カーネル法による引用解析: 複数コミュニティが存在する場合

伊藤　敬彦† 　新保　仁† 　持橋　大地†† 　松本　裕治†

† 奈良先端科学技術大学院大学 情報科学研究科
630-0192 奈良県生駒市高山町 8916-5
†† ATR 音声コミュニケーション研究所 音声言語処理研究室
E-mail: †{takahi-i,shimbo,matsu}@is.naist.jp, ††daichi.mochihashi@atr.jp

あらまし　本稿では Kandola らのノイマンカーネルを複数のコミュニティが存在する引用関係グラフに適用する際に生ずる問題点について議論する. ノイマンカーネルはパラメタ調整によって, 個々の論文に対して, 関連論文あるいは重要論文をランク付けして提示できる. しかし, パラメタを重要度に偏らせると引用グラフ全体における支配的なコミュニティの重要論文のみが, 個々の論文が属するコミュニティにかかわらず, 上位にランキングされてしまう. これに対して我々は, 引用の生成過程を Hofmann の pLSI (probabilistic Latent Semantic Indexing) によりモデル化し, その結果を用いて, コミュニティ (トピック) ごとの重みつき引用グラフ (隠れトピックグラフ) を計算する. これらのグラフは, 引用の各トピックに対する帰属確率を計算し, 弧の重みとするため, 同一の引用が, 複数の隠れトピックグラフに異なった重みのもとで存在することができる (多重トピック). これらの隠れトピックグラフにノイマンカーネルを適用することによって, 対象論文が属するコミュニティを考慮して重要論文を推薦，提示できる.

キーワード　HITS, グラフカーネル, pLSI, リンク解析

Kernel-based Citation Analysis for Graphs with Multiple Communities

Takahiko ITO†, Masashi SHIMBO†, Daichi MOCHIHASHI††, and Yuji MATSUMOTO†

† Graduate School of Information Science, Nara Institute of Science and Technology
†† ATR Spoken Language Communication Research Laboratories
E-mail: †{takahi-i,shimbo,matsu}@is.naist.jp, ††daichi.mochihashi@atr.jp

Abstract　In this paper, we discuss issues raised by applying Neumann kernels to large citation graphs that have multiple communities. Neumann kernels can identify not only papers related a given paper but also the most important papers in a graph. However, when Neumann kernels are biased towards importance, top-ranked papers are the important papers in the dominant community of the graph irrespective of the communities where the target paper is cited. To solve this problem, we model the generation process of citations by pLSI (probabilistic Latent Semantic Indexing), and construct a weighted graph (hidden topic graph) for each community (topic). Applying Neumann kernels to hidden topic graphs, we can rank papers on the basis of the communities in which they appear.

Key words　HITS, Graph kernels, pLSI, Link Analysis

1. ま え が き

科学技術文献間の引用関係は, 個々の文献を特徴付ける有用な情報である. この情報を活用して文献間の関係を測定する各種の尺度 (引用解析尺度あるいは計量書誌尺度) が, 古くから計量書誌学 (bibliometrics) において提案されてきた. その中で, 共引用や書誌結合といった文書間の ‘関連度’ はもっとも有名な尺度である.

また, PageRank [3] や HITS [14] に代表される Web ページの重要度算出法も, ページ間の引用 (リンク) 関係を基に個々のページの重要度を算出するため, 引用解析尺度の一種と見なせる. 以下, 科学文献, Web ページをまとめて単に ‘文書’ と呼ぶ.

現在までに我々は, グラフ上で定義されたカーネル法を引用解析に適用することで, 二つの引用解析尺度 (重要度, 関連度) に統一的な解釈を与えた. 具体的には, グラフ上で定義されるパラメタつきカーネル関数族 (ノイマンカーネル [12]) におけるパラメタのとりうる範囲の両端点にあたる二つのカーネルは, 各々, 共引用に基づく関連度と, (HITS における権威度で与えられる) 文書の重要度に対応する. 結果として, ノイマンカーネルは二つの端点 (重要度と関連度) 間に存在する任意の中点とみなせ, パ

ラメタを調整することで重要度あるいは関連度へ偏らせることができる.

しかし, ノイマンカーネルのパラメタを重要度に偏らせた場合, 引用グラフ全体における支配的なコミュニティの重要文書のみが個々の文書が属するコミュニティにかかわらず, 上位にランキングされてしまうという問題が存在することが分かった.

このため, 我々は引用の生成過程を Hofmann の pLSI (probabilistic Latent Semantic Indexing) によりモデル化し, その結果を用いて, コミュニティ (トピック) ごとの重みつき引用グラフ (隠れトピックグラフ) を計算する. これらグラフは, 引用の各トピックに対する帰属確率を計算, 弧の重みとするため, 同一の引用が, 複数の隠れトピックグラフに, 異なった重みのもとで存在することができる (多重トピック).

これらの隠れトピックグラフにノイマンカーネルを適用することによって, 対象文書が属するコミュニティを考慮して重要文書を推薦, 提示できることを示す.

2. 背　　景

本章では, 始めに本稿を通して使用する用語, 記法を導入する. 次に, 本論文を通して議論する 2 種類の引用解析尺度 (関連度, 重要度) 及び, ノイマンカーネルを説明する.

2.1 記　　法

引用解析尺度はいずれも**引用グラフ**を用いて定義される. 引用グラフとは, 単純有向グラフ $G=(V,E)$ であり, 節点 ($\in V$) は文書を, 弧 ($\in E \subset V \times V$) はそれらの間の引用をモデル化したものである. 文書対 $(i,j) \in V \times V$ に対して, $(i,j) \in E$ となるのは文書 i が文書 j を引用する場合に限られる.

