社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＩＣＳ－135　(21)
2004／3／16

カーネル法による計量書誌尺度の統一的解釈

伊藤　敬彦†　　新保　　仁†　　工藤　　拓†　　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
E-mail: †{takahi-i,shimbo,taku-ku,matsu}@is.aist-nara.ac.jp

あらまし　グラフ上で定義されたカーネル法を引用解析に適用し，過去独立に議論されてきた書誌計量尺度 (関連度, 重要度, 相対的重要度) に対し統一的解釈を与える．具体的には，相対的重要度は関連度と重要度を二つの端点とする線分上の任意の中点として定義され，重要度もしくは関連度に対する偏りはパラメータによって制御される．

キーワード　カーネル法，リンク解析，引用解析，共引用解析，計量書誌学

A Kernel-based Account of Bibliometric Measures

Takahiko ITO†, Masashi SHIMBO†, Taku KUDO†, and Yuji MATSUMOTO†

† Graduate School of Information Science, Nara Institute of Science and Technology
E-mail: †{takahi-i,shimbo,taku-ku,matsu}@is.aist-nara.ac.jp

Abstract　We explore the application of kernel methods to citation analysis. We show that a family of kernels on graphs provides a unified perspective on the three bibliometric measures that have been discussed independently: relatedness between documents, overall importance of individual documents, and importance of documents relative to one or more (root) documents (relative importance). The framework provided by the kernels establishes relative importance as an intermediate between relatedness and overall importance, in which the degree of 'relativity,' or the bias between relatedness and importance, is naturally controlled by a parameter characterizing the kernel family.

Key words　Kernel Methods, Link Analysis, Citation Analysis, Bibliometrics

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 (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．PageRank [2] や HITS [10] 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，計量書誌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．

'相対的重要度' は，検索クエリの内容に応じて柔軟に Web ページの順位付けを行うことを念頭に提案された，新しい計量書誌尺度である．しかし，この尺度に関する従来の議論 [6], [14] は計算方法の提案が主であり，この尺度自体の位置づけには曖昧な部分があった．ことに，共引用や書誌結合といった，従来からある，文書間の '関連度' との関係については一切触れられておらず，実際，[14] の議論に基づくと，これら関連度も相対的重要度の基準を満たすことになる．

本報告では，3 つの計量書誌尺度 (相対的重要度，重要度，関連度) に統一的な解釈を与える．この解釈は，グラフ上で定義されたパラメトリックなカーネル関数族 (Neumann カーネル [8]) によって提供される．パラメータのとり得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 authority 度に類似の) 文書の重要度に対応する．結果として，相対的重要度は二つの端点 (重要度と関連度) 間に存在する任意の中点と考えられ，パラメー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可能である．

本報告のもう一つの目的は，カーネル法を用いて，古典的な関連度算出法の欠点を克服することである．これらの古典的手法では二つの論文 (Web ページ) が同一の論文をする (書誌結合)，もしくは同一の論文から引用される (共引用解析) 関係にない場合，それらに関連度を与えることができない．この欠点はグラフ上の Laplace 作用素をカーネルに導入することで解決される．

2. 準　　備

準備として，本報告を通して議論する，3 種類の計量書誌尺度について復習する．これらは，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．重み付きグラフは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，$w : E \to \mathbb{R}\backslash\{0\}$ は，弧のラベル (重み; 非0の実数) を定める重み付け関数である．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．重み付きグラフ $G = (V,E,w)$ 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，$A_{ij} = w(i,j)$，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．重み付きグラフ 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．

2.1 関 連 度

書誌結合 [9] と共引用解析 [9] 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，科学文献検索システム CiteSeer [11] は，個々の引用に対するヒューリスティクスな重み付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 の隣接行列とする時，対称行列 $A^{\mathrm{T}}A$ を G の**共引用行列**と呼び，この行列によって導出される重み付き無向グラフを，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．

このとき $A^{\mathrm{T}}A$ 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 (図1)．

2.2 重 要 度

各文書の重要度を算出するための基準として，各文書の被引用数が古くから用いられてきた．Kleinberg の HITS [10] 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，各文書を二つのスコア (authority 度 と hub 度) で評価する．直観的には，高い authority 度を持つ文書とは hub 度の高い文書から多く引用される文書であり，逆に hub 度の高い文書とは authority 度の高い文書を多く引用する文書である．

