Small World 構造に基づく文書からのキーワード抽出

松尾 豊[†,††] 大澤幸生[†††,††] 石塚満[†]

本論文では，Small World 構造を利用した文書からのキーワード抽出法を提案する．Small World とは，ノードがクラスタ化されているにも関わらず，ノード間の平均パス長が短いグラフ構造である．文書中の単語の共起関係により構成したグラフが，Small World の特徴を備えていることを示す．さらに，ある語を取り除くことによって平均パス長が大きく増加するような語をキーワードとして取出す．このような語は，離れたクラスタ，すなわち概念を橋渡しする語であり，文書の主旨において重要な語である可能性が高い．

Keyword Extraction using Small World Structure in a Document

Yutaka Matsuo,[†,††] Yukio Ohsawa[†††,††] and Mitsuru Ishizuka[†]

We develop a new keyword extraction algorithm which utilizes small world structure of a document. In a graph with small world structure, nodes are highly clustered yet the path length between them is small. A term co-occurrence graph, where nodes represent terms in a document and edges represent the co-occurrence of terms, is shown to have small world characteristics. Furthermore, terms are extracted as keywords that have high contribution to the graph being small world. Such words connect multiple clusters i.e., concepts, thus they are important for the point of a document.

1. はじめに

キーワード抽出は，文書検索，Web ページ検索，文書クラスタリング，要約文抽出など，情報検索において重要な技術である．適切なキーワードを自動的に抽出することができれば，読むべき文書を選択しやすくなったり，文書間の関係を把握することが容易になるなどのメリットがある．また，近年着目を集めているテキストマイニングの視点からも，文書の意図や特徴を端的に表すキーワードを的確に抽出する技術は，文書の傾向をつかむ，特徴的な意見を見つける，新しい知見を得るといった用途に必要不可欠である．実際に，多くの情報検索／テキストマイニング手法が何らかの形でキーワード抽出技術を用いている．

† 東京大学
The University of Tokyo
†† 科学技術振興事業団
Japan Science and Technology Corporation
☆ 現在，産業技術総合研究所サイバーアシスト研究センター
Presently with Cyber Assist Research Center, National Institute of Advanced Industrial Science and Technology
††† 筑波大学
The University of Tsukuba

キーワード抽出法としてよく用いられている tfidf[24] は，当該文書中の語の出現頻度 (tf) と，その語がコーパス中でどのくらいの文書に出現するかという文書頻度 (df) を用いた尺度である．tfidf は，他の文書と比べて多く出現する語を重要と考えようというもので，その根底には「何度も繰り返し言及される概念は重要な概念である」という仮定[14] がある．idf はある情報量の単純で頑健な推定値となっており，tfidf は出現確率と情報量をかけあわせた特徴量であるという指摘もされている[2]．tfidf の他にも，ある文書集合にだけ偏って出現する語は特徴的である[21)5)]，文書集合中で共起する語が少ないほど特徴的である[25]，共起する単語分布の偏りが大きい語ほど特徴的である[9] などの手法も提案されている．これらは，文書集合中の語の分布をもとに，統計的／経験的な尺度を用いてある文書（集合）を代表する語彙を自動識別する方法であり[11]，文書中に各語が出現する事象を単独の事象として処理している．

一方，23) では，文書から構成した語の共起グラフを用いてキーワードを取り出す手法を提案している．これは文書中の語の構造に着目していたアプローチであり，複数の概念クラスタと共起する語を重要と考え，語の構造的な重要性を計っている．テキストから得た

語のネットワークを用いて情報検索の支援をするシステムとして，3)22) などがある．また，29) には，人間関係などの社会的なネットワークを対象として，ネットワークの力と影響についてさまざまな側面から述べられている．社会的なネットワークでは，重要な位置を占めると大きな影響力を得られるが，この視点からは，文書から得た語のネットワークにおいて重要な位置を占める語は，もとの文書においても重要性が高いと考えるのが自然であろう．

本論文では，文書から得た語の共起グラフを利用し，最近注目を集めている Small World 構造に基づいたキーワード抽出アルゴリズムを提案する．Small World とは，ノードがクラスタ化されているにも関わらず，任意の 2 点間のパス長が短いグラフである．文書から得られた語の共起グラフもこのような構造をしており，Small World 構造に対する貢献の高い語をキーワードとして抽出する．

なお，本論文におけるキーワードとは,「文書中に出現し，著者が自分の主張を伝える上で重要であると考える語」を指す．一般的な検索における網羅性，特定性という視点とは異なる部分もあるが，文書は著者が何かを伝えるために書いている以上，キーワードとしては著者の主張の上で重要な語を取り出すべきであると考える．テキストマイニングの視点からも，テキストに書かれた意図を取り出すために，著者の主張を表す語を取り出すことが重要である．

次章では，本手法の基礎となる Small World について紹介し，3 章では論文から得た語の共起グラフが Small World 構造を備えていることを示す．Small World 構造に基づいたキーワード抽出法とその評価を 4 章に述べ，5 章で議論を行う．なお，本論文で扱うグラフは，重みなし無方向グラフであり，社会学的な話題との整合性から，節をノード，弧をリンクという．