あわせて, 以下の定義と記法を用いる. 重みつきグラフは 3 項組 (V,E,w) と定義され, ここで (V,E) は (重みなし) 単純有向グラフ, $w : E \to (0,\infty)$ は, 弧のラベル (重み; 正の実数) を定める重みづけ関数である. 重みなしグラフは, 全ての $(i,j) \in E$ に対し $w(i,j)=1$ であるような, 重みつきグラフの一種とみなせる. 重みつきグラフ $G=(V,E,w)$ に対し, その節点集合 V を $V(G)$, 弧集合 E を $E(G)$ と表す. また, $|V| \times |V|$ 行列 A で, 全ての $(i,j) \in E$ に対して, $A(i,j)=w(i,j)$, それ以外は 0 なる行列を G の隣接行列と呼び, $A(G)$ と書く. 重みつきグラフ G が無向グラフであるとは, $A(G)$ が対称行列であることを言う. また, 経路のコストを, 経路に含まれる全ての弧の重みの積と定義する. したがって, 弧の重みが 2 節点間を結ぶ弧の多重度 (本数) を表す整数の場合 (いわゆる多重グラフ) には, 経路コストは経路の総数 (多重度) を表す.

2.2 引用解析尺度

2.2.1 関　連　度

書誌結合 [13] と共引用解析 [16] は引用関係から文書間の関連度 (類似度) を求める最も一般的な手法である. 例えば, 科学文献検索サービス CiteSeer [15] は, 個々の引用に対するヒューリスティクな重みづけと共引用解析を併用して関連文献を提示する.

共引用解析において, 文書間の関連度は双方を同時に引用する文書数と定義される. 逆に, 書誌結合において, 文書間の関連度は同一の文書を双方が引用する数によって与えられる. また, これらの尺度は引用グラフに基づいて定義できる.

定義 1　$A=A(G)$ を引用グラフ G の隣接行列とする. 対称行列 $A^{\mathrm{T}}A$ を G の**共引用行列**と呼び, この行列を隣接行列とみなすことで得られる重みつき無向グラフを, G の**共引用グラフ**と呼ぶ. G の**書誌結合行列**と**書誌結合グラフ**は $A^{\mathrm{T}}A$ の代わりに AA^{T} を用いて同様に定義される.

図 1 に引用グラフの一例とそれに対応する書誌結合グラフ及び共引用グラフを示す. 共引用行列 $A^{\mathrm{T}}A$ の (i,j)-要素は文書 i, j 間の共引用解析の値と一致し, 同様に, AA^{T} の各要素は書誌結合の値を表す.

2.2.2 HITS 重要度

文書の重要度をその内容から推定することは困難であり, 古くから各文書の被引用数が重要度の指標として用いられてきた. Kleinberg の HITS [14] は, 同様の考えに基づくが, より洗練された重要度算出法である.

HITS は, 個々の文書を二つのスコア (権威度とハブ度) で評価する. 直観的には, 権威度が高い文書とはハブ度の高い文書から多く引用される文書であり, 逆にハブ度の高い文書とは権威度の高い文書を多く引用する文書である. 以下のようにハブ度と権威度は相補的に定義される.

定義 2　G と A をそれぞれ引用グラフとその隣接行列とすると, **HITS アルゴリズム**は以下の再帰式を $n=0,1,2,\ldots$ について計算する.

$$a_{n+1}=\frac{A^{\mathrm{T}}h_n}{|A^{\mathrm{T}}h_n|}, \quad h_{n+1}=\frac{Aa_{n+1}}{|Aa_{n+1}|}. \tag{1}$$

ただし, $a_0=h_0=\mathbf{e}$ とする ($\mathbf{e}$ は全ての要素が 1 のベクトル). すると, 文書 i の**権威度**は**権威度ベクトル** $\lim_{n\to\infty}a_n$ の i-要素で与えられ, **ハブ度**は**ハブ度ベクトル** $\lim_{n\to\infty}h_n$ の i-要素で与えられる.

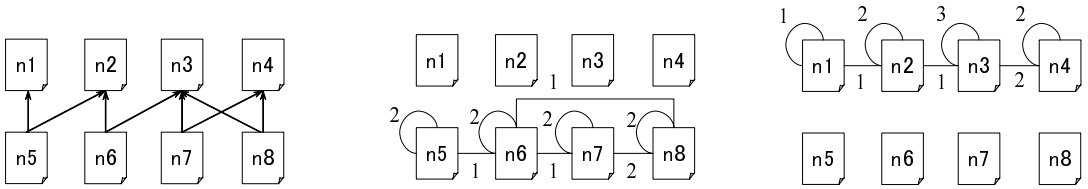
Kleinberg は, HITS の権威度ベクトル $\lim_{n\to\infty}a_n$ とハブ度ベクトル $\lim_{n\to\infty}h_n$ がそれぞれ, 共引用行列 $A^{\mathrm{T}}A$ と書誌結合行列 AA^{T} の最大固有値に対応する固有ベクトル (最大固有ベクトル) に一致することを示した.

2.3 ノイマンカーネル

Kandola ら [12] は文書間の類似性を文書内の単語を元に計算するために**ノイマンカーネル**を提案した.