定義2 (HITS) G を引用グラフとする．HITS アルゴリズムは以下の再帰式を計算する ($n = 0, 1, \ldots$)．

$$a^{(n+1)} = A(G)^{\mathrm{T}} h^{(n)}, \quad h^{(n+1)} = A(G) a^{(n+1)}, \tag{1}$$

ただしここでベクトル $a^{(n)}, h^{(n)}$ は大きさが1になるように毎回正規化される．文書 i の **authority 度**は **authority ベクトル** $\lim_{n\to\infty} a^{(n)}$ の i-要素で与えられ，**hub 度**は **hub ベクトル** $\lim_{n\to\infty} h^{(n)}$ の i-要素で与えられる．

引用グラフ G に対して，HITS の authority ベクトルと hub ベクトルはそれぞれ，共引用行列 $A(G)^{\mathrm{T}}A(G)$ と書誌結合行列 $A(G)A(G)^{\mathrm{T}}$ の最大固有ベクトルと一致することが知られている．

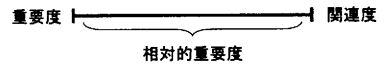


図2 相対的重要度: 重要度と関連度の混合

2.3 相対的重要度

HITS と PageRank は共にサーチエンジンから返される Web ページのランキングに使用される．しかし，この二つは (大域的な) 重要度算出アルゴリズムであるために共通の，topic drift と呼ばれる問題 [1] を持つ．これは，大域的な重要度アルゴリズムに固有の問題で，個々のページがクエリと関係するかという情報を考慮せずランキングを決定してしまうことから生ずる問題である．

この問題を解決するため，ユーザの発するクエリに合わせたランキング (クエリ依存ランキング) を返す手法が，ページの内容情報を用いて，多く開発されてきた [1], [3], [4], [12]．

最近，White と Smyth [14] は引用グラフ上の新しい尺度 '相対的重要度' を提案した．この尺度は '根節点 (の集合) に対する各節点の重要度' と定義され，明らかに，クエリ依存ランキングへの応用を念頭に考案されたものである．もっとも他の手法と違い，この手法は文書の内容情報を使用せずとも，根節点集合 (クエリに相当) さえ与えられれば，引用関係のみからクエリ依存ランキングを計算可能，という特徴がある．Haveliwala [6] は一般的な重要度のランキングをトピック (根節点) に偏らせたが，この手法も相対的重要度算出アルゴリズムの一種と見なせる．

ただし，これら先行研究ではもっぱら計算手法に関する議論が中心であり，相対的重要度の，大域的重要度や関連度といった既存の計量書誌学尺度に対する位置づけが曖昧なまま残されている，という問題がある．

3. 関連度と重要度の混合としての相対的重要度

相対的重要度の位置づけを明確にするため，我々は，この尺度を '重要度と関連度の混合' として捉え (図2)，三つの計量書誌尺度に統一的な解釈を与えることを試みる．本節の残りと4節では，グラフ上に定義されたカーネルがそのような統一的な尺度を提供することを示す．

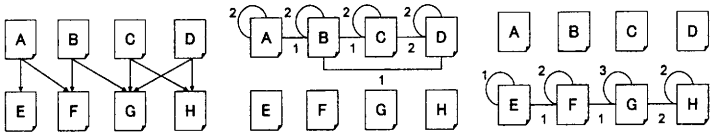
3.1 Neumann カーネルに基づく相対的重要度

Kandola ら [8] 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 **Neumann カーネル** を提案した．我々は，このカーネルを引用グラフに対して適用し，相対的重要度を計算する．文書内の単語情報は一切用いない．

本来，Neumann カーネルは，与えられた文書-単語行列 X (X_{ij} は単語 j の文書 i 内での出現数) から，まず文書共起行列 $K = X^{\mathrm{T}}X$ と単語共起行列 $M = XX^{\mathrm{T}}$ を生成する．K の (i,j)-要素は文書 i, j 間の類似度を，M は単語間の類似度を与えることになる．Neumann カーネルは K, M を用いて以下のように定義される．

定義3 X を文書-単語行列とし，$K = X^{\mathrm{T}}X, M = XX^{\mathrm{T}}$ とすると，減衰係数 λ の Neumann カーネル行列，$\hat{K}_\lambda$ および $\hat{M}_\lambda$ は以下の方程式の解として与えられる．

$$\hat{K}_\lambda = \lambda X^{\mathrm{T}} \hat{M} X + K, \quad \hat{M}_\lambda = \lambda X^{\mathrm{T}} \hat{K} X + M \tag{2}$$



