2008－NL－188 (3)
2008／11／26

Small World構造を用いた関連記事集合中の重要記事の判定

江越良太† 永井秀利† 中村貞吾†

近年，インターネット上のニュースサイトは情報取得のツールとしての役割が高まり，多くの人に利用されている．毎日膨大な量の記事が掲載されるので，情報を効率よく処理するための技術が求められる．記事から情報を取得する過程での問題のひとつとして，システムによって得た関連記事には内容の重複する冗長な記事が多く含まれ，必要な情報を得るための記事が見つけにくいという問題がある．

本論文では，関連記事集合から冗長な記事を取り除き，話題の移り変わりや全体の情報を理解するために必要な記事を抽出する方法として，Small World 構造を用いた手法を提案する．Small World 構造とは，グラフ内にいくつかのクラスタが形成されているにも関わらず，2 ノード間の平均最短パス長が短いという性質を持ったグラフ構造である．記事をノードとし，関連のあるものにリンクを結んだ関連グラフが，この Small World 構造の性質を備えていることを示し，その性質を利用した重要記事の抽出を行う．この手法により，関連記事内の細かな話題を結びつける記事を見つけることができると考える．

Extraction of Important Articles from Related Articles using Small World Structure

Ryota Egoshi,† Hidetoshi Nagai† and Teigo Nakamura†

In recent years, news sites on the Internet becomes important as the tool of information acquisition. Because a lot of articles are published every day, the efficient technology that acquire information is necessary. A lot of redundant articles are included in related articles that a web search system acquired. So it is difficult to find articles which contain truly required information.

We propose a technique that uses the Small World structure as a method of extracting necessary articles to understand change of topic and whole information in related articles. The Small World structure is a graph structure with characteristics that the average shortest path length between two nodes is short though there are some clusters in the graph. A related graph, where nodes represent related articles and edges represent relation between the nodes, have characteristics of Small World. We think that we can find important articles that connect topics in related articles by this technique.

1. はじめに

近年，インターネット上のニュースサイトは情報取得のツールとしての役割が高まり，多くの人に利用されている．新聞やテレビなどと異なり記事の量を制限する必要がないため，毎日膨大な量の記事が掲載される．利用者は様々な情報を得ることができるが，一方でそれらの大量の情報を効率よく処理することが求められる．

本論文では，ニュースサイトの大量の記事から情報を取得するためのシステムの機能のひとつとして，関連記事集合を縮小する手法を提案する．ここでの関連記事とは当該の話題についての記述がある記事をいい，関連記事集合とはその記事の集まりのことを指す．方法として，Small World 構造を用いる．Small World 構造とは，グラフ内にいくつかのクラスタが形成されているにも関わらず，2 ノード間の平均最短パス長が短いという性質を持ったグラフ構造である．本論文では，記事をノードとし，関連のあるものにリンクを結んだ関連グラフが，この Small World 構造の性質を備えていることを示し，その性質を利用した重要記事の抽出を行う．この手法により，関連記事集合内の細かな話題を結びつける記事を見つけることができ，関連記事集合を縮小できると考える．

† 九州工業大学大学院
Kyushu Institute of Technology

2. 情報取得支援システム

我々は，ニュースサイトの記事から情報を取得する

ためのシステムの開発を行っている．これは，日常テレビや新聞でニュースを見るのと同様に情報収集することを目的としたシステムであり，情報を完全に網羅することよりも，ある程度の情報の網羅性を確保した上で利用者が手軽に使えることに重点をおいている．

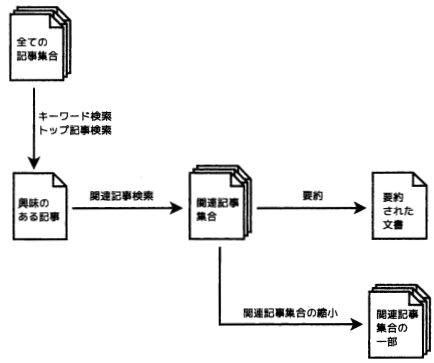

図 1 システムでの主な情報取得の流れ

ひとつの情報取得の流れを説明する．利用者はまずキーワード検索やトップ記事検索によって興味のある記事を見つける．その記事の内容についてさらに詳しく知りたいときには，関連記事検索によって類似の記事や関連する記事を検索する．このとき，話題性の高いものは内容の重複する記事を含む何十件もの記事が関連記事として判定されることがある．それらの記事を全て読むのは，時間がかかる上に何度も繰り返し同じ情報を読むことになり，効率が悪い．そのような場合には，それらの記事を要約して１つの文書にし，利用者はそれを読むことで当該の話題の全般を知ることができる．