我々は先行研究 [10] において, このノイマンカーネルを文書の類似性を分析するために使用するのではなく, 引用グラフに対して適用し, 引用解析尺度の観点からその性質を分析した. 具体的には, 引用グラフにおける隣接行列 A を文書-単語行列の代わりに用い, $K=A^{\mathrm{T}}A$ と $M=AA^{\mathrm{T}}$ を生成する. 行列 K, M はそれぞれ共引用行列と書誌結合行列と一致する. このとき引用解析におけるノイマンカーネルは次式で与えられる.

$$\hat{K}_\gamma=\sum_{n=1}^{\infty}\left(\frac{\gamma}{\lambda}\right)^{n-1}\left(A^{\mathrm{T}}A\right)^n, \quad \hat{M}_\gamma=\sum_{n=1}^{\infty}\left(\frac{\gamma}{\lambda}\right)^{n-1}\left(AA^{\mathrm{T}}\right)^n. \tag{2}$$



(a) 引用グラフ　(b) 書誌結合グラフ　(c) 共引用グラフ

図 1　(a) 引用グラフとその (b) 書誌結合グラフ, および (c) 共引用グラフ

ここで, $0 \leqq \gamma < 1$ はパラメータ (拡散係数), λ は $A^{\mathrm{T}}A$, 及び AA^{T} のスペクトル半径を表す. 任意のグラフ G に関して, $(A^{\mathrm{T}}A)^n$ の (i, j)-要素は, G の共引用グラフ $A^{\mathrm{T}}A$ 上で, 節点 i と節点 j を両端点に持ち, かつ長さが n であるような経路の総数を表している. このことから, 式 (2) のノイマンカーネルは共引用 (書誌結合) グラフ上の各節点間の全ての経路の数の重みつき和を求めていることになる.

n が小さいとき, 節点 i, j を結ぶ経路の数は, これら 2 節点間の距離がどれだけ近いかに左右される. 実際, 共引用グラフ上の節点 i, j 間の距離が n 以上であれば, $(A^{\mathrm{T}}A)^n$ の (i, j)-要素は 0 となる. グラフ上の節点間の距離が近いほど, 対応する文書は内容的に近いという仮定のもとでは, n が小さい場合の $(A^{\mathrm{T}}A)^n$ の各要素は, 各節点間の関連度を表すことがわかる. 特に $n = 1$ のときには, 共引用行列と一致する.

一方, n が十分大きい場合の $(A^{\mathrm{T}}A)^n$ の各要素は, 節点間の距離よりもむしろ, 各節点の大域的重要度を反映したものとなり HITS の重要度に偏る [10] [8].

総合すると, ノイマンカーネル $\hat{K}_\gamma$ は $n = 1$ から ∞ までの $(A^{\mathrm{T}}A)^n$ の重みつき和を計算しており, 共引用行列と HITS の権威度の一種の混合を表していると言える. なお, ノイマンカーネル行列の関連度と重要度への偏りは拡散係数 γ によって決定される. 拡散係数 γ が小さい場合, 短い経路の長さ n の重みが大きくなるため関連度に偏る. 一方で, 拡散係数 γ が 1 に十分近い場合, ノイマンカーネルの各行 (あるいは列) の要素間の大小は HITS の権威度, ハブ度で与えられる順位づけに近づくことが示せる [10].

以下の議論は, 権威度及び $\hat{K}_\gamma$ を中心に行うが, ハブ度及び $\hat{M}_\gamma$ についても成り立つ.

3. ノイマンカーネルの問題点

ノイマンカーネルを引用解析尺度の観点から見ると共引用 と HITS の混合となる尺度であることが分かった. しかしノイマンカーネルは, 共引用 と HITS の混合であるために, HITS の問題点を受け継いでしまう. HITS の問題点とは, 複数コミュニティを持つ (巨大な) 引用関係グラフを扱う場合, 引用グラフ全体における支配的なコミュニティに属する論文のみが上位にランキングされてしまうという問題である.

ここで引用グラフのコミュニティとはコミュニティに属する文書間にコミュニティの外の文書よりも多くの引用がなされている文書集合を表す.

例えば, 図 2 のように複数のコミュニティを持つ引用グラフに対して HITS を計算する. 図 2 では, 文書 n_1 が属するコミュニティ (コミュニティ1) と文書 n_6 が属するコミュニティ (コミュニティ2) の 2 つが存在する. このグラフにおける権威度ランキングは $n_2 > n_1 > n_3 > n_5 > n_4 > n_6$ であるが, コミュニティ2 に存在する文書が不当に低い重要度が与えられてしまっている.

以下, ノイマンカーネルが HITS の問題点を引き継いでしまうことを簡単な例題を用いて検証する.

例 1　複数のコミュニティを持つグラフ (図 2) から導出される共引用グラフ (図 3) に対してノイマンカーネルを適用しその問題点を指摘する.

ノイマンカーネルを拡散係数を大きく設定し重要度に偏らせ ($\gamma = 0.99$) て計算すると, 以下のカーネル行列が得られる. また, 簡単のためにカーネル行列の文書 n_1 から n_6 に対応する部分行列のみを示している. この行列の第 i 要素が文書 n_i に対応する. 他の要素は全て 0 であるため省略した.

$$\hat{K}_{0.99} = \begin{pmatrix} 108.53 & 225.98 & 59.64 & 7.16 & 29.30 & 1.36 \\ 225.98 & 477.37 & 127.64 & 15.33 & 62.70 & 2.90 \\ 59.64 & 127.64 & 37.87 & 5.30 & 21.67 & 1.00 \\ 7.16 & 15.33 & 5.30 & 5.16 & 7.34 & 2.17 \\ 29.30 & 62.70 & 21.67 & 7.34 & 23.74 & 1.39 \\ 1.36 & 2.90 & 1.00 & 2.17 & 1.39 & 1.60 \end{pmatrix}. \tag{3}$$

上の行列の各 i 行が文書 n_i に対する重要度ベクトルである. つまり, i 行目のベクトル中の各要素の相対的な大きさが文書 i から見た各文書の重要度を表す. 例えば 文書 n_3 (3 行目) から見ると, 文書 n_2 の値 (127.64) が最も大きく重要であるといえる.