(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

もし $\lambda < \min(\|K\|^{-1}, \|M\|^{-1})$ であれば $\hat{K}_\lambda$ および $\hat{M}_\lambda$ は存在し,

$$\hat{K}_\lambda = K(I-\lambda K)^{-1}, \quad \hat{M}_\lambda = M(I-\lambda M)^{-1} \tag{3}$$

となる [8]. このとき, 文書 i, j 間の類似度は $\hat{K}_\lambda$ の (i,j)-要素で与えられ, 同様に, M の各要素は単語間の類似度を表す.

式 (3) は, $\hat{K}_\lambda$, $\hat{M}_\lambda$ がいずれも Neumann (幾何) 級数を用いて表せることを示唆している. すなわち,

$$\hat{K}_\lambda = K\sum_{n=0}^{\infty}(\lambda K)^n, \quad \hat{M}_\lambda = M\sum_{n=0}^{\infty}(\lambda M)^n. \tag{4}$$

さて, 式 (2) の二重再帰式は HITS (式 (1)) の authority 度と hub 度と類似の相補的な関係である. 我々はこの類似性に着目し, Neumann カーネルを引用解析に応用する.

具体的には, 文書単語行列 X の代わりに, 引用グラフの隣接行列 A を使用して二つの行列 $K = A^{\mathrm{T}}A$ と $M = AA^{\mathrm{T}}$ を生成し (これらの行列は共引用行列, 書誌結合行列とそれぞれ等しい), 式 (3) に代入して Neumann カーネル行列 $\hat{K}_\lambda$, $\hat{M}_\lambda$ を得る. 式 (4) に照らしてみると, $\hat{K}_\lambda$ はパラメータ λ が十分大きい場合には $\lim_{n\to\infty}(A^{\mathrm{T}}A)^n$, すなわち HITS の authority 度により大きな比重をおくことに対応し, 逆に λ が 0 に近い場合には, 書誌結合行列 $K = A^{\mathrm{T}}A$ に近づくことがわかる. ゆえに, $\hat{K}_\lambda$ は, 共引用と HITS の authority 度のある種の混合と見なせ, これは, 3 節冒頭で述べた観点に立つと, 相対重要度そのものに他ならない. ここで, 根節点 i から見た文書 j の相対的重要度は $\hat{K}_\lambda$ の (i,j)-要素で与えられる. 同様に, $M = AA^{\mathrm{T}}$ を用いて得られる $\hat{M}_\lambda$ の各要素は, 書誌結合と HITS の hub 度の混合となる相対的重要度を表す.

3.2 解　　釈

文書 (単語) 共起行列の代わりに共引用行列と書誌結合行列を使用したが, 実際に Neumann カーネルは何を計算しているのであろうか? 以下, Neumann カーネルによって導出される素性空間と, この空間において, カーネルが何を計算しているかについて概略を述べる.

(有限) グラフ G 上の Neumann カーネルは G 中の与えられた 2 節点間の全ての経路コスト (経路を構成する弧の重みの積(注1)) を, 重みをつけて足し合わせていると考えることができる. このカーネルは正定値 (positive semidefinite) [13] であり, したがって, ある (素性) 空間において内積を計算していることになる. Neumann カーネルによって導出される素性空間 $\mathcal{F}$ の基底 (素性) は, おのおのが G 上の 1 本の経路に対応する. i-番目の基底に対応する経路を π_i とすると, グラフの節点は, 経路 π_i の終端となる場合のみ i-番目の要素が非ゼロとなるような, $\mathcal{F}$ 上のベクトルとして表現される. ベクトルの各成分の大きさは, 個々の基底に対応する経路の長さ (と減衰係数 λ) に応じて減衰させられる. 具体的には, ある節点が, 長さ n, コスト c の経路 π_i の端点ならば, この節点に対応する素性ベクトルの第 i 要素は, $\lambda^{(n-1)/2}c^{1/2}$ となる (λ は Neumann カーネルの減衰係数パラメータ). Neumann カーネルはこのように二つの節点間の全ての経路の重み付き和を, 素性空間 $\mathcal{F}$ における内積として, その節点対に割り当てる. 一般に, G 上の 2 節点間には無限の数の経路が存在するので, $\mathcal{F}$ もまた無限の次元数を持つことになるが, 式 (3) からわかるように, Neumann カーネルはグラフの大きさの多項式時間で計算可能である.