2. Small World とは

初めて会った人と共通の知人を発見して,「せまい世界ですね．(It's a small world.)」と言った経験は誰もがあるだろう．Small World のひとつの定式化は，任意の 2 人がどのくらいの確率で知りあいであるかというものであり，もうひとつの定式化は，任意の 2 人が平均何人の知り合いの「鎖」を通じてつながっているか，というものである．1960 年代，著名な社会心理学者である Stanley Milgram は，米国のネブラスカ州オマハに住む住人 160 人をランダムに選び，1300 マイル以上離れたボストンの株主仲買人まで手紙を転送してもらうという実験を行った[18]．手紙は，first-name を

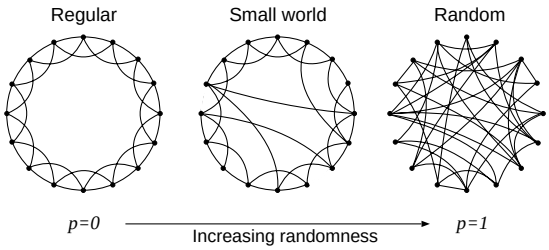

図 1 regular ring lattice のランダムなつなぎかえ．

知っている知り合い，友人だけに転送され，1 通の手紙が届くまで平均 5 人が転送を仲介したことが明らかになった．つまり，米国に住む互いに全く面識のない 2 人が，平均 5 人を間に介して結ばれているのである．この結果は大きな驚きをもって迎えられ，six-degrees of separation として米国では広く知られるところとなった☆．その後，Mark S. Granovetter[7]，Manfred Kochen[13] ら社会学者によって，この現象にさまざまな考察が加えられた．

長らく社会心理学的な研究対象であった Small World は，1998 年に Duncan Watts らがグラフにおける 2 つの特徴量として定式化を行って以来，コンピュータサイエンスの分野でも注目を集めるようになった．2 つの特徴量とは，以下である[28)4)]．

- L (characteristic path length)：すべてのノードの組についてのパス長の平均．パス長とは，最短パスの長さである．
- C (clustering coefficient)：ひとつのノードが k 個のノードと隣接しているとき，この k 個のノード間に存在するリンク数を ${}_kC_2$ で割ったものを，すべてのノードについて平均をとったものである．つまり，リンクを知り合い関係に例えると，C は自分の知り合い同士が知り合いである確率を表す．

ノード数，リンク数が一定であるとすると，C が大きいグラフは，近傍同士でのリンクが多いので L は大きくなる傾向がある．また，C が小さいランダムなグラフでは L は小さい．したがって，L と C は対応していると考えがちであるが，Watts によると，Small World は $L \geq L_{rand}$ (または $L \approx L_{rand}$) かつ $C \gg C_{rand}$ であるようなグラフとして定義される．L_{rand}，C_{rand} は，同じノード数，リンク数のランダムグラフにおける L と C である．つまり，ランダムグラフと同程度に L が小さいにも関わらず，近傍同士のリンクが非常に多いのである．

☆ "Six-degrees of separation"（邦題：私に近い 6 人の他人）という映画も 1993 年に公開されている[8]．

図 1 は，β-Graph と呼ばれる Small World のモデル[27] である．まず，n 個（ここでは 16 個）のノードがそれぞれ近傍の k ノード（ここでは 4 ノード）にリンクが張られている規則的なグラフ（regular ring lattice）を考える．次に，各リンクに対して，確率 p でリンクのつなぎかえを行う．リンクのつなぎかえとは，ランダムに選んだノードへとリンクを張りかえる操作である．これを全てのリンクに対して 1 回ずつ適用する．

$p = 0$ のときは，全くつなぎかえを行わないことに相当し，$p = 1$ のときは，すべてのリンクをつなぎかえたランダムグラフとなる．この中間の値では，近傍を結ぶ規則的なリンクとランダムな長いリンクが混在するグラフが得られ，C が大きく L が小さい Small World となる．ランダムなリンクがショートカットの役割を果たし，急激に L を減少させるのである．このモデルは，Small World を直観的に把握しやすく，さまざまな文献で紹介されている．

表 1 は，社会ネットワーク，送電網，神経回路網それぞれのグラフに対しての Watts らの解析結果である．いずれのグラフにおいても，L は L_{rand} と同程度（もしくは少し大きい）であるが，C は C_{rand} よりかなり大きく，Small World の特徴を備えていることが分かる．これらのグラフの他にも，WWW[1] や食物連鎖のグラフ[20] などが Small World であることが次々と報告されており，自然現象や人工物において Small World が遍在することが明らかになりつつある．