この行列で, 我々は n_6 (6 行目) に注目する. すると コミュニティ2 に存在する n_6 から見た他の文書のランキングでも, コミュニティ1 に存在する文書 n_2 の値 (2.90) が最も大きく, HITS の権威度ランキング ($n_2 > n_1 > n_3 > n_5 > n_4 > n_6$) に偏った値であることが分かる. しかし, これは我々の直感と大きく異なる. このグラフでは n_6 が属しているコミュニティと n_2 が属しているコミュニティは異なるため, n_6 に対する重要度は n_6 と同一のコミュニティにあり被引用数が多い n_5 の方が n_2 よりも大きいことが望ましい. また一方で, n_6 と同一のコミュニティに属す文書 n_4, n_5 を比較すると n_5 の被引用数は n_4 より多いにもかかわらず, n_6 に対して n_4 が n_5 よりも上位にランクされてしまっている.

さらに拡散係数を大きく設定すると ($\gamma = 0.999$), カーネル行列の全ての行における要素間の大小は HITS の権威度と完全に

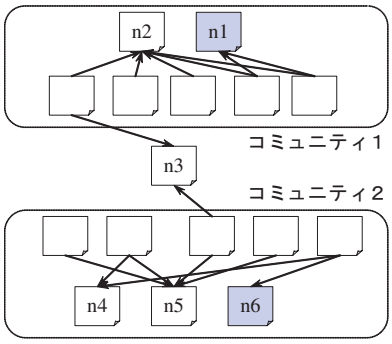


図 2　複数コミュニティを持つ引用グラフ.

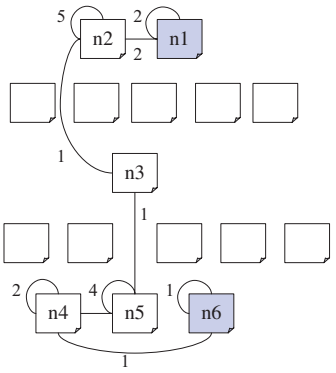


図 3　図 2 より導かれる共引用グラフ.

一致し, 対象文書が存在するコミュニティの情報は完全に無視されてしまう.

このような現象を避けるため, Kleinberg はグラフ全体ではなく, ユーザが入力したキーワードを含む文書, およびそれらと直接の引用関係にある文書のみからなる部分グラフを生成しこの部分グラフに対して HITS アルゴリズムを適用することを提案している. 生成された部分グラフに対して式 (1) を計算することで, ユーザが検索したいトピックと HITS の重要度のランキングにずれが生じるのを防いでいる.

Kleinberg と同様に, ノイマンカーネルを適用する際にユーザに興味のあるキーワードを入力してもらい, キーワードを含む文書からなる部分グラフに対してノイマンカーネルを適用することも考えられる. しかし, これを行うには文書の内容情報を入手し単語を索引する必要があるが, 技術論文は引用関係のみが公開されていることが多く, Web 文書のように内容情報を簡単に収集することは難しい. さらに, たとえユーザが入力したキーワードを含む文書からなる部分グラフを基に HITS を計算してもユーザの検索したいトピックと上位にランキングされる文書が属すコミュニティのずれを完全に防げないことが知られおり (topic drift 問題 [1]) ノイマンカーネルにおいてもこの問題が生じることが考えられる.

4. ノイマンカーネルが持つ問題点の解決

前節で説明した問題を解決するため, 引用の生成モデルを用いて文書の属するコミュニティ毎に重み付き引用グラフを生成し, それらにノイマンカーネルカーネルを適用する. この引用グラフの導出では, Kleinberg の提案手法と異なり, 引用情報のみを用いる (文書の内容情報は必要としない). まず, 本稿で用いる引用の生成モデルについて簡単に説明する.

4.1 引用の生成モデル

pLSI [6] (probabilistic Latent Semantic Indexing) は文書の生成過程をモデル化する文書の生成モデルの一つである. 文書の生成確率は観測できない隠れトピック t を仮定して計算される. たとえば, pLSI において文書 d_i と単語 w_j の共起確率は以下で与えられる.

$$p(d_i, w_j) = \sum_{t=1}^{N} p(t)p(d_i|t)p(w_j|t) \tag{4}$$

ここで, $t \in \{1,2,3...N\}$ は隠れトピックを表す.

Cohn ら [4] は引用の生成過程を文書の生成モデルと同様にモデル化することで各コミュニティ t (文書の生成モデルにおける「トピック」) 毎に重要度を計算できることを示した (pHITS).

具体的には pLSI を引用グラフの隣接行列 A に適用することで引用の生成確率を計算する. また, 文書 d_i に対して行われる引用を c_i で表す. このモデルでは, 文書 d_i が 文書 d_j を引用する確率 (つまり, 文書 d_i と引用 c_j の共起確率) は式 (4) と同様に以下で与えられる.

$$p(d_i, c_j) = \sum_{t=1}^{N} p(t)p(d_i|t)p(c_j|t)$$

Cohn らは pLSI から計算される確率 $p(c|t)$, すなわちコミュニティ t で文書が引用される確率をコミュニティ t における文書の重要度 (pHITS) として使用した.

pHITS はコミュニティ毎に重要度を算出することができるが, ノイマンカーネルのようにパラメタ設定によって文書間の関連性を同時に考慮することができない. そこで我々は, 各コミュニティ毎に引用グラフの弧の重みを変更した隠れトピックグラフを生成し, それらにノイマンカーネルを適用する. こうすることで, pHITS のように単純にコミュニティ毎に重要度のランキングを示すのではなく, ノイマンカーネルの持つ, 重要度と関連度の混合という性質を受け継ぎ, 対象文書に同一コミュニティで重要な文書を推薦，提示できるようになる.