4. Laplace 作用素のカーネルへの適用

4.1 古典的関連度算出手法 (共引用解析/書誌結合) の欠点

前節で見てきたように, Neumann カーネルは共引用解析/書誌結合関連度と HITS 重要度との関係を明らかにすると同時に, それらの中間としての相対的重要度の定式化を与えた. この節ではカーネル法を用いた, 別の重要度, 関連度, 相対的重要度の定式化を提案する. この定式化において古典的な関連度算出手法 (共引用解析/書誌結合) の欠点を克服する. 具体的には, 共引用解析/書誌結合は以下の二つの欠点を持つ.

問題 1　共引用解析は二つの文書が共通に同一の文書から引用されない場合, 関連度を算出できない. 同様に, 書誌結合は 2 文書が共通に同一の文書を引用しない場合関連度を算出できない.

図 1(a) の引用グラフにおいて, 文書 E と G を同時に引用する文書は存在しない(注2). そのため, 共引用解析に基づくとこれらの文書には全く関連はない. しかし, 文書 A, B, F を通じて, 微弱な関連性が E, G 間に存在すると考える方が自然である.

問題 2　共引用解析は 2 文書間の関連度を共通に引用された数のみにもとづいて決定し, 他の文書と引用している文書との関係を無視する. 例えば, 引用している文書が他の文書から受ける引用の数を考慮しない.

図 3 に示す Web ページ間のリンク関係を考える. ページ D と E は Google や Yahoo のような頻繁にリンクされるページを表す. 一方でページ F と G は B と C 以外どこからも引用されていない. 一般に, Google や Yahoo を共通に引用したからといってそれらのページが (内容的に) 似ているとは言えず, したがってこのような状況では, ページ B とより関連性の高いのは A ではなく C であると考えるのが自然である. しかし, ページ対 (A, B) と (B, C) は, ページを 2 個共通に引用しているという点では同

(注1) : 重みが多重度を表す整数の場合には, 経路の本数に相当する.

(注2) : したがって, 図 1(c) において E と G を結ぶ弧は存在しない.

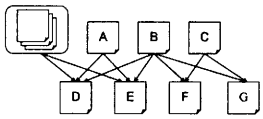


図3　例：書誌結合の問題点

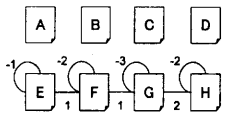


図4　負の Laplace 作用素 $-L(G)$ のグラフ表現. ここで G は図 1(c) の共引用行列.

じであるので, 書誌結合において, これら 2 対の関連度は等しくなってしまう.

4.2　グラフの Laplace 作用素を用いた拡張

以下で, **グラフの Laplace 作用素 (graph Laplacian)** を使って 4.1 節で提起した問題を解決する.

定義 4　(G,w) を重み付き無向グラフとする. このとき, **G の Laplace 作用素** ($L(G)$ と表記される) は以下で与えられる.

$$L(G) = D(G) - A(G) \tag{5}$$

ここで $D(G)$ は $D(G)_{ii} = \sum_k w(i,k)$ なる対角行列である.

式 (3) において隣接行列 $A(G)$ の代わりに, $-L(G)$ (以下, 負の Laplace 作用素と呼ぶ) を使用することで Geometric Laplacian カーネルを得る.

定義 5　G を重みつき無向グラフとする. λ を減衰係数とする **Geometric Laplacian カーネル行列** $\hat{L}_\lambda(G)$ は以下で与えられる.

$$\hat{L}_\lambda(G) = \sum_{n=0}^{\infty} (-\lambda L(G))^n.$$

式 (5) が示すように, 負の Laplace 作用素 $-L(G)$ はグラフ G の自己ループの重みを変更したものと考えることができる. 例えば, 図 4 は図 1(c) の共引用グラフから導かれる負の Laplace 作用素をグラフ表現したものである. 自己ループの弧の重みを除外すれば, グラフ自体はもとの共引用/書誌結合グラフと同一であるので, Geometric Laplacian カーネルの素性空間もまた, 3.2 節で述べた Neumann カーネルの素性空間と同一となる. すなわち, 空間の各基底 (素性) はグラフの経路に相当する. ただし, 各素性 (すなわち経路) には違った重みが与えられている. というのも, Neumann カーネルにおける長さ n の経路の重みは λ^{n-1} で与えられるが, Geometric Laplacian カーネルでは長さ n の経路の重みは λ^n で与えられる. さらに, Neumann カーネルは長さ 0 の経路に重みを与えないが Geometric Laplacian カーネルは長さ 0 の経路も数える.