では，なぜ Small World が自然界や人工物に遍在するだろうか．これについては，局所的かつ大域的な情報伝達効率が高い[15]，グラフの連結性の最大化 (maximal connectivity) とコストの最小化 (minimal cost) のトレードオフである[16] などの指摘がされている．ノード間の距離には，物理的な距離（例えば航空路でいえば，New York と L.A. の物理的な距離）とグラフ上の距離（飛行機を何回乗り継がなければいけないか）の 2 つがある．信号や情報，物質などの伝達効率からは，グラフ上の距離は短い方が望ましい．すなわち，L は短い方が望ましい．一方，リンクを張るコストはノード間の物理的な距離に比例すると考えると，リンクの物理的な長さの平均 W は短い方が望ましい．L と W はトレードオフの関係にあるが，16) では関数

$$E = \lambda L + (1 - \lambda)W$$

(ただし，λ は 0 以上 1 以下の重みを表すパラメータ) を想定し，これを最小化するようなリンクの張り方を求めると，λ が 0 のときには regular lattice が，λ が 1 のときはランダムグラフが得られ，その中間では Small World が出現することを明らかにした．多くのグラフでは，連結性の最大化（L が小さいこと）とコストの最小化（W が小さいこと）という相反する両方の要求があるため，Small World が遍在すると考えられている．

表 1 Small World 構造を持つ様々なグラフの L,C[28]
Table 1 L and C for graphs with a small world topology[27]

	L	L_{rand}	C	C_{rand}
Film actor	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
C. elegans	2.65	2.55	0.28	0.05

Film actor は，ハリウッドの俳優についてのグラフ（$n = 224,225$, $k = 61$）で，2 人の俳優が同じ映画で共演していればリンクが張られる．Power grid はアメリカ西部の送電網についてのグラフ（$n = 4941$, $k = 2.67$）で，ノードは発電機，変電所等を表し，リンクは高圧送電線を表す．*C. elegans* は，*Caenorhabditis elegans* という線虫のニューロンのネットワーク（$n = 282$, $k = 14$）で，シナプスもしくはギャップ結合でつながれたニューロン間にリンクが張られる．

以上，Small World に関連して，本論文と関係する重要な部分だけを紹介するに留めたが，詳細については各文献を参照されたい．

3. 文書中の語の共起グラフ

本論文では，ひとつの文書から生成する語の共起グラフについて考察する．英語の文書を対象とした場合，語の共起グラフは以下のような手順で構成することができる．なお以下では，グラフのノード数を n，1 ノードあたりのリンク数の平均を k とする．

(1) 前処理：ステミング☆1を行う．ストップワード☆2を取り除く．N-gram によりフレーズを抽出する☆3．

(2) ノードの生成：規定回数 (f_0 回) 以上出現する語（フレーズも含む）をノードとして取り出す．

(3) リンクの生成：2 つのノードに対応する語の，同一文中での共起が多ければリンクを張る．共起は Jaccard 係数☆4を用いて計り，この上位か

☆1 語幹の形を得る処理．” ···ing”,”···ed”, 三単元の s などを取り除く．ここでは Porter の方法[26] を用いた．

☆2 “a”,”the”,”that” などのあらかじめ決められた不要語．ここでは，Salton の SMART System のリストを用いた．

☆3 $N = 4$ を用いた．5 単語以上から成るフレーズが f_0 回以上出現することは実験ではほとんどなかった．

☆4 語 a と b に対する Jaccard 係数は，$Jaccard(a,b) = \#sentence(a \cap b)/\#sentence(a \cup b)$ である．ここで，$\#sentence(a \cap b)$ は，a, b 両方の語を含む文の数，$\#sentence(a \cup b)$ は少なくとも一方の語を含む文の数である．Jaccard 係数は，例えば 12) で Web 上の文書から人名の共起グラフを構成する際に用いられている．