これらが全て円滑に進めば問題はないが，何十件もの記事からの自動文書要約は精度が悪く，利用者がスムーズに読める程の要約文書作成は難しい．また，計算コストの面でも実用に耐えない．

そこで，内容の重複する記事やあまり重要でない記事を取り除き，関連記事集合を縮小する．冗長な内容が記事内に多少含まれていても，利用者が全てに目を通せる程度の数にまで記事数を減らすことができれば，自動要約の機能を必要としない．

本論文では以降で関連記事集合の縮小の手法について述べる．

3. Small World 構造

ここでの関連記事集合は，同じ話題について書かれた記事の集まりである．同じ話題の集まりの中でも，それぞれの細かい話題に分かれるため，これを関連性の強い記事同士でリンクを結んだグラフ構造としてみたとき，関連記事集合の中にはそれぞれの細かい話題を持った複数のクラスタが形成されると予想される．このクラスタ内の記事は互いに似通った情報を持っているので，その中から代表となるような記事を選び，それぞれのクラスタの代表だけに絞り込むことで，内容の重複するような冗長な記事を取り除くことができると考える．また，全ての記事は同じ話題について書かれているので，それぞれのクラスタは結び付いている．その結び付きが細かい話題同士を結び，話題のつながりを捉えるのに有効であると考える．

以上の仮定に基づくと，関連記事集合の関連グラフは，全体にある程度のまとまりがあり，その中に複数のクラスタが形成されているグラフであるといえる．このような性質を持ったグラフとして，Small World 構造がある．

Small World とは社会心理学者の Stanley Milgram が任意の 2 人が平均何人のつながりによってつながっているかという実験により提唱されたものである[1]．実験というのは，アメリカネブラスカ州に住む 160 人の住人から，2000km 以上離れたマサチューセッツ州に住むまったく面識の無いゴールの人物まで，知人経由で手紙を転送するというものである．この結果，平均 5.5 人の仲介によって手紙が届けられた．この実験結果に対して，Duncan Watts らがグラフにおける特徴量 L，C の定式化を行なった[2][3]．L と C の定義を以下に示す．

L(Caravteristic Path Length): すべてのノードの組についての最短パス長の平均.

C(clustering coefficient): まず，あるノードが n 個のノードと直接リンクしているとき，この n 個のノード間にあるリンクの数を ${}_nC_2$ で割る．この値についてすべてのノードにおける平均.

C は，あるノードから直接リンクがあるノード同士が直接リンクしている割合である．これはクラスタが形成されているかどうかの指標となる．同じノード数，リンク数でランダムにリンクを結んだランダムグラフでの L と C を L_{rand}，C_{rand} とすると，Small World は $L \geq L_{rand}$，$C \gg C_{rand}$ であるというのが，Watts らの Small World の定式化である．

このようにして定義されたグラフは，複数のクラスタに分かれているのに，任意の 2 ノードの最短パス長がランダムグラフと同程度に短いという性質を持っている．これは，Stanley Milgram の実験により示された，いくつもの知人関係のまとまりに分かれているのに，面識の無い人物へのつながりが近いことと一致し，また，関連記事集合において，複数の細かい話題のまとまりに分かれているのに，それらのまとまりがつながっている状態とも近いと考えられる．

松尾ら[4] は，Small World 構造の性質を持つグラフには，最短パス長の平均 L を小さくする要因であ

る，クラスタを結ぶノードがあることに着目した．松尾らは，単一文書内の単語をノードとした共起グラフがSmall World構造の性質を備えているとして，共起グラフ中にあるクラスタ間を結ぶノードは異なる話題を結びつける単語であると考えた．これを見つけることで文書からキーワードを抽出する手法を提案した．