Geometric Laplacian カーネルはこの節の始めで述べた共引用解析/書誌結合の欠点を克服したものになっている. 実際, 二つの文書が共に同一の文書を引用する, もしくは共引用されることがなくとも (問題 1). 共引用解析/書誌結合グラフ中の二つの節点間に経路が存在する限り, Geometric Laplacian カーネルはそれらの関連度を計算可能である. なぜなら Geometric Laplacian カーネルは共引用解析/書誌結合グラフ内の全ての経路をその素性空間において数えているからである. 例えば, 図 1(a) において, 共引用解析では文書 E と文書 G の間に関連度を定義できないが, 共引用グラフにおける Geometric Laplacian カーネルは関連度 $(2\lambda^3 - 16\lambda^4 + 94\lambda^5 - \ldots)$ を与える.

自己ループに割り当てられた負のコストによって, Geometric Laplacian カーネルは, 問題 2 も克服している. 文書が他の多くの文書と共に引用する (される) ほど, より強い負の重みがその節点の自己ループに付与される. Geometric Laplacian カーネルは 2 節点間の全ての経路の重み付き和を計算するため, (被) 引用数の大きな節点を含む経路がある場合, 同じ経路の途中に, 大きな負の重みを持つ自己ループを加えた別の経路についても考慮されることを意味する. このため, (被) 引用数の大きい節点を含む経路の重みは大きく減少させられ, そのような経路のみで結ばれた節点間の関連度は, 結果的に低く見積もられる. 例えば, G として 図 1(c) の共引用グラフを用いると, Geometric Laplacian カーネルは文書 F と E の関連度に, F と G 間の関連度よりも大きな値を与える.

4.3　Geometric Laplacian カーネルにおける偏りの調節

Neumann カーネルにおいてパラメータ λ は '関連度と重要度の偏り' を調整したが, Geometric Laplacian カーネルにおける λ を, 同じ目的のために使用することはできない. なぜなら, グラフの負の Laplace 作用素を隣接行列の代わりに使用しているため, 行列の対角要素に負の値が与えられるが, これによって, 前節で述べた通り (被) 引用数による重要度への偏り効果が打ち消されてしまうからである. 後ほど, 5.3 節の実験によって, λ が変化しても尺度が重要度に偏らないことを実証する.

λ が関連度, 重要度間の偏り調整に使えないとはいえ, 新しいパラメータを導入することで, Geometric Laplacian カーネルに基づく (相対的) 重要度も定義できる.

定義 6　G を重みつき無向グラフとする. **変更版 Geometric Laplacian カーネル行列** $\hat{L}_{\lambda,\alpha}(G)$ は以下の式で与えられる.

$$\hat{L}_{\lambda,\alpha}(G) = \sum_{n=0}^{\infty} (-\lambda(\alpha D(G) - A(G)))^n \quad (0 \leq \alpha \leq 1)$$

ここで α, λ はパラメータ, $D(G)$ は定義 4 と同じ対角行列である.

変更版 Geometric Laplacian カーネルを引用解析に用いる際には, G として共引用 (あるいは書誌結合) グラフを用いる. このとき, λ を十分に大きく設定し, 重要度 (関連度) への偏りは α の値を変えて調整する. $\alpha = 1$ のとき, $\hat{L}_{\lambda,\alpha}(G)$ は通常の Geometric Laplacian カーネル $\hat{L}_\lambda(G)$ と等しくなるが, α が減少すると, カーネル行列の各行 (列) ベクトル中の要素の大きさによる順位は, 重要度に近くなる. 特に, $\alpha = 0$ の場合には, $\hat{L}_{\lambda,\alpha}(G)$ は減衰係数 λ の Neumann カーネルとほぼ一致し, HITS に偏った相対的重要度を算出する.

5.　実　　験

我々は, Neumann カーネルと Geometric Laplacian カーネルを

表 1　根論文 'Empirical studies in discourse' に対する Neumann カーネル ($\lambda = 0.005$) の出力.