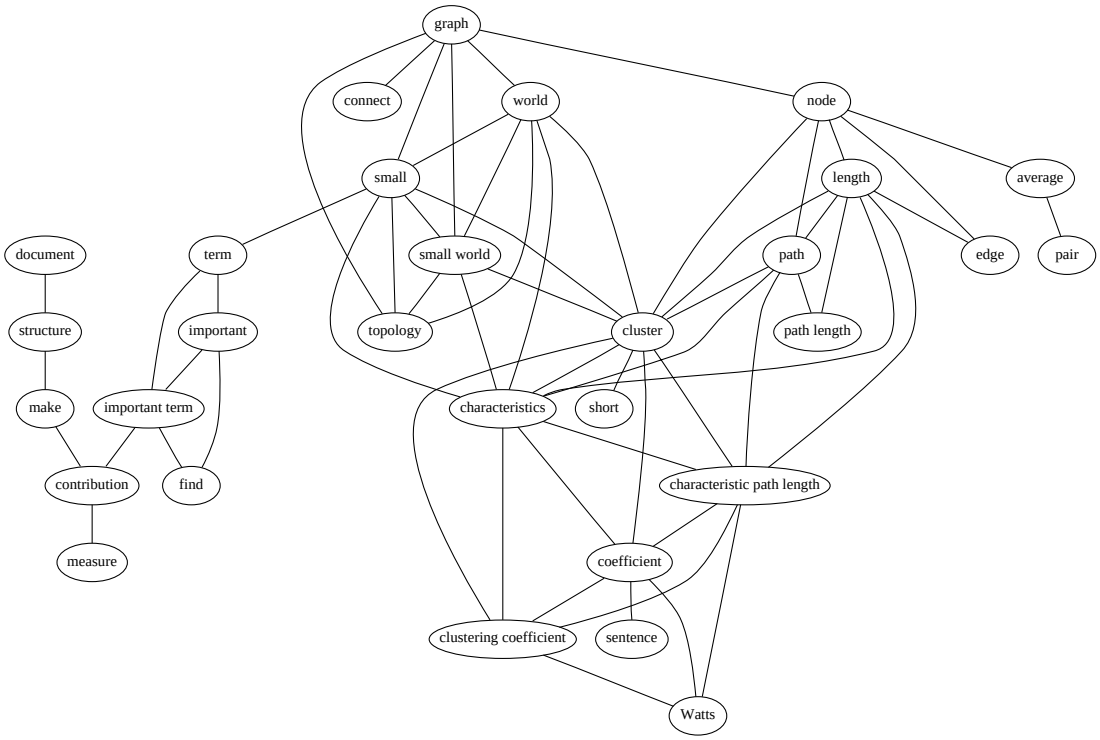

図 2 論文 17) から得た語の共起グラフ．$f_0 = 5$, $k_0 = 4.0$ とし，最大連結サブグラフだけを取り出している．

ら順に k が既定値 (k_0) に達するまでリンクを張る．

図 2 に，本論文とほぼ同じ内容の論文 17) に対して得た語の共起グラフを示す．いくつかの語の集まりがクラスタを作っている様子が分かる．右側の大きなクラスタには Small World に関連する語が集まっており，左側には “important term” “document” などキーワード抽出に関する語が集まっている．そして 2 つのクラスタを “small” と “term” のリンクがつないでいる．また右側のクラスタの中にも，“small world” “topology” などの語の集まり，“node” “path” “length” などの語の集まり，“characteristic path length” “clustering coefficient” “Watts” などの語の集まりが見てとれる．関連する概念の語がリンクで結ばれながら，全体としてもまとまりのある構造になっている．実際，このグラフに対して $L = 3.63$，$C = 0.524$（$L_{rand} = 2.40$, $C_{rand} = 0.0856$）であり，L が L_{rand} と同程度でありながら C が C_{rand} より非常に大きいという Small World の特徴を備えている．

一般的に，論文が Small World であるかどうかを，WWW9[☆] の論文 57 篇および JAIR[☆☆] の論文 166 篇を用いて検証した．結果を**表 2**，**表 3** に示す．WWW9 の論文は，平均で L は 5.60 で L_{rand} の 1.5 倍程度だが，C は C_{rand} と比べて 15 倍以上大きい．JAIR の論文は WWW9 のものと比べて，論文の長さのばらつきが大きいが，やはり L は L_{rand} の 1.4 倍程度であるのに対して C は C_{rand} の 14 倍である．したがって，L が L_{rand} と同程度で，C が C_{rand} よりも非常に大きい[☆☆☆]という Small World の性質を備えていることがわかる．なお，同じ規模の regular lattice に対して，L は 20 以上，C は 0.6 程度である．

論文から得た語の共起グラフが Small World の性質を持っている理由は，次のように説明することがで

☆ 9th International World Wide Web Conference. http://www9.org/.

☆☆ Journal of Artificial Intelligence Research の 93 年 (Vol.1) から 2001 年 (Vol.14) までの論文.

☆☆☆ 著者らの知る限り，C_{rand} と比べ C がどの程度大きければ Small World といえるのかという定量的な報告はされていない．表 1 では，非常に n の大きい Film actors で C は C_{rand} の約 3000 倍，Power grid では 16 倍，*C. elegans* では 5.6 倍である．WWW9 や JAIR の論文では，C は C_{rand} の十数倍であり，グラフの規模が同程度である *C. elegans* が Small World であると認められていることから考えて，十分 Small World といえると判断した．

表 2 WWW9 の論文 57 篇についての L と C
Table 2 L and C for 57 graphs of papers in WWW9

	L	L_{rand}	C	C_{rand}
Max.	7.67	4.21	0.509	0.0432
Ave.	5.60	3.71	0.355	0.0211
Min.	4.17	3.03	0.196	0.0113

全ての論文に対して，各項目それぞれの最大値，平均，最小値を示している．共起グラフは，$f_0 = 3$, $k_0 = 4.0$ として生成し，最大連結サブグラフ（平均 79%のノードをカバーする）だけに着目した．得られたグラフは，平均で $n = 275$, $k = 5.04$ であった．(k が k_0 と異なるのは，最大連結サブグラフにだけ着目しているためである．)