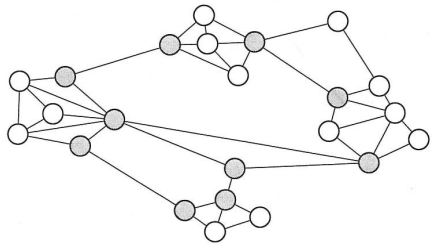

図 2 Small World とクラスタを結ぶノードの例

図2でSmall Worldの例を示し，クラスタ間を結ぶノードを灰色で示している．これを関連グラフとしてみたとき，各クラスタのノードはおおよそ同じ話題についての情報を持っていると考えると，灰色で示されたノードは関連グラフ全体の骨組みを形成していると見ることもできる．

我々の目的は，関連記事集合内の各話題の中から代表となる記事を選ぶことであり，それらの記事が各話題を結び付けていることが必要である．関連記事集合から得られるグラフがSmall World構造の性質を備えていれば，クラスタを結ぶノードがこのような各話題の代表となる記事であると考えられる．そこで，松尾らが提案しているクラスタ間を結ぶノードを見つけるという手法を，関連記事集合から作った関連グラフに応用し，記事集合の縮小を行う．

4. 記事集合の関連グラフの作成

関連記事集合の，記事をノードとし，内容の重複する部分の多い記事の組にリンクを結んだ関連グラフを作る．

4.1 ノードの特徴ベクトルの作成

ノード間の関連にはベクトル空間モデル[5]を用いる．まず，ノードの特徴ベクトルを作るため，記事に含まれる単語を抽出する．単語の抽出には形態素解析器MeCab[6]を使用した．抽出する語は名詞全般であり，名詞が連続して出現する場合には複合名詞として使用する．全ての記事の単語集合を特徴ベクトルの軸とし，各記事における各単語の$tfidf(i)$[7]を軸の値とする．$tfidf(i)$とは以下の式により求められる，単語の重要度である．

$$tfidf(i) = tf(i) \cdot idf(i)$$

$$tf(i) = \frac{n(i)}{\sum_k n(k)}$$

$$idf(i) = \log \frac{|D|}{|d : d \ni t(i)|}$$

$n(i)$は記事中での単語iの出現数である．$|D|$はデータベース上の記事の総数，$|d : d \ni t(i)|$は単語iを含む記事数である．$tf(i)$(Term Frequency)は単語iが当該記事中に出てくる頻度を表している．一方$idf(i)$(Inverse Document Frequency)は単語iの稀少さを表している．

4.2 リンクの作成

リンクを作成するために，まず全てのノードの組についての関連度を求める．特徴ベクトルのコサイン相関値を用いて，ベクトルの類似度を計り，ノードの関連度とする．ベクトルA，Bのコサイン相関値$sim(A, Q)$は以下のようになる．

$$sim(A, B) = \frac{\vec{A}\vec{B}}{|\vec{A}||\vec{B}|} = \frac{\sum_{i=1}^{n}(w_i \cdot q_i)}{\sqrt{\sum_{i=1}^{n}(w_i)^2 \sum_{i=1}^{n}(q_i)^2}}$$

コサイン相関値$sim(A, Q)$は2つの特徴ベクトルがどのくらい似ているかを示しているが，特徴ベクトルを作っている記事内の単語集合は記事の内容を抽象的に表していると考えられるので，$sim(A, Q)$は記事の内容がどのくらい重複しているかを示しているといえる．

次に，関連度の高いノードの組から順に，1つのノードから出るリンク数の平均が閾値kになるまでリンクを張る．

4.3 LとCの検証

関連記事集合を関連グラフにしたとき，そのグラフがSmall World構造の性質を備えているか，LとCを検証した．関連記事集合10件の平均と，ランダムグラフ50件の平均を表1に示す．ここでは$k = 3.0$でグラフを作成している．LはL_{rand}の1.1倍と，同程度であるのに対して，CはC_{rand}の約8.4倍である．$C \gg C_{rand}$がSmall Worldの定式化のひとつであるが，Wattsらの解析[2]では，CがC_{rand}の5.6倍であるノード数200程のグラフもSmall World構造であるとしている．今回作成した関連グラフはノードが30から50程度と少ないのにWattsらの解析したグラフよりもCがC_{rand}に比べて大きく，ノードの偏りがあるため，$C \gg C_{rand}$であると考えられる．以上より，今回作成した関連グラフはSmall World構造の性質を備えているといえる．