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

表 2　根論文 'Empirical studies in discourse' に対する Neumann カーネル ($\lambda = 0.001$) の出力.

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

実際の論文間の引用関係に適用した. 引用グラフは, OCR 処理した参考文献一覧から, [7] 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.

評価は, 以下のように行った. 上記引用グラフから共引用グラフを作り, これに対してまず, 様々なパラメータ設定のもとで, Neumann カーネル行列, Geometric Laplacian カーネル行列, 変更版 Geometric Laplacian カーネル行列を計算する. 根節点 (根論文) を選び, 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 10 個を取り出し, 各々を与える列番号から, 該当する論文を抽出する. 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 (authority 度) 順位との相関について調べる.

5.1 Neumann カーネルの性能評価

根論文 'Empirical studies in discourse' (Marilyn A. Walker and Johanna D. Moore; Computational Linguistics 23(1):1–12, 1997) に対する Neumann カーネルの出力結果を, 表 1 ($\lambda = 0.005$), 表 2 ($\lambda = 0.001$) に掲げる. 表中, 最左列 (K) が Neumann 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 (authority 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.

表 3　根論文 'Empirical studies in discourse' に対する Geometric Laplacian カーネル ($\lambda = 0.001$) の出力.

K	C	H	論文題目
1	1	771	Empirical studies in discourse
2	2	1061	Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue
3	2	604	Message Understanding Conference (MUC) tests of discourse processing
4	2	201	The reliability of a dialogue structure coding scheme
5	2	76	Assessing agreement on classification tasks: the Kappa statistic
6	2	50	Attention, intentions, and the structure of discourse
7	2	1	Building a large annotated corpus of English: the Penn Treebank
8	-	198	A prosodic analysis of discourse segments in direction-giving monologues
9	-	96	Centering: a framework for modeling the local coherence of discourse
10	-	115	Combining multiple knowledge sources for discourse segmentation

表 1 で示されているように, $\lambda = 0.005$ の時, Neumann カーネルの順位付けは HITS の順位付けと一致している.

表 2 において, λ を 0.001 に下げると Neumann カーネルの出力は HITS の順位付けとかなり違ったものになる. 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, Neumann カーネルの順位付けは, λ 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, Neumann 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.

5.2 Geometric Laplacian カーネルの性能評価

5.1 節と同一の根論文に対する Geometric Laplacian カーネルの出力を, 表 3 に掲げる. なお, K 列が Geometric Laplacian カーネルによる順位であり, C 列, H 列の意味は前節と同じである.

Geometric Laplacian カーネルの出力が共引用解析で出力された 7 件にとどまらない点は, 前節で見た $\lambda = 0.001$ の場合の Neumann カーネル (表 2) と同様である. しかも, 表 3 の Geometric Laplacian カーネルの出力中, 共引用解析では出力されていない 3 件の論文題目には, いずれも 'discourse' という根論文の題目 ('Empirical studies in discourse') と共通のキーワードが含まれていることから, 根論文とまったく内容的に無関係な論文が追加されたわけではないと判断できる.

それに対し, 表 2 において Neumann カーネルで追加された 3 件のうち 2 件は discourse 関連の論文ではなく, HITS スコアが極めて高い論文である. Neumann カーネルは λ が十分小さい場合でも重要度に偏った尺度を与えるのに対し, Geometric Laplacian カーネルはより関連度に近い尺度であることが推測できる.

5.3 Geometric Laplacian カーネルにおける減衰係数の影響

我々は, 4.3 節において, λ を大きくしても Geometric Laplacian

表 5 Neumann カーネル, 変更版 Geometric Laplacian カーネルと他の尺度との相関 (平均 K-min 距離).

	Neumann カーネル					変更版 Geometric Laplacian カーネル			
	$\lambda = 0.005$	0.0045	0.004	0.0035	0.003	$\alpha = 0.01$	0.0316	0.1	0.316
HITS	20.4	76.9	86.2	87.8	88.7	20.4	27.6	48.3	95.8
Geometric Laplacian カーネル ($\lambda = 0.001$)	92.6	60.3	57.7	56.4	55.7	92.6	92.6	92.8	29.5

表 4 異なるパラメータ λ を用いた Geometric Laplacian カーネル間の相関 (平均 K-min 距離).