表 3 JAIR の論文 166 篇についての L と C
Table 3 L and C for 166 graphs of papers in JAIR

	L	L_{rand}	C	C_{rand}
Max.	9.22	3.97	0.588	0.0931
Ave.	4.73	3.38	0.326	0.0230
Min.	3.07	2.41	0.149	0.0129

$f_0 = 3$, $k_0 = 4.0$ とした．最大連結サブグラフ（平均 88%のノードをカバーする）だけに着目し，得られたグラフは平均で $n = 196$, $k = 4.81$ であった．

きる．論文の一文中で同時に用いる語は関連が強い方が分かりやすい．例えば,「ノード」や「パス」という語は，グラフ構造に関する語であるので同時に用いても分かりやすいが,「パス」と「文書」は関連が弱く，同時に用いると分かりにくい．一方で，なぜ「パス」と「文書」がひとつの論文に出現するのか明らかであることも重要である.「パス」「ノード」などのグラフに関する語と「文書」「キーワード」などの文書に関する語が,「共起グラフ」という語によって結びついていることが明らかであれば，筆者の主張が伝わりやすいだろう．したがって，読者に分かりやすく，しかもまとまりのある論文を推敲しながら書き上げるという作業は，コストの最小化と連結性の最大化の両方を考慮していると考えられる．

次章では，複数の語のクラスタを結び付けている語をキーワードとして取り出す手法について述べる．

4. Small World 構造を用いたキーワードの抽出

4.1 アルゴリズム

文書から得られた語の共起グラフが Small World であるとすると，いくつかのノードは L を減少させるのに大きく貢献しているはずである．このような語は，互いに関連のうすい語のクラスタ同士をつないでいるのであるから，文書の論旨において重要な意味を担ったキーワードであると考えることができる．本節では，まず L の定義を非連結グラフに拡張した後，ひとつのノードの Small World 構造に対する貢献を示す contribution という指標について述べる．

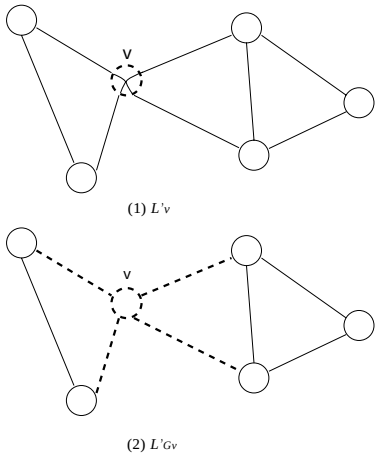

図 3 L'_v と L'_{G_v} の例

定義 4.1
ノード i，ノード j に対する *extended* path length $d'(i,j)$ を次のように定義する．

$$d'(i,j) = \begin{cases} d(i,j), & \text{if } (i,j) \text{ are connected,} \\ w_{sum}, & \text{otherwise.} \end{cases} \quad (1)$$

ただし，$d(i,j)$ は，連結したグラフにおけるノード i とノード j のパス長である．w_{sum} は定数で，すべての連結していないサブグラフの幅の和である．グラフの幅とはグラフ中の 2 ノード間のパス長の最大値であり，w_{sum} は，サブグラフが新たなリンクにより連結されたときの 2 ノード間のパス長の上限を与えている．

この定義を用いて，L を自然に拡張することができる．

定義 4.2
extended characteristic path length L' は，すべてのノードの組についての extended path length の平均である．

さらに，ひとつのノードの L に対する寄与を計るために，次の定義を行う．

定義 4.3
L'_v は，ノード v 以外のすべてのノードの組についての extended path length の平均である．L'_{G_v} は，ノード v を取り除いたグラフにおける extended characteristic path length である．

これを**図 3** に簡単に図示する．L'_v の計算ではノード v はグラフに接続されているが平均には含めない．L'_{G_v}

表 4 論文 17) の頻出語
Table 4 Frequent terms in 17)

Term	Frequency
graph	39
small	37
world	37
term	34
small world	30
node	29
paper	21
length	21
document	19
edge	19

表 5 論文 17) における contribution が上位の語
Table 5 Terms with 10 largest CB_v in 17)

Term	CB_v	Frequency
small	3.05	37
term	2.80	34
important term	1.93	7
contribution	1.64	6
node	1.00	29
make	0.82	6
cluster	0.57	15
graph	0.54	39
coefficient	0.52	8
average	0.50	8

の計算ではノード v と v を含むリンクはグラフから除外される．この差をとることで，ノード v が L の減少にどれくらい影響を与えているかを求めることができる．

定義 4.4
ノード v の *contribution* CB_v は，次のように定義される．

$$CB_v = L'_{G_v} - L'_v \tag{2}$$

すなわち，contribution はノード v の Small World 構造への貢献を計る指標である．この値が大きいノードは，離れたクラスタをつなぐ重要なノードであると考えられる．

