社会ネットワーク分析指標を用いた包括的Webナビゲーションの実現と評価

Evaluation of the Comprehensive Web Navigation System Using Social Network Analysis

島田 諭[*1] Satoshi SHIMADA
福原 知宏[*2] Tomohiro FUKUHARA
佐藤 哲司[*1] Tetsuji SATOH

[*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

We discuss the characteristics of terms that is suitable for web navigation by using network centrality measures. Based on the user behavior in our system, we show that the words presented by our system is more appropriate than the words inputted by users.

1. はじめに

コミュニティの内部における情報共有を主眼としたSNSや電子掲示板においては，用字用語の独自性や省略の多さによって，途中から参加するユーザにとって内容の把握が困難であることが多い．従来の情報検索では，ユーザにとって未知である内容に到達することは難しく，ユーザに対してクエリの入力を明示的に要求せずに，文書集合の内部をナビゲーションできる手法が求められている．

我々は，トピック間遷移に適した特徴語を文書集合から抽出し，多様な遷移経路をユーザに提示する，包括的Webナビゲーションを提案している．ナビゲーションに用いる特徴語には，従来の情報検索と比較し，特定性（トピックを特定する性質）または網羅性（トピック間を橋渡しする性質）のいずれかが顕著であることが求められる．本論文では，文書集合から生成した共起語グラフにおける次数中心性および媒介中心性の指標を用いて，提案手法により抽出された特徴語と，ユーザが入力した検索語を比較する．

2. 包括的Webナビゲーション

2.1 用字用語の独自化と検索困難性

Web上では，特定の地域に居住するユーザや，特定の分野に関心を持つユーザが，時間的，地理的な制約を受けずに集まることが容易である．何らかの共通の関心を持つユーザが，SNSやブログ，電子掲示板等を利用して情報を共有することで，一定のコミュニティが形成されていく．このようなコミュニティにおいては，参加者間で既に共有されている内容は省略される傾向が見られる．時間の経過とともに，より多くの予備知識を読者が持っていることが期待されるようになり，予備知識を持たないユーザにとって内容の把握が困難になる．

このような，読者側に一定の予備知識があることを前提とする表現により，予備知識を有しないユーザによる内容把握が困難になる問題は，どのような文書集合においても生じる一般的な問題である．多くの読者を想定する一般紙においても，大きな事件や事故に関する第一報では，前提となる知識の解説にも多くの紙面が割かれるが，続報では第一報に対する差分となる新事実の記述が中心となり，前提となる知識は既に共有されていることが前提とされる．いわば，現在進行形ですべての記事をシーケンシャルに受け取る読者を念頭に書かれており，ポータルサイト上で個別の記事を断片的に参照するユーザや，記事データベースの検索によりランダムな順序で閲覧する将来のユーザにとっての可読性はほとんど考慮されていない．

連絡先: 島田諭，筑波大学大学院図書館情報メディア研究科，〒305-8550 茨城県つくば市春日1-2，Tel: 029-859-1391，Fax: 029-859-1391，sat@slis.tsukuba.ac.jp

2.2 ナビゲーションの必要性

この問題に対応するには，文書集合に対してランダムな順序で閲覧できる従来型の検索だけでなく，文書集合が有している文脈を再現し，書き手が意図した文脈に沿って部分的にシーケンシャルな閲覧を可能とするようなナビゲーションの機能を備えておくことが必要である．このようなナビゲーションがあれば，ユーザは新聞や書籍を手に取るような感覚で，明確なクエリを持たずに文書集合を探索できる．

一般に，書かれた時期や内容，著者の関心や属性が近い文書は，似通った用字用語で記述されることが多い．このことから，シソーラスなどの外部知識を導入せず，語の局所的な共起関係のみを用いても，文脈をたどることができると考えられる．特にコミュニティにおいては，円滑な意思疎通を図るために，参加者間で一定の用字用語に収束させていくことがよく見られる．この場合，細かい用字用語の違いは，同一の内容であっても著者の性質が異なることを示していると考えられ，コミュニティにおける局所的な文脈やサブコミュニティの把握に有用と考えられる．

2.3 包括的Webナビゲーションとは

我々が提案している包括的Webナビゲーションは，ユーザによる包括的な内容把握を支援するための手法である．包括的な内容把握とは，文書集合の概略が把握でき，ユーザが関心を持つ主なトピックを列挙できる状態に至ることを指す．文書集合にどのような文書が含まれているのか不明な状態から探索を開始して包括的な内容把握に至るためには，ユーザに明示的な検索語の入力を要求せず，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ハイパーリンクを付与しておくことが必要である一方，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，文書あたりの出リンク数を少数に留めながら，文書集合全体を