5. 重要ノードの抽出

関連グラフのSmall World構造を用いて，欠けて

表 1　関連グラフの Small World 特徴量

L	L_{rand}	C	C_{rand}
2.05	1.97	0.56	0.067

しまうと話題のつながりや内容の網羅性が失われてしまうような記事を判定する．このような記事に相当するノードを重要ノードと呼ぶ．

先に述べたように，Small World には L を減少させる要因となっている，クラスタ間を結ぶノードが存在し，これらのノードが我々の検出したい重要ノードである．これらのノードは，欠けてしまうとクラスタ間が切り離されたり，ノード間の最短パス長が長くなるなどして，L が増加することになる．松尾らの手法と同様に，この性質を利用して重要ノードを検出する．

最短パス長の増加を正しく計るため，ノード間にパスがない場合のパス長を定める必要がある．ノード i とノード j の最短パス長 $d'(i,j)$ を次のように再定義する．

$$d'(i,j) = \begin{cases} d(i,j), & x \text{ 連結している場合} \\ w_{sum}, & x \text{ それ以外} \end{cases}$$

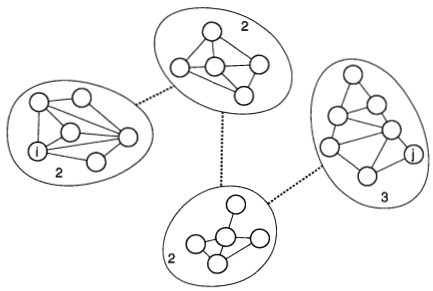

図 3　w_{sum} の例

$d(i,j)$ は連結しているノードでの通常の最短パス長である．w_{sum} は定数であり，すべての連結サブグラフの幅の和である．グラフの幅とは，グラフ中の最短パス長の最大となるものである．つまり，連結していない 2 ノードが，新たなリンクにより連結される場合のうち，最も最短パス長が長いものを与えていることになる．例えば図 3 では 4 つのサブグラフがあるが，i と j の最短パス長 $d'(i,j)$ は 9 である．当然，連結しているどの 2 ノード間最短パス長よりも短くなることはない．

さらに，L を拡張して $L'(v)$，$L'_G(v)$ を以下のように定義する．

$L'(v)$:　ノード v 以外のすべてのノードの組についての最短パス長 $d'(i,j)$ の平均．

$L'_G(v)$:　グラフからノード v を取り除いたときのすべてのノードの組についての最短パス長 $d'(i,j)$ の平均．

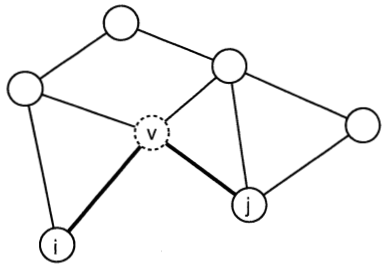

図 4　$L'(v)$ の例

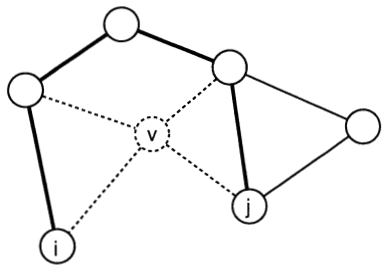

図 5　$L'_G(v)$ の例

$L'(v)$ では，ノード v を経由するパスはそのままで，2 ノード間平均最短パス長を計算する．パスは消えていないので，例えば図 4 では i と j の最短パス長は 2 である．一方 $L'_G(v)$ では，ノード v がグラフから取り除かれてしまうため，ノード v を経由して最短パスを得ていたノードの組は，他の迂回するパスが最短パスとなったり，あるいは連結が途切れてしまう．図 5 では迂回するパス長 4 の経路が最短パスとなる．つまり，$L'(v)$ と $L'_G(v)$ の差 $CB(v)$ は，いかにノード v が L の減少に貢献しているかを示している．

$$CB(v) = L'_G(v) - L'(v)$$

すべてのノードについてこの重要度 $CB(v)$ を計算し，閾値 t よりも大きいノードを重要ノードとする．この手法はクラスタが形成されていることに基づいているが，クラスタ自体を見つけるわけではないため，クラスタの厳密な定義を必要としない．