4.2 キーワード抽出アルゴリズムとしての評価

前述の論文 17) に対して，頻出語の上位と contribution の高い語の上位を**表 4**, **表 5** に示す．頻出語は，“graph”, “small world” など，論文中に多く出現する一般的な語が得られているのに対し，contribution の高い語は “important word” や “contribution”, “cluster” など，出現回数は多くないものの論文中で重要な役割を果たす語が得られていることが分かる．

contribution の上位語が，論文の主旨を表すキーワードとしてどの程度優れているか評価実験を行った．ここでのキーワードとは，文書検索において他の文書と区別するための語ではなく，文書の著者の主張を表す語という意味で用いている．したがって評価実験は，実際に論文の著者にキーワードかどうか判定してもらうという方法をとった．

表 6 論文 17) の $CB_v \times idf$ 上位語
Table 6 Terms with 10 largest $CB_v \times idf$ in 17)

Term	CB_v	Frequency
small world	2.58	37
shortest path	1.76	34
short cuts	0.94	7
contractor	0.90	6
rare	0.80	29
co-occurrence	0.73	6
sentence	0.63	15
path length	0.50	39
important term	0.48	8
document	0.15	8

実験は，人工知能の分野の 7 著者 20 論文に対して行い，tf，tfidf☆，idf☆☆，KeyGraph☆☆☆と比較した．各手法でキーワード 15 個を出力し，各手法から得られたキーワードの上位語を混ぜてシャッフルし，著者に「論文を構成する重要な概念を表すと思う語にチェックをして下さい」という質問を行った．各手法による出力語中でキーワードであると判定された割合が precision である．さらに，「提示した全ての語（提示した以外の語でも覚えているものがあれば含めてよい）のうち，論文中で不可欠な概念を表す語 5 つ以上を選び A，B，C，D，E と印をつけ，それと同義の語にも同じ印をつけてください」という指示を行った．5 つ（以上）の概念のうち各手法で提示した語にいくつ含まれているかで coverage を測定した．なお，$f_0 = 3$，$k_0 = 4.0$ とした．

結果を**表 7** に示す．本手法は，precision，coverage ともに 50%程度と，tf と同程度の性能しか得られておらず，tfidf より悪い結果となっている．precision, coverage の他に frequency index という指標を表示しているが，これは各手法が提示した語の出現頻度の平均を示しており，これが高いほど「当たり前」の語を出力していることになる．本手法の frequency index は低いので，出現頻度が低いにも関わらず重要な語を，tf と同程度の割合で取り出しているという点で，一定の評価はできるだろう．

☆ コーパスは前述の JAIR の論文とした．また，語 v に対する idf の重みづけ $idf(v) = \log(N/df(v)) + 1$ とした．ただし N は全文書数，$df(v)$ は語 v が出現する文書数である．
☆☆ 後述のように本手法と idf を組み合わせて用いるため評価に加えた．なお，出現回数が 3 回以上の語を対象とする．
☆☆☆ 本手法と同様に構造的な特徴からキーワードを抽出するため，コーパスは不要である．

表 7 Precision と Coverage
Table 7 Precision and Coverage

	tf	idf	KeyGraph	**本手法**	tfidf	**本手法・idf**
precision	0.53	0.44	0.42	**0.47**	0.55	**0.73**
coverage	0.48	0.52	0.44	**0.52**	0.61	**0.68**
frequency index	28.6	6.9	17.3	**13.8**	18.1	**11.1**

本手法の性能が良くない理由として，出力の上位に，例えば表 5 の “make” のような非常に一般的な語が含まれることが挙げられる．確かに，“make” などの語は多くの語と共起し，共起グラフにおいて複数のクラスタをつなぐのは当然である．しかし，このような語は，ある論文中でだけ出現するわけではなく，他の論文でも同様に多数回出現する．そこで，一般的であるにも関わらずストップワードには含まれていない語を取り除く目的で，語 v の重みを

$$CB_v \times idf(v)$$

とする工夫を行った．語 v が多数の文書に出現する語であれば，$idf(v)$ の値は低くなる．

結果を本手法・idf として表 7 中に示している．precision，coverage とも大幅に良くなっており，tfidf と比べてもよい性能が得られている．また，frequency index も，本手法単独のときよりもさらに下がっており，一般語を取り除くという目的が果たせていることが分かる．実際に，論文 17) に対して本手法・idf で得たキーワードを**表 6** に示す．表中のほぼ全ての語が，論文のキーワードとして適切な語となっている．

結論として，本手法は頻度が低いにも関わらず重要な語を抽出することができる．しかしながら，キーワード抽出アルゴリズムとしての精度をあげるために idf と組み合わせることにより，よい性能が得られる．

4.3 本手法の計算量

本手法は，最短路の探索を繰り返し行うので計算コストは比較的高い．ある語 v について式 (2) に示される CB_v を求めるには，全てのノードペアの間の最短路の計算が必要である．最短路の計算にはダイクストラ法やワーシャル-フロイド法（例えば 10)）などのアルゴリズムを用いることができるが，全ノードペアの最短路の計算量はノード数 n に対して $O(n^3)$ の計算量となる．本手法ではこの部分の計算量が最もオーダが高く，これが本手法の計算量となる．