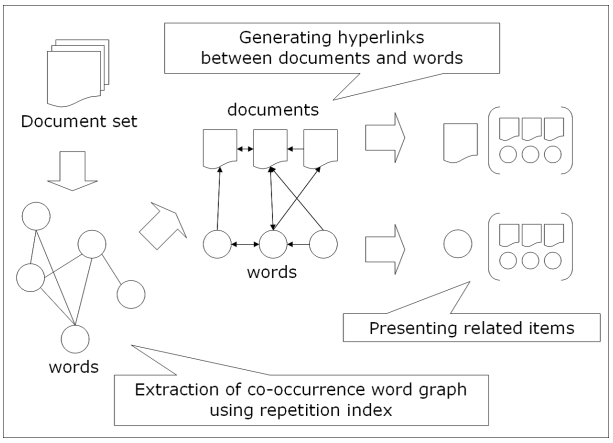

図 1: 提案システムにおける処理の流れ

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	(その他)	(その他)	0.01

網羅できることが求められる．

提案手法では，文書集合から抽出された語の共起と反復度に基づき，文書およびキーワードをノードとする Small-World ネットワークを生成してナビゲーションを行う [島田 08a]．Small-World ネットワークとは，多様なノードへ短い距離で到達可能な構造を持つネットワークをいう [Watts 98]．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，平均パス長（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 [松尾 02]．新聞記事，同一著者によるブログ記事，不特定多数の著者による CGM テキストを用いた実験で，いずれのデータに対しても Small-World ネットワークが生成できることを確認している．

3. 提案システムの概要

3.1 キーワードの抽出

提案手法では，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）[武田 01]，および df，df_2 の閾値を用いて,「特定性を示すキーワード」（区分 I）および「網羅性を示すキーワード」（区分 II）を抽出する．この閾値は，約 300 件のブログ記事を用いた予備調査により決定した [島田 08b]．

区分 I，II に含まれない語については,「周辺的なキーワード」（区分 III）および「その他のキーワード」（区分 IV）に区分する．なお，$df < 3$ となる低頻度語は除外する．

抽出されたキーワードに対し，当該の文書集合における性質を表すスコアとして，表 1 に示すスコアを付与する．

3.2 ハイパーリンクの生成

提案手法では，抽出されたキーワードに対し付与されたスコア，およびキーワードの共起関係を用いて，文書間，キーワード間，文書－キーワード間にハイパーリンクを生成する．

文書間のハイパーリンクは，以下に示す方法で生成する．基点となる文書 d_i と文書 d_j との間で共起するキーワード集合を $T_{ij} = \{t_1, ..., t_n\}$ とし，文書 d_i-d_j 間の関連度 $r(d_i, d_j)$ 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．

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

基点となる文書 d_i と 1 個以上の語が共起する全文書について，関連度 r を算出する．基点となる文書から，関連度 r の上位 8 件もしくは上位 20%のいずれか少ない方の数の文書に対し，ハイパーリンクを付与する．

キーワード間のハイパーリンクは，以下のように生成する．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 (2) により算出する．ここで，w_k は語 t_k のスコアである．

$$r(t_i, t_k) = mw_k \tag{2}$$

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．基点となるキーワードから，関連度 r の上位 8 件もしくは上位 20%のいずれか少ない方の数のキーワードに対し，ハイパーリンクを付与する．

文書からキーワードへのハイパーリンクは以下のように生成する．基点となる文書 d_i において出現するキーワード群 $T_i = \{t_1, ..., t_n\}$ を，基点となる文書 d_i から別の文書へ遷移するために有用な語の候補とする．個々のキーワード t_k のスコア w_k と，文書集合全体における文書頻度 $df(t_k)$ を用いて，基点となる文書における重要度を決定する．キーワード群 T_i から，区分 I，II，III，IV の順に，キーワードを取得する．区分 I のキーワードについては，df の降順でソートし，上位から順に取得する．その他の区分のキーワードについては，df の昇順でソートし，上位から順に取得する．取得した順に 8 個目もしくは取得したキーワード集合 T_i の 20%の数までのキーワードを，基点となる文書における重要語とする．基点となるキーワードから，これらの重要語に対し，ハイパーリンクを付与する．

キーワードから文書へのハイパーリンクは以下のように生成する．基点となるキーワード t_i に対し，t_i が出現する文書群 $D_j = \{d_1, ..., d_m\}$ を取得する．文書群 D_j に含まれる文書 d_j において出現するキーワード群を $T_j = \{t_1, ..., t_n\}$ とし，文書 d_j のスコア S_j を式 (3) により算出する．ここで，w_k は語 t_k のスコアである．

$$S_j = \sum_{k=1}^{n} w_k \tag{3}$$

文書群 D_j に含まれる全文書について，スコア S を算出する．基点となるキーワード t_i から，スコア S の上位 8 件もしくは上位 20%のいずれか少ない方の数の文書に対し，ハイパーリンクを付与する．