4.2 隠れトピックグラフに対するノイマンカーネルの適用

pLSI を引用グラフに適用する事で, 文書 d_i が d_j を引用した時に (つまり, 文書 d_i と引用 c_j が共起した時に) 引用 c_j がコミュニティ t を表現している確率 $p(t|d_i, c_j)$ が得られる.

この確率 $p(t|d_i, c_j)$ を引用グラフにおける弧の重み $w(i,j)$ とすることで **隠れトピックグラフ** G_t が生成される. 隠れトピックグラフの隣接行列 $A_t = A(G_t)$ は, 引用グラフ上の全ての弧 $(i,j) \in E(G)$ に対して, $A_t(i,j) = p(t|d_i, c_j)$, それ以外は 0 なる行列である. 隠れトピックグラフの隣接行列 A_t を利用して, コミュニティt における共引用行列 $A_t^T A_t$ が生成できる. このとき, $A_t^T A_t$ の (i,j)-要素は文書 d_i, d_j の共引用の値をトピック t 毎に分解したものになっている.

次に我々は隠れトピックグラフの隣接行列 A_t を組み合わせ

てできる共引用行列 $A_t^T A_t$ に対してノイマンカーネルを適用する. トピック t におけるカーネル行列は以下で与えられる.

$$\hat{K}_{t,\gamma} = \sum_{n=1}^{\infty} \left(\frac{\gamma}{\lambda}\right)^{n-1} \left(A_t^{\mathrm{T}} A_t\right)^n \tag{5}$$

ここで, ここで, $0 \leqq \gamma < 1$ はパラメータ拡散係数, λ は $A_t^{\mathrm{T}} A_t$ のスペクトル半径を表す.

最終的に我々は各コミュニティ毎に生成されるノイマンカーネル, 式 (5) を足しあわせて使用する. 正定値カーネルの和は正定値カーネルであることが示されているため [5], 以下の式は一つの正定値カーネルを表している.

$$R_\gamma = \sum_{t=1}^{N} \hat{K}_{t,\gamma} \tag{6}$$

4.3 解　　釈

以下, このカーネルの解釈について簡単な例題を用いて述べる.

例 2　節 3 で使用した引用グラフ (図 2), に対して提案手法 (式 6) を適用する.

はじめに, 拡散係数を小さく設定し関連度に偏らせた ($\gamma = 0.1$) 提案手法を (図 2) に適用した. なお, この実験では隠れコミュニティの数を 2 として pLSI を計算した.

$$R_{0.1} = \begin{pmatrix} 2.14 & 2.25 & 0.04 & 0.00 & 0.00 & 0.00 \\ 2.25 & 5.54 & 1.01 & 0.00 & 0.00 & 0.00 \\ 0.04 & 1.01 & 1.86 & 0.02 & 1.03 & 0.00 \\ 0.00 & 0.00 & 0.02 & 2.14 & 1.15 & 1.07 \\ 0.00 & 0.00 & 1.03 & 1.15 & 4.43 & 0.03 \\ 0.00 & 0.00 & 0.00 & 1.07 & 0.03 & 1.05 \end{pmatrix}. \tag{7}$$

この行列の各要素は各文書間の共引用解析の値とほぼ等しい値を持つことがわかる.

次に提案手法を重要度に偏らせて ($\gamma = 0.99$) 実験を行った. 結果は以下のようになる.

$$R_{0.99} = \begin{pmatrix} 115.04 & 235.81 & 43.56 & 0.00 & 0.00 & 0.00 \\ 235.81 & 489.90 & 91.63 & 0.00 & 0.00 & 0.00 \\ 43.56 & 91.63 & 40.82 & 37.59 & 90.96 & 10.18 \\ 0.00 & 0.00 & 37.59 & 69.38 & 157.23 & 20.07 \\ 0.00 & 0.00 & 90.96 & 157.23 & 375.79 & 42.60 \\ 0.00 & 0.00 & 10.18 & 20.07 & 42.60 & 6.70 \end{pmatrix}. \tag{8}$$

この結果は, 拡散係数の小さい場合 (式 (7)) と近いように見える. しかし, n_6 に注目すると両者の違いがわかる. 拡散係数の小さい場合, n_6 に対して n_4 が直接の共引用関係にあるため最も大きな値を持つが, 式 (8) ではコミュニティ 2 で最も被引用数が多い n_5 が最大値をとる. このように, 提案手法では重要度に偏らせることで同一コミュニティ内での重要度に偏ってゆくことがわかる. ただし, このランキングは隠れトピック 2 における pHITS の重要度ランキングとは異なる. pHITS における隠れトピック 2 の重要度ランキングは $n_4 > n_5 > n_6 > n_3$ である. しかし n_4 よりも, n_5 の方が被引用数が多いため, 提案手法のように n_5 がコミュニティ 2 で最も重要であると考える方が妥当である.

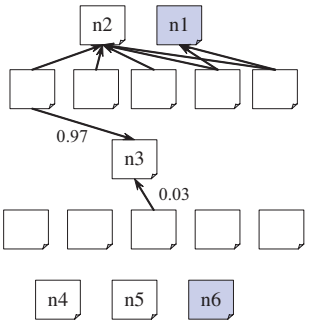


図 4　隠れトピックグラフ 1.

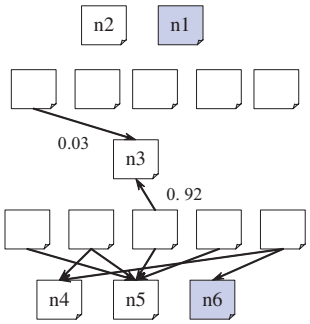


図 5　隠れトピックグラフ 2.

また, 式 (8) で二つのコミュニティの中間に存在する n_3 に注目すると, 最も大きな値をもつ文書はコミュニティ 1 に存在する n_2 だが, 次に大きな値をもつ文書はコミュニティ 2 に存在する n_5 であり, 二つのコミュニティのランキングを混合したものとなっている. これは我々の直感と一致する結果である.

