（担当編集委員　太田 学）

島田　　諭（学生会員）

2004 年東洋大学社会学部メディアコミュニケーション学科卒業．2009 年筑波大学大学院図書館情報メディア研究科博士前期課程修了．同年より同研究科博士後期課程に在籍．グループウェア，テキストマイニング，Web ナビゲーションに関する研究に従事．日本データベース学会学生会員．

福原　知宏（正会員）

2003 年奈良先端科学技術大学院大学情報科学研究科博士後期課程単位取得認定退学．工学博士．科学技術振興機構社会技術研究開発センター研究員，東京大学人工物工学研究センター特任助教を経て，2010 年産業技術総合研究所サービス工学研究センター特別研究員．Web テキストマイニング，スパムフィルタリングの研究に従事．

佐藤　哲司（正会員）

1980 年山梨大学工学部電子工学科卒業．同年日本電信電話公社武蔵野電気通信研究所に入所．以来，論理回路の大規模一括集積技術，データベースマシン，マルチメディアデータベースの研究・開発に従事．1994 年工学博士（大阪大学）取得．2007 年 4 月より現職．情報検索，社会インタラクションに興味を持つ．電子情報通信学会会員．