本論文の実験では n は 200〜300 程度であり，CPU Pentium II 333MHz の計算機に実装した Linux 上の C 言語のプログラムで 30 秒以内でキーワードが得られる．しかし，計算量をいかに減らしながら同様のキーワードを得るかも今後の課題のひとつであろう．

5. 議　　論

本手法では，文書から共起グラフを生成し，その構造的な重要性に着目してキーワードを抽出する．しかし，文書からどのように共起グラフを生成するか，また，どのような構造的な重要性に着目するかにはいくつかのバリエーションが考えられる．

まず，評価実験では $f_0 = 3$ という値を用いた．つまり，文書から共起グラフを生成する際，出現回数が 3 以上の単語をノードとしている．この値を大きくすれば，頻度による足切りを行うことになり，より頻度を重視した結果となる．逆に，この値を 3 より小さくすると，ノード数が極端に多くなり，その文書に偶然出現したような必然性のない語まで出てきてしまう．評価実験では，本手法の特徴を明らかにするために，なるべく小さな $f_0 = 3$ という値を用いた．

次に，ノード間のリンクを張る際，本論文では単語間の共起関係の強さを Jaccard 係数を用いて計っている．ここでの Jaccard 係数は，語が出現する文の集合がどのくらい類似しているかを計る指標であるが，他にも共起頻度を用いる方法や相互情報量を用いる方法などがある[26]．これらの指標についても比較を行ったが，共起頻度を用いた場合には出現頻度の高い語からリンクが多く張られ，結果的に出現頻度の高い語が抽出されやすくなる．一方，相互情報量は，2 つの語が独立に生起する場合の確率と共起する確率を比較するので，出現回数が小さく，しかも同時に出現する語のペアに対して大きな値が出やすくなる．その結果，出現頻度の小さい語に偏ってリンクが張られ，キーワードとして選ばれやすくなる．以上の知見を踏まえて，本論文では Jaccard 係数を用いた．

また，本論文ではグラフの構造における重要性を，Small World 構造に着目して contribution という指標で計っているが，グラフの中心性を計る指標としては Freeman の centrality の概念[6] が有名である．これは，人間関係などのネットワークにおける情報伝達やコントロールにおいて，あるノードがどのくらい中心的かを計る指標である．Freeman は以下の 3 つの指標を提示している．

- *degree*: ひとつのノードがいくつのノードとリン

クしているか，つまりノードの次数 (degree) によって中心性を計る．

- *betweenness*: あるノードが他のノードのペアの最短パスにどのくらいの割合で含まれているかを計る．つまり，情報を伝える際に，そのノードを通らなければいけない度合を表す．
- *closeness*: あるノードから他のノードへのパス長の合計である．グラフ中のどのノードにも近いノードが中心的であるとするものである．

これらの指標を用いた語の重み付けの検討も行ったが，得られる語はグラフの中心部に偏ってしまう．例えば，論文 17) に対して適用すると，“path length”, “characteristic path”, “characteristic path length”, “path length averaged” など，お互いに近い位置にあるフレーズが大量に出てきてしまう．文書中では，中心部の語だけではなくグラフの端の方の語も重要であるし，中心部と端の語の関係を表す語も重要であるが，これらの指標はもともと中心性の指標であるため，キーワード抽出のための特徴量としては適当でない．

本手法の欠点として，文書からグラフ構造を正確に取り出すために，文書の長さがある程度必要である点が挙げられる．論文の抄録程度の長さでは，適切なグラフを抽出することは難しいが，WordNet[19] などの語義のグラフ構造や，文書集合全体の語のグラフ構造を利用することにより解決できる可能性もあるだろう．

6. ま と め

本論文では，文書から抽出した語の共起グラフが Small World 構造であることを示し，構造的に重要な役割を担う語をキーワードとして提示する手法を提案した．頻度による語の重み付けの背景にある「何度も繰り返し言及される概念は重要な概念である」という仮定は，簡単でありながら非常に強力である．しかし，本論文では「概念のネットワーク上で重要な位置を占める概念は重要な概念である」というもう一つの仮定を提案している．今後はさらに共起グラフのリンクに長さを付与するなど，改良を行っていく予定である．