表 2: 実験に用いるデータの概要

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

表 3: タスク 1 におけるユーザ入力語とシステム提示語

ユーザ入力語		システム提示語
第一	第二以降	
O157	ユニーク　結果的	O157
ウイルス	結果　逆に	感染
病気	増　対照的	細菌
風邪	売れ　不安	雑菌
流行	不思議　面白い	大腸菌
大腸菌	落ち込み　売り上げ	豚肉
繁殖	好調　対策	予防
	商品　新商品	
	関連商品	

表 4: タスク 2 におけるユーザ入力語とシステム提示語

ユーザ入力語		システム提示語
第一	第二以降	
ビッグバン	今年	ODA
円安	株価	格付
外国人投資家		株安
景気		金制調
景気回復		金融持
景気回復宣言		金融政策
経済		個人消費
重大		財政構造改革
消費		三洋電機
消費税		資金量
消費税率		住友銀行
世界経済		中央銀行研究会
		東京三菱銀行
		東京市場
		抜本改革
		不良債権額

4. 語の中心性の分析

本研究では，ユーザがクエリを想起しにくい検索課題に対して，ユーザにクエリの入力を要求する従来の情報検索システムに比べ，幅広い探索を容易することを目指している．このことについて，曖昧な検索課題を用いた被験者実験により得られたクエリログを用いて評価を行った．

4.1 実験に用いるデータ

本論文では,「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．

「朝日新聞」では，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．

4.2 被験者実験の概要

協力を得た被験者 4 名に対し，以下のタスクを示して，提案システムを用いた情報探索を行ってもらった．

- タスク 1「この年に猛威を振るった「O157」に関連して，急によく売れるようになった商品や，対策のために売り出された新商品などを網羅的に列挙してください．また，その商品について書かれた記事も挙げてください.」
- タスク 2 「この年の経済分野での三大ニュースと思われるトピックを 3 つ見つけて，漢字 4〜6 字程度で表現してください．また，それぞれのトピックを象徴すると思われる記事を 3 件程度ずつ挙げてください.」

タスク 1 は，特定のテーマに関して網羅的に情報を探索する例である．タスク 2 は，漠然としたテーマに関して，探索を進めながら情報を分類していく例である．いずれも，従来の情報検索でも一定の探索が可能ではあるが，提案手法を用いることによって探索が容易になったり，ユーザにとって見つけにくい情報を提示できることが期待される．

それぞれのタスクにおいて，ユーザにより入力された検索語（ユーザ入力語），および提案システムにより提示された語のうち，ユーザにより選択された語（システム提示語）を表 3,4 に示す．

4.3 次数中心性および媒介中心性に基づく評価

ネットワークが Small-World 性を示すためには，次数が少ないながら，ネットワーク上の平均パス長を大幅に短縮するノードの存在が不可欠である．つまり，共起語グラフにおける「次数中心性が低い割に媒介中心性が高い語」が，そのようなノードであると考えられる．

本論文では，式 (4) により得られる値を次数中心性（degree centrality），式 (5) により得られる値を媒介中心性（betweenness centrality）として用いる[*1]．

$$C_d(v_i) = \frac{deg(v_i)}{N-1} \tag{4}$$

$$C_b(v_k) = \sum_{i,j(i\neq j)} \frac{\sigma_{ij}(v_k)}{\sigma_{ij}} \tag{5}$$

ここで，$deg(v_i)$ は，ノード v_i と隣接するノードが持つリンク数，N はノードの総数である．また，σ_{ij} は，ノード v_i と v_j 間の最短経路の総数，$\sigma_{ij}(v_k)$ は，ノード v_i と v_j 間の最短経路のうち，ノード v_k を経由する経路の数である．

予備調査により，全体では次数中心性が高いほど媒介中心性も高くなる傾向が見られた．しかし，単に次数中心性が高いことによって媒介中心性が引き上げられているだけの語は，ナビゲーションにおいて重視しない．このため，次数中心性および媒介中心性の値を直接用いるのではなく，式 6 に示す指標を用いて，各語のナビゲーションにおける有用性を評価した．