以下, 提案手法がなぜ重要度に偏らせても各文書の属すコミュニティを考慮できるのかについて, 引用グラフ (図 2) 及びこの引用グラフから節 4.2 の手法で導出される, 二つの隠れトピックグラフ (図 4,5) を基に考察する.

隠れトピック 1 に対応する隠れトピックグラフ 1 (図 4) において, n_6 が含まれるコミュニティ間の孤が消滅していることが分かる. これは弧の重みが 0 となったことを表している. また, 重みが表記されていない孤の重みは引用グラフの場合から変更が無く 1.0 である. この隠れトピックグラフから共引用グラフを生成しノイマンカーネルを適用する. このグラフではコミュニティ 2 には孤が存在しなくなっているため, 重要度に偏らせるとコミュニティ 1 の重要度に偏る. 一方で, ノイマンカーネルは共引用グラフ上の各節点間の全ての経路の数の重みつき和を求めているにもかかわらず n_4, n_5, n_6 は共引用グラフで弧を持たないため, 他の節点に対する内積はゼロとなる.

次に隠れトピック 2 に対応する隠れトピックグラフ 2 (図 5)

表 1　根論文 ‘Empirical studies in discourse’ に対するノイマンカーネル ($\gamma = 0.001$) の出力.

$\hat{K}$	C	H	論文題目
1	1	771	Empirical studies in discourse
2	2	1	Building a large annotated corpus of English: the Penn Treebank
3	2	50	Attention, intentions, and the structure of discourse
4	2	76	Assessing agreement on classification tasks: the Kappa statistic
5	2	201	The reliability of a dialogue structure coding scheme
6	2	604	Message Understanding Conference (MUC) tests of discourse processing
7	2	1061	Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue
8	–	3	Statistical decision-tree models for parsing
9	–	4	A new statistical parser based on bigram lexical dependencies
10	–	96	Centering: a framework for modeling the local coherence of discourse

を考える. 今度はコミュニティ1 内に存在する孤が消滅していることが分かる. この隠れトピックグラフから共引用グラフを生成しノイマンカーネルを適用する. 今度はコミュニティ1 内に存在する孤が消滅しているためこのノイマンカーネルは重要度に偏らせるとコミュニティ2 における重要度に偏る. このとき, n_1, n_2 は共引用グラフで弧を持たないため, 他の節点に対する内積はゼロとなる.

提案手法 (式 (6)) は隠れトピック 1,2 に基づくノイマンカーネルを足し合わせるが, 隠れトピックグラフ 1 だけで共引用が存在する節点 n_1, n_2 はコミュニティ1 の重要度のみに偏り, 隠れトピックグラフ 2 だけで共引用が存在する節点 n_4, n_5, n_6 はコミュニティ2 の重要度に偏る. そしてコミュニティの中間に存在する節点 n_3 はどちらのコミュニティに存在する節点とも共引用関係にある. それゆえコミュニティ1,2 に対して適用されたノイマンカーネルそれぞれで重要度に偏るため, 二つのコミュニティにおける重要度を混合した重要度に偏っている.

5. 実　　験

我々は, 提案手法の性能を実際の引用関係データを用いて検証した. 実験ではノイマンカーネルを実際の論文間の引用グラフに適用した. 引用グラフは, OCR 処理した参考文献一覧から, [9] の手法で抽出したものであり, 自然言語処理学分野の学術誌論文, 国際学会論文 2867 件からなる. 本実験を通して, 隠れコミュニティは 5 に固定した.

まず, 上記引用グラフから共引用グラフを作り, 様々なパラメタ設定のもとで, ノイマンカーネル行列を (i) 隠れトピックを考慮しない場合, (ii) する場合の両方を計算した. 次に実験で着目する論文 (以下根論文と呼ぶ) を選び, 計算されているカーネル行列中の根論文に対応する行の要素の中から, 大きさ順に 上位 10 件を取り出す. そして各々を与える列番号から, 該当する論文を抽出する. このとき上位 10 件には根論文自身も含まれる可能性がある. この処理を各カーネル行列に対して行い, 個々の 10 件の論文リスト (以下, ‘カーネルの出力’ と呼ぶ) と, 共引用解析や, HITS (権威度) 順位との相関について調べる.

5.1　ノイマンカーネル (隠れトピックを考慮しない場合)

この節では 隠れトピックを考慮しない場合におけるノイマンカーネルの性能を評価する. ノイマンカーネルの拡散係数 γ の値は 2 種類の設定 (0.001, 0.9999) で行った.

表 2　根論文 ‘Empirical studies in discourse’ に対するノイマンカーネル ($\gamma = 0.9999$) の出力.

$\hat{K}$	C	H	論文題目
1	2	1	Building a large annotated corpus of English: the Penn Treebank
2	–	2	A stochastic parts program and noun phrase parser for unrestricted text
3	–	3	Statistical decision-tree models for parsing
4	–	4	A new statistical parser based on bigram lexical dependencies
5	–	5	Unsupervised word sense disambiguation rivaling supervised methods
6	–	6	Word-sense disambiguation using statistical models of Roget’s categories trained
7	–	7	The mathematics of statistical machine translation: parameter estimation
8	–	8	Three generative, lexicalised models for statistical parsing
9	–	9	Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging
10	–	10	Integrating multiple knowledge sources to disambiguate word sense: an exemplar-based approach

論文 ‘Empirical studies in discourse’ を根論文として選択した. この根論文に対する ノイマンカーネルの出力結果を, 表 1 ($\gamma = 0.001$), 表 2 ($\gamma = 0.9999$) に掲げる.