参 考 文 献

1) Adamic, L. A.: The Small World Web, *Proc. ECDL'99*, pp. 443–452 (1999).
2) 相澤彰子: 語と文書の共起に基づく特徴度の数量的表現について, 情報処理学会論文誌, Vol. 41, No. 12, pp. 3332–3343 (2000).
3) Belkin, N. J., Oddy, R. N. and Brooks, H. M.: ASK for information retrieval: Part II. Results of a design study, *Journal of Documentation*, Vol. 38, No. 3, pp. 145–164 (1982).
4) Collins, J. J. and Chow, C. C.: It's a small world, *Nature*, Vol. 393, pp. 409–410 (1998).
5) Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol. 19, No. 1, pp. 61–74 (1993).
6) Freeman, L. C.: Centrality in social networks: Conceptual clarification, *Social Networks*, Vol. 1, pp. 215–239 (1979).
7) Granovetter, M.: Strength of Weak Ties, *American Journal of Sociology*, Vol. 78, pp. 1360–1380 (1973).
8) Guare, J.: *Six Degrees of Separation: A Play*, Vintage Books, New York (1990).
9) Hisamitsu, T., Niwa, Y. and Tsujii, J.: A Method of Measuring Term Representativeness — Baseline Method Using Co-occuences Distribution —, *Proc. Coling 2000*, pp. 320–326 (2000).
10) 伊里正夫, 古林隆: ネットワーク理論, 日科技連出版社 (1976).
11) Kageura, K. and Umino, B.: Methods of Automatic Term Recognition, *Terminology*, Vol. 3, No. 2, pp. 259–289 (1996).
12) Kautz, H., Selman, B. and Shah, M.: The Hidden Web, *AI magagine*, Vol.18, No.2, pp.27–35 (1997).
13) Kochen, M.(ed.): *The Small World*, Ablex Publishing Corporation, New Jersey (1989).
14) Luhn, H. P.: A statistical approach to mechanized encoding and searching of literary information, *IBM Journal of Research and Development*, Vol. 1, No. 4, pp. 390–317 (1957).
15) Marchiori, M. and Latora, V.: Harmony in the small-world, *Physica A*, Vol. 285, pp. 539–546 (2000).
16) Mathias, N. and Gopal, V.: Small worlds: How and why, *Physical Review E*, Vol. 63, No. 2 (2001).
17) Matsuo, Y., Ohsawa, Y. and Ishizuka, M.: A Document as a Small World, *Proceedings the 5th World Multi-Conference on Systemics, Cybenetics and Infomatics (SCI2001)*, Vol. 8, pp. 410–414 (2001).
18) Milgram, S.: The small-world problem, *Psychology Today*, Vol. 2, pp. 60–67 (1967).
19) Miller, G. A.: WordNet: A lexical database for English, *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41 (1995).
20) Montoya, J. M. and Solé, R. V.: Small World Patterns in Food Webs, *to appear* (2000). URL://www.santafe.efu/sfi/publications/Working-Papers/00-10-059.pdf.

21) 長尾真, 水谷幹男, 池田浩之: 日本語文献における重要語の自動抽出, 情報処理, Vol. 17, No. 2, pp. pp.110–117 (1976).
22) Niwa, Y., Iwayama, S. and Takano, A.: Topic graph generation for query navigation: Use of frequency classes for topic extraction, *Peoc. of NLPRS'97*, pp. 95–100 (1997).
23) 大澤幸生, ネルス E. ベンソン, 石塚満: KeyGraph: 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会誌, Vol. J82-D-I, No. 2, pp. 391–400 (1999).
24) Salton, G. and Yang, C. S.: On the Specification of Term Values in Automatic Indexing, *Journal of Documentation*, Vol. 29, No. 4, pp. 351–372 (1973).
25) 寺本陽彦, 宮原豊, 松本俊二: 類似文書検索のためのタームの共起語分布分析による計算, 情報処理学会第 59 回全国大会論文誌, IP-06 (1999).
26) 徳永健伸: 情報検索と言語処理, 東京大学出版会 (1999).
27) Watts, D.: *Small worlds: the dynamics of networks between order and randomness*, Princeton (1999).
28) Watts, D. and Strogatz, S.: Collective dynamics of small-world networks, *Nature*, Vol. 393, pp. 440–442 (1998).
29) 安田雪: ネットワーク分析, 新曜社, 東京 (1997).

(平成 年 月 日受付)
(平成 2002 年 3 月 20 日採録)

松尾 豊

1997 年東京大学工学部電子情報工学科卒業．2002 年同大学院博士課程修了．博士（工学）．現在，独立行政法人産業技術総合研究所サイバーアシスト研究センター勤務．推論，数理計画法，テキストマイニングに興味を持つ．人間にとって価値の高い情報の提示を目指している．人工知能学会，電気学会，AAAI 各会員．

大澤 幸生（正会員）

1990 年東京大学工学部電子卒業．1995 年同大学院博士課程修了．博士（工学）．大阪大学助手を経て，現在，筑波大学大学院経営システム科学専攻助教授．AAAI，IEEE，人工知能学会各会員．人工知能学会において，1994 年・1998 年全国大会優秀論文賞，1998 年人工知能学会誌の優秀論文賞受賞．

石塚 満（正会員）

1971 年東京大学工学部電子卒業．1976 年同大学院博士課程修了．工学博士．同年 NTT 入社，横須賀研究所．1978 年東京大学生産技術研究所助教授，1992 年工学部電子情報工学科教授．2001 年より情報理工学研究科電子情報学専攻．研究分野は人工知能，知識処理，マルチモーダル擬人化エージェント，ネットワーク化知的情報環境．IEEE，AAAI，人工知能学会，映像情報メディア学会，画像電子学会等の会員．