		Geometric Laplacian カーネル			HITS
		$\lambda = 0.001$	0.0005	0.0001	
Geometric	$\lambda = 0.001$	0.0	3.6	3.9	95.9
Laplacian	0.0005		0.0	1.9	95.9
カーネル	0.0001			0.0	96.5

カーネル行列の各要素は重要度に偏らない, と述べた. この裏付けのため, 減衰係数 λ を変えた Geometric Laplacian カーネルが出力する上位 10 件の論文リストの間の相関を引用グラフ中の全論文を対象に調べた. あわせて Geometric Laplacian カーネルと HITS アルゴリズムの出力ランキングとの相関も調べた.

出力ランキング間の相関を計る尺度として, K-min (minimizing Kendall) 距離 [5] を使用した. K-min 距離が小さい場合, 二つのランキングリストは高い類似性を持つ. 特に, K-min 距離が 0 の場合, 二つのランキングは完全に等しい.

各手法の出力 (各論文を根節点と見た際の上位 10 件の論文リスト) を全 2867 件の論文それぞれを根節点として計算し, 個々の根論文に対する出力の K-min 距離を求めた. 表 4 に示す結果は, 全論文に対する平均値を表している. この表は, Geometric Laplacian カーネルによる順位付けが λ にほとんど影響を受けないことを示している. 一方, Geometric Laplacian カーネルによる順位付けは, λ の値を大きく設定しても, HITS 順位とは, 全く似ていないことがわかる. 以上を 5.2 節の結果とあわせて考えると, Geometric Laplacian カーネルは, λ によらず関連度を表す尺度と見なせる.

5.4 Neumann カーネル, 変更版 Geometric Laplacian カーネルと関連度, 重要度との相関

本節では, 混合パラメータの変化が Neumann カーネルと, 変更版 Geometric Laplacian カーネルに与える影響, および, その他の関連度, 重要度尺度との相関を調べる. 関連度, 重要度算出法として共引用行列に対する Geometric Laplacian カーネル ($\lambda = 0.001$) と HITS authority 度をそれぞれ使用した. 各手法の出力するランキングの相関を計るため, 再び K-min 距離を使用する. 実験は (関連度と重要度の) 混合率 (Neumann カーネルにおける λ, 変更版 Geometric Laplacian カーネルにおける α) を変化させて行なった. なお, 変更版 Geometric Laplacian カーネルにおいては, 各 α に対して, λ を取りうる最大の値 (G を共引用グラフとして, $\|A(G) - \alpha D(G)\|^{-1} - \varepsilon$) に設定した. 実験結果を表 5 に掲げる.

この表から, λ を大きくした場合の Neumann カーネルと, α を減少させた場合の変更版 Geometric Laplacian カーネルが, ともに HITS のランキングに近付いていることがわかる. 逆に, λ を減少, あるいは α を増加させた場合には, 両手法ともに Geometric Laplacian カーネル ($\lambda = 0.001$) と似たランキングを出力していることがわかる. これにより, 両手法ともに混合パラメータによって重要度と関連度の間の偏りを調整できることがわかる.

6. むすび

カーネル法に基づく引用解析手法を提案した. この中で, グラフ上に定義されるカーネル法が三つの計量書誌尺度 (相対的重要度, 関連度, 重要度) に統一的解釈を与えることを示した. この解釈により, 相対的重要度は重要度と関連度の混合として定義される. 具体例として, 相対的重要度算出アルゴリズムとして Neumann カーネルを用い, 共引用解析/書誌結合 と HITS の関係を, 従来とは異なった視点から説明した. さらに Neumann カーネルにグラフの Laplace 作用素を導入することで, 古典的な関連度算出アルゴリズムである共引用解析/書誌結合の欠点を解消した. 最後に本論文で提案したカーネルに基づく手法の性能を, 実際に論文間の引用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] 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.

[3] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. 11th International World Wide Web Conference*, 1998.

[4] D. Cohn and T. Hofmann. The missing link—a probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2001.

[5] R. Fagin, R. Kumar, and D. Sivakumar. Comparing *k* lists. In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pp. 28–36, 2003.

[6] T. H. Haveliwala. Topic-sensitive PageRank. In *Proc. 11th International World Wide Web Conference*, 2002.

[7] 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.

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

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

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

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

[12] M. Ricardson and P. Domingos. The intelligent surfer: probabilistic combination of link and content information in PageRank. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2001.

[13] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[14] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 266–275, 2003.