表中, 最左列 ($\hat{K}$) が ノイマンカーネルによる順位であり, C と題された列は根論文における共引用解析による順位, H 列は HITS における大域的な重要度 (権威度) の順位, をそれぞれ表す. 最右列は各論文の題目である. C 列の ‘–’ は, 該当論文と根論文が 1 度も共引用されていないため, 順位付けができなかったことを表す.

γ を 0.001 に設定した場合 (表 1), ノイマンカーネルの出力は HITS の順位付けとかなり違ったものであることがわかる. この表の最上位は, 根論文自身であり, 以下 7 位までは全て共引用関係にある論文である (C 列を参照). このことは, ノイマンカーネルの順位付けは, γ を減少させると, より関連度に近づくことを裏付けている. 実際, 根論文と共引用されている論文は, これら (根論文を除く) 6 件以外にはない.

また, 共引用解析の結果 (C 列) によると, これら 6 件の論文は同順位 (全て 2 位) とされている. それに対して, ノイマンカーネルでは, これら 6 件に対して HITS 順位 (H 列) を反映した順位付けが行われており, このカーネルが HITS 重要度と共引用関連度の混合である, という特徴がうかがえる.

しかし, 表 2 で示されているように, 拡散係数を大きく設定すると ($\gamma = 0.9999$ の時), ノイマンカーネルの順位付け は HITS の順位付けと一致してしまう. 例えば, 5 位にランクされた論文, ‘Unsupervised word sense disambiguation rivaling supervised methods’ のように上位にランクされた論文の大半が ‘discourse’ というトピックと関連しないことがわかる.

5.2　隠れトピックを考慮したノイマンカーネル

この節では 隠れトピックを考慮したノイマンカーネル (式 (6)) の性能を評価する. 本実験では, 隠れコミュニティ数を 5 に固定して実験を行った. また, 拡散係数 γ の値は 2 種類の設定 (0.001, 0.9999) で行った.

表 3　根論文 ‘Empirical studies in discourse’ に対する ノイマンカーネル ($\gamma = 0.001$) の出力 (コミュニティを考慮した場合)

R	C	H	論文題目
1	1	771	Empirical studies in discourse
2	2	201	The reliability of a dialogue structure coding scheme
3	2	76	Assessing agreement on classification tasks: the kappa statistic
4	2	1061	Effects of variable initiative on linguistic behavior in Human-Computer spoken natural language dialogue
3	2	50	Attention, intentions, and the structure of discourse
6	2	1	Building a large annotated corpus of English: the Penn Treebank
7	2	604	Message Understanding Conference (MUC) tests of discourse processing
8	–	96	Centering: a framework for modeling the local coherence of discourse
9	–	374	A trainable document summarizer
10	–	60	Evaluating a focus-based approach to anaphora resolution

表 4　根論文 ‘Empirical studies in discourse’ に対する ノイマンカーネル ($\gamma = 0.9999$) の出力 (コミュニティを考慮した場合).

R	C	H	論文題目
1	2	50	Attention, intentions, and the structure of discourse
2	–	61	Multi-Paragraph segmentation of expository text
3	–	77	Lexical cohesion computed by thesaural relations as an indicator of the structure of text
4	–	111	Combining multiple knowledge sources for discourse segmentation
5	–	194	A prosodic analysis of discourse segments in direction-giving monologues
6	2	76	Assessing agreement on classification tasks: the Kappa statistic
7	–	150	An automatic method of finding topic boundaries
8	–	162	Text segmentation based on similarity between words
9	–	317	Intention-Based segmentation: human reliability and correlation with linguistic cues
10	–	340	Replicability of transaction and action coding in the map task corpus

先程の実験と同じように, 根論文 ‘Empirical studies in discourse’ に対する提案手法 (式 (6)) の出力結果を分析した. 結果を, 表 3 ($\gamma = 0.001$), 表 4 ($\gamma = 0.9999$) に掲げる. 表中, 最左列 (R) が 提案手法 (式 (6)) による順位である. それ以外の列については前節の実験と同一の事象を表している.

表 3 において, 拡散係数を小さく設定 ($\gamma = 0.001$) した時の結果が示されている. 表 1 のように, 上位にランクされた論文の多くが題目に ‘discourse’ や ‘dialogue’ 等の単語を含んでおり, 根論文に関係するものであることがわかる.

また, 表 4 で示されているように, $\gamma = 0.9999$ と拡散係数を大きく設定し, 重要度に偏らせた時でも, 上位にランクされる多くの論文のタイトルに ‘discourse’ が含まれている. このことから提案手法は拡散係数を大きく設定し重要度に偏らせると, 文書が属しているコミュニティに応じた重要度に偏ってゆくことが分かる. 例えば 5 位にランクされた論文 ‘A prosodic analysis of discourse segments in direction-giving monologues’ は表 1, 及び表 3 では根論文と共引用関係に無いため上位にランク付けされなかったが, この論文は題目に ‘discourse’ を含んでおり根論文のトピックから大きく外れていないことがわかる.

6. 議　　論

本節では, 提案手法と関連する研究について議論する.

6.1 フィッシャーカーネル

フィッシャーカーネル [11] は生成モデルから導出されるカーネルである. 生成モデルが与えられたとき, フィッシャーカーネルを導出するには生成モデルのフィッシャースコア $u(d,\theta)$ すなわち生成モデルの対数尤度関数のパラメタ θ に関する勾配, 及び, フィッシャー情報行列 I を求める必要がある (注1). 生成モデルが持つ各パラメタ θ で対数尤度関数を偏微分した値からなるベクトル $u(d,\hat{\theta})$ が与えられた時, フィッシャーカーネルは以下で定義される.

$$K(d,d') = u(d;\hat{\theta})^{\mathrm{T}} I(\hat{\theta})^{-1} u(d';\hat{\theta})$$

Hofmann [7] は文書の類似度を計算する上で問題となる多義語を処理するため, pLSI に基づいたフィッシャーカーネルを提案した.