*1 Pajek（http://vlado.fmf.uni-lj.si/pub/networks/pajek/）を用いて算出した．

表 5: 実験結果

		反復度	df	df_2	共起語数	C_d	C_b	$C_{b/d}$
ユーザ入力語	平均	0.209	111.31	23.66	2925.3	0.03296	**0.00076**	-1.317
	分散	0.023	25039.79	1076.59	6964868.0	0.00165	0.00000	0.035
システム提示語	平均	**0.323**	**38.34**	**12.85**	**985.4**	**0.00895**	0.00059	**-1.143**
	分散	0.048	21597.67	3194.46	3052906.6	0.00120	0.00003	0.035
網羅性を示す語	平均	0.425	**18.65**	**7.98**	878.2	**0.00379**	0.00002	**-1.054**
	分散	0.003	80.64	16.85	129698.7	0.00001	0.00000	0.020
特定性を示す語	平均	**0.685**	22.27	14.50	**832.4**	0.00481	0.00008	-1.075
	分散	0.010	931.16	376.83	825693.6	0.00007	0.00000	0.016
その他の語	平均	0.241	46.11	13.93	1037.5	0.01098	0.00082	-1.209
	分散	0.036	30407.92	4484.21	4186528.7	0.00168	0.00004	0.039

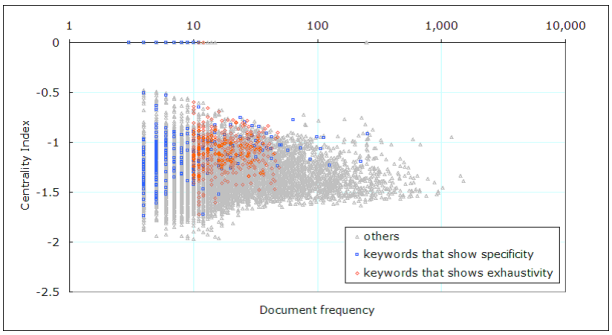

図 2: すべてのキーワードにおける中心性と文書頻度の分布

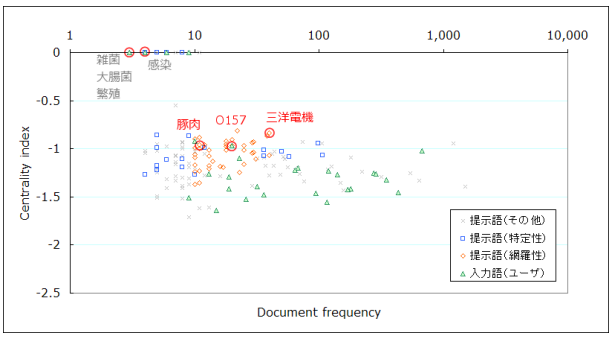

図 3: ユーザ入力語とシステム提示語における中心性の分布

$$C_{b/d}(v_m) = -\frac{\log_2(C_b(v_m))}{\log_2(C_d(v_m))^2} \tag{6}$$

この値について，対象データ全体における分布を図 2 に，ユーザ入力語およびシステム提示語における分布を 3 に示す．全体では，df が低いほど $C_{b/d}$ の分散は大きい．提案システムでは，概ね $C_{b/d}$ が-1.5 より高くなる語を「網羅性を示す語」として提示できている．「特定性を示す語」においては，df が高くても $C_{b/d}$ が下がらない語を提示できている．

ユーザ入力語とシステム提示語について，反復度，df，df_2，共起語数，C_d，C_b，$C_{b/d}$ の平均および分散を表 5 に示す．提案システムでは，ユーザ入力語と比べ反復度が高く共起語数が少ない，より局所的な関連性を示すと考えられる語を提示している．単純に媒介中心性 C_b を見るとユーザ入力語のほうが高いが，次数中心性を考慮した $C_{b/d}$ で見るとシステム提示語のほうが高い．

システム提示語においては，$C_{b/d}$ は「網羅性を示す語」が最も高く，「特定性を示す語」「その他の語」の順に下がっている．

5. おわりに

本論文では，文書集合から生成した共起語グラフにおける，次数中心性を考慮した媒介中心性の指標である $C_{b/d}$ を用いて，提案手法により抽出された特徴語と，ユーザが入力した検索語を比較した．その結果，システムによる提示語はユーザによる入力語と比べ，$C_{b/d}$ の平均が高く，ナビゲーションにおいて特に重要である網羅性（トピック間を橋渡しする性質）を強く示す語を提示できていることが確認できた．

システムにより抽出されたキーワードは，ユーザに対して遷移経路の候補を提示するためのハイパーリンクのアンカーテキストとして提示されるが，同時に内容の要約（スニペット）や，現在位置（トピックの粒度や，文書集合における希少性）も示していると考えられる．今後，これらそれぞれの効果を分離して調べられる，より精密な実験が必要である．

参考文献

[Watts 98] Watts, D. and Strogatz, S.: Collective dynamics of 'small-world' networks, *Nature*, Vol. 393, No. 6684, pp. 440–442 (1998)

[松尾 02] 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)

[島田 08a] 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)

[島田 08b] 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)

[武田 01] 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)