6. 複数のノードがクラスタを結ぶ際の問題と回避方法

松尾らのキーワード抽出では，対象が論文中の単語

である．論文では，著者が読者に分かりやすいような話題の展開を行うため，比較的形の整ったグラフとなる．一方，関連記事集合では，各記事の記事同士の関わりについて考慮されて発行されているわけではないので，一部の記事の内容がほとんど同じであることがある．

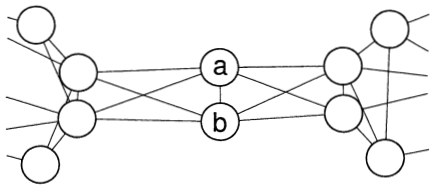

図 6 2 つのノードがクラスタを結ぶ例

図 6 では，ノード a とノード b がクラスタを結んでいる．この状態でノード a の重要度 $CB(a)$ を計算しようとノード a をグラフから取り除いても，ノード b がクラスタ間を結んでいるため $L'_G(a)$ は全く増加せず，ノード a の重要度 $CB(a)$ は 0 となってしまう．同様に，ノード b の重要度 $CB(b)$ も 0 である．しかし実際にはこの 2 つのノードはクラスタ間を結んでいるため，2 つの話題を結びつける記事であり，どちらか一方を重要ノードとして判定する必要がある．このようなことは，ノード a とノード b の内容がほとんど同一のときに起こる．内容がほとんど同一であれば，リンクのしかたがほとんど変わらず，どちらのノードを通っても最短パス長が同じになる．ノード a とノード b が小さなクラスタを形成していると考えることもできるが，$CB(b)$ ではこのようなノード対をクラスタとみなして，このクラスタの代表となる重要ノードの抽出を行うことができない．

この問題を回避する手段として，関連度が非常に高い記事の組がある場合には，グラフを作るときにそれらの記事のうち 1 つだけをノードとして追加するという方法もあるが，そのためには関連度がどのくらい高い場合にグラフ構造として上記のような問題が起こるかが分かっている必要がある．しかし，関連度の計算に用いられている $tfidf(i)$ はコーパスの文書の量などに依存する上，各記事の関連度の高さの分布が関連記事集合によってまちまちで，閾値をあらかじめ与えることは難しい．

そこで，問題が起こる場合を直接検出するために，複数のノードを同時にグラフから取り除いた場合の重要度 $CB(\{u,v\})$ を計算して問題となるノードを見つけ，拡張した重要度 $CB'(v)$ を与える．

(1) グラフ中の全てのノードの $CB(v)$ を計算する．
(2) 直接リンクのあるノードの組 $\{u,v\}$ を取り出す．
(3) $CB(\{u,v\})$ を計算する．
(4) $CB(\{u,v\}) > CB(u)$ または $CB(\{u,v\}) > CB(v)$ のとき，$CB(u)$，$CB(v)$ の値の大きい方を CB' として $CB(\{u,v\})$ の値に修正する．$CB(u) = CB(v)$ の場合には日付の新しい記事の重要度を CB' として $CB(\{u,v\})$ の値に修正する．

このようにして新たに得た重要度 $CB'(v)$ が閾値 t よりも大きければ重要ノードとして追加する．ノード 3 つ以上によってクラスタが結ばれているような場合も考えられるが，今回は 2 つのノードの組までで重要度 $CB'(v)$ の計算を行った．