Hofmann は pLSI が持つ 2 種類のパラメタ $p(w_j|t)$, $p(t)$ (注2) それぞれで pLSI の対数尤度関数を偏微分して, 2 種類のフィッシャーカーネルを定義し, それら二つのカーネルの和を 2 文書間の類似度を算出するカーネルとして使用した. ここで, $p(w_j|t)$ はトピック t のもとでの単語 w_j の生起確率 $p(t)$ はトピック t の生起確率を表す. このうちパラメタ $p(w_j|t)$ に基づくカーネル $\bar{K}(d,d')$ は以下で与えられる.

$$\bar{K}(d,d') = \sum_{j=1}^{M} \hat{p}(w_j|d)\hat{p}(w_j|d') \sum_{t=1}^{N} \frac{p(t|d,w_j)p(t|d',w_j)}{p(w_j|t)} \quad (9)$$

ここで, $\hat{p}(w_j|d)$ は実際に文書 d に単語 w_j が存在する頻度を d 中に存在する単語数で割った数である.

このカーネルは, $\hat{p}(w_j|d)\hat{p}(w_j|d')$ の項を除くと, 我々の定義した隠れトピックグラフに類似する重みつきグラフ (注3) において長さ 1 の経路の重みを計算していることに対応する. したがって, このカーネルを引用解析に適用するとグラフ上で長さ 2 の経路以上の経路を介してつながっている, つまり, 二つの文書が共通に同一の文書から引用されない場合, 内積を算出できないという (共引用解析と同様の) 問題が存在する.

これに対して提案手法は, 隠れトピックグラフにノイマンカーネルを適用することで全ての長さの経路の重みつき和を計算する. この方法により, 同一の文書から引用された論文同士でなくとも内積の値を付与できる.

Hofmann は式 (9) とは別に pLSI に存在するパラメタ $p(t)$ に基づくカーネルを同時に使用することで上記の問題に対処している. しかしこの場合には, 我々がノイマンカーネルを用いて行った関連度と重要度の混合という議論は行えない.

(注1)：フィッシャー情報行列は計算コストが大きいため一般に単位行列で代用される.

(注2)：正確にはパラメータ $p(w_j|t)$, $p(t)$ そのものではなく, $2\sqrt{p(w_j|t)}$, $2\sqrt{p(t)}$ で偏微分する.

(注3)：ただし, フィッシャーカーネルでは提案手法で使用する隠れトピックグラフに存在する各弧に $\sqrt{1/p(w_j|t)}$ という重みが付与される.

6.2 他の引用生成モデルを用いた拡張

本稿では, Cohn らの引用の生成モデルを用いて隠れトピックグラフを生成したが, この引用生成モデルは Hofmann の pLSI という文書生成モデルに基づくことはすでに述べた. しかし, pLSI 以外にも, Blei らの LDA (Latent Dirichlet Allocation) [2] や山本らの DM (Dirichlet Mixture) [17] 等の隠れトピックを仮定した文書生成モデルが多数提案されており, これらも引用生成モデルとして利用することが可能である. 今後, これらのモデルを基にした場合に提案手法の振舞いに違いがあるか調べる予定である.

7. むすび

Hofmann の pLSI (probabilistic Latent Semantic Indexing) の結果を用いて, コミュニティ (トピック) ごとに重みつき引用グラフ (隠れトピックグラフ) を生成した. そして各隠れトピックグラフにノイマンカーネルを適用することによって, 対象文書が属するコミュニティを考慮して重要文書を推薦, 提示できることを示した. 提案したカーネルの有効性を引用関係データを用いて検証した.

文　献

[1] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked enviornment. In *Proc. 21st ACM SIGIR Conference*, 1998.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. In *Neural Information Processing Systems 14*, 2001.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual (web) search engine. *Computer Network and ISDN Systems*, 30(1–7):107–117, 1998.

[4] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proc. 18th International Conference of Machine Learning*, 2001.

[5] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, Department of Computer Science, University of California at Santa Cruz, Santa Cruz, CA 95064, USA, 1999.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *Proc. 22th ACM SIGIR Conference*, pp. 50–57, 1999.

[7] T. Hofmann. Learning the semilarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in Neural Information Processing Systems 12*, pp. 914–920. MIT Press, 2000.

[8] 伊藤, 新保, 工藤, 松本. カーネル法による計量書誌尺度の統一的解釈. 人工知能学会論文誌, 第 19 巻, 2004.

[9] 伊藤, 堀部, 新保, 松本. 複数尺度を用いた参考文献の同定. 情報処理学会研究会報告 2003-DBS-130, pp. 181–188, 5 月 2003.

[10] T. Ito, M. Shimbo, T. Kudo, and Y. Matsumoto. Application of kernels to link analysis. In *Proc. 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2005.

[11] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, 1998.

[12] J. Kandola, J. Shawe-Taylor, and N. Cristianini. Learning semantic similarity. In *Advances in Neural Information Processing Systems 15*. MIT Press, 2002.

[13] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963.

[14] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, pp. 604–632, 1999.

[15] S. Lawrence, K. Bollacker, and C. L. Giles. Autonomous citation matching. In *Proc. 3rd International Conference on Autonomous Agents*, New York, 1999. ACM Press.

[16] H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. American Society for Information Science*, 24:265–269, 1973.

[17] 山本, 貞光, 三品. 混合ディレクレ分布を用いた文脈のモデル化と言語モデルへの応用. 情報処理学会研究会報告 2003-SLP-48-5, 2003.