7. 評　　価

重要度 $CB'(v)$ を用いて必要な記事が抽出できるかの評価実験を行った．評価の基準となるのは，以下の 2 つである．

1. 情報が網羅されていて，話題の移り変わりが読み取れるように記事が抽出されているか．
2. 内容の重複する記事が含まれていないか．

2 つめの条件に関しては，一部重複する箇所があっても，その他の部分に固有の情報が含まれていれば抽出するという条件とする．

実験は，7 件の事件に関連するそれぞれの記事集合を，先述の基準に基づき人手で縮小し，重要度 $CB'(v)$ を用いた手法と比較することで行った．テストデータは Yahoo!ニュース (http://headlines.yahoo.co.jp) に 2008 年 4 月から 9 月の間に掲載された記事 13783 件から，先述の $tfidf(i)$ による特徴ベクトルのコサイン相関値で関連記事検索を行って得た関連記事集合を用いた．本手法で抽出され人手でも必要であると判断された記事数を正解数とし，人手で必要であると判断されたが本手法では抽出されなかった記事数を不足数，本手法で抽出されたが人手では不要と判断された記事数を不正解数として，結果を表 2 に示す．1 ノードからのリンク数平均は $k = 3.0$ とし，重要度 $CB'(v)$ の閾値は $t = 0$ として，少しでも重要度のあるものは全て重要ノードとした．

表 2 本手法と人手との比較評価

記事集合	記事数	正解数	不正解数	不足数
A	30	3	2	2
B	35	4	1	2
C	27	5	0	3
D	34	5	2	2
E	32	3	0	3
F	45	5	2	3
G	37	4	1	2
Total	240	29	8	17

7 つの関連記事集合全体において，本手法により抽

出された記事が人手でも必要であると判断された割合は約 78.4 %である．残りの約 22.6 %は冗長な記事が含まれている割合であり，全体の約 84.2 %を占める冗長な記事を減らしたという点ではまずまずの成果であるといえる．

一方，人手で必要であると判断された記事が本手法で抽出された割合は約 63.0 %で，4 割近くの必要な記事を取りこぼしている．閾値は $t = 0$ で行っているため，重要度 $CB'(v)$ では抽出できないことになる．この原因としてまず，クラスタ内に複数の話題が含まれていることが挙げられる．記事の中には 1 つだけの情報ではなくさまざまな情報が含まれているので，それぞれの部分同士が関連しあい，クラスタを作ってしまう．その結果複数の話題が 1 つのクラスタ内に存在することになり，クラスタ間を結ぶノードを見つける本手法では，クラスタ内に埋もれてしまっている記事を見つけることができない．また，リンクが 1 つしかないノードの重要度 $CB'(v)$ は 0 となってしまう問題もある．どのノード対も結んでいないので他のノード対の最短パスとなることがないためである．実際にはこのようなノードは固有の情報を持っていることが多い．複数の話題を結ぶ記事を見つけるこの手法では新しく分かった話題についての記事は見つけにくいという欠点があるといえる．

8. おわりに

本論文では，関連記事集合の関連グラフが Small World 構造の性質を備えていることを示し，その性質を利用して話題と話題を結びつける重要な記事を判定する手法を提案した．また，関連記事集合はそれぞれの記事同士の関わりを考慮して作られたものではないため，類似度が高すぎる記事が混在している場合の構造上の問題があり，その問題に対する対処法を示した．本手法は，冗長な記事を取り除くという点では有効であるが，一方で重要な記事の取りこぼしがある．今後は必要な記事を抽出するため，手法の改善を行う予定である．

参考文献

1) Milgram, Stanley.: The small-world problem, Psychology Today, Vol.2, pp.60-67, 1967.
2) Watts, Duncan J. and Strogatz, Steven H.: Collective dynamics of small-world networks, Nature, Vol.393, pp.440-442, 1999.
3) Collins, J. J. and Chow, C. C.: It's a small world, Nature, Vol.393, pp.409-410, 1998.
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.143, No.12, pp.1825-1833, 2002.
5) G. Salton.: The Vector Space Model, Automatic Text Processing, pp.312-325, 1985.
6) MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/
7) G. Salton. and M. J. McGill.: Introduction to Modern Information Retrieval, McGraw-Hill Book Company, 1983.