E-033

ANP を用いた索引語ランキング手法の提案

The Proposal of the Ranking Method for the Indexing Words Using ANP

横井 健† 保福 一郎† 柳本 豪一‡
Takeru Yokoi Ichiro Hofuku Hidekazu Yanagimoto

1 はじめに

近年、急速に情報化が進み、数多くの文書情報がインターネットに代表されるネットワークを通じて公開されるようになった。上記の流れの中で、文書集合中から重要語を抽出する重要語抽出の研究が行われている。重要語抽出では、単語の出現頻度や共起関係などに着目し、統計的な尺度を用いて単語間に優劣を付ける手法が一般的である[1][2]。一方、文書集合が与えられた場合には、既存の手法のように単語の持つ情報だけではなく、文書間の相互関係も重要度の決定に利用できると考えられる。

一方、オペレーションズ・リサーチの分野では、アンケートなどの意志決定問題において、AHP（Analytic Hierarchy Process）や ANP（Analytic Network Process）[3]と呼ばれる評価手法がある。上記の手法では、2 つの側面から統合的な評価値を得ることができる。

本研究では、文書を「単語をどのように配置するか？」という意志決定問題と見なすことで、ANP を文書集合に適用し、その文書集合中における単語の重要度を文書間の関連性も考慮して評価する手法を提案する。

2 提案手法

以下では、ANP を用いた単語のランキング方法について説明する。まず、ANP の概略を説明し、本研究で提案する ANP の超行列（Super Matrix）と評価値について説明する。最後に、ANP の評価として用いる超行列の主固有ベクトルの求め方について述べる。

2.1 ANPの概要

ANP は AHP の階層性をネットワーク構造に拡張した手法である。AHP は総合目的、評価基準、代替案に分けられた 3 層の階層を有している。代替案とは評価対象を示す。AHP では、評価基準ごとに代替案を一対比較で相対的に評価し、それらの評価値を階層的に統合化する。一方 ANP では、先の AHP に対して、代替案から評価基準に対しても重み付けを行い、ネットワーク構造を構成する。そのネットワーク構造を超行列と呼ばれる行列により表現し、その行列の主固有ベクトルを統合的な評価値と見なす。ANP における超行列 S は、以下のように定義される。

$$S = \begin{bmatrix} 0 & V \\ W & 0 \end{bmatrix} \cdots (1)$$

ここで、V は、評価基準数×代替案数の行列で、代替案からみた評価基準に対する重みを要素とする。また、W は代替案数×評価基準数の行列で、評価基準からみた代替案に対する重みを要素とする。

2.2 超行列の作成方法

本研究では、文書間の関連性を考慮して索引語の重要度を決定できるように ANP の超行列を定義する。文書を ANP の枠組みに適用するために、文書中に含まれる単語を代替案、文書そのものを評価基準と見なす。まず、文書 $\mathbf{d}_i$ を文書中に含まれる索引語の出現確率により以下のような列ベクトルで表現する。

$$\mathbf{d}_i = [p(w_1|\mathbf{d}_i) \quad p(w_2|\mathbf{d}_i) \quad \cdots \quad p(w_V|\mathbf{d}_i)]^T \cdots (2)$$

ここで、$[\cdot]^T$ はベクトルの転置、$p(w_j|\mathbf{d}_i)$は、文書 $\mathbf{d}_i$ における単語 w_jの出現確率を示す。V は、文書集合中における索引語数を示す。なお、$p(w_j|\mathbf{d}_i)$は、以下のように定義する。

$$p(w_j|\mathbf{d}_i) = tf_{ij} / \sum_{j=1}^{V} tf_{ij}$$

tf_{ij}は、文書 $\mathbf{d}_i$ 中における単語 w_jの出現頻度とする。

さらに、(2)のベクトルを用いて、W を以下のように定義する。W は、文書が与えられた場合の単語に対する重みと考えることができる。なお、N は文書集合に含まれる全文書数を示す。

$$W = [\mathbf{d}_1 \quad \mathbf{d}_2 \quad \cdots \quad \mathbf{d}_N]$$

一方、代替案から評価基準を評価、つまり、単語が与えられた場合は、単語 $\mathbf{w}_j$ を w_j が与えられた際の文書の出現確率を用いて、以下のように定義する。

$$\mathbf{w}_j = [p(\mathbf{d}_1|w_j) \quad p(\mathbf{d}_2|w_j) \quad \cdots \quad p(\mathbf{d}_N|w_j)]^T \cdots (3)$$

なお、$p(\mathbf{d}_i|w_j)$は、ベイズの定理を用いて以下のように計算する。

$$p(\mathbf{d}_i|w_j) = \frac{p(w_j|\mathbf{d}_i)P(\mathbf{d}_i)}{P(w_j)}$$

ここで、$p(\mathbf{d}_i)$は、文書 $\mathbf{d}_i$ の文書集合中における出現確率、$p(w_j)$は、文書集合全体における単語 w_j の出現確率を示す。本研究では、$p(\mathbf{d}_i)$、$p(w_j)$をそれぞれ以下のように定義する。

$$p(\mathbf{d}_i) = 1/N$$

$$p(w_j) = \sum_{i=1}^{N} p(w_j|\mathbf{d}_i)\, p(\mathbf{d}_i)$$

さらに、(3)を用いて、超行列 S 中の V を以下のように定義する。

$$V = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \cdots \quad \mathbf{w}_V]$$

なお、本研究で提案する超行列 S は、列の要素の和が 1 であることから、確率行列の条件を満たしている。

次に、超行列 S から得られる主固有ベクトル $\mathbf{v}$ は、以下のようになる。

$$\mathbf{v} = [v_1 \quad \cdots \quad v_N \quad v_{N+1} \quad \cdots \quad v_{N+V}]^T$$

ここで、v_1 から、v_N までが、評価基準、つまり、文書に対

† 東京都立産業技術高等専門学校ものづくり工学科
‡ 大阪府立大学工学部知能情報工学科

する統合的な評価を示しており、v_{N+1} から v_{N+V} までが、代替案、つまり、単語の統合的な評価を示している。本研究では、索引語の重要性に着目しているので、v_{N+1} から v_{N+V} までの要素を用いる。

2.3 主固有ベクトルの求め方

ANPでは、超行列Sの主固有ベクトルのみを求めれば良いので、べき乗法を用いる。べき乗法を用いて主固有ベクトルを求めるためには、超行列Sが既約行列かつ原始行列でなければならない。本研究で提案している超行列Sは既約行列ではあるが原始行列ではない。しかし一方で、確率行列の条件を満たしているので、次の定理を用いて、原始性を付加することができる[4]。

定理　超行列（確率行列）Sが既約であれば

$$S_\alpha = \alpha I + (1-\alpha)S \quad (0 < \alpha < 1) \cdots (4)$$

は確率行列であって、原始行列となり、S の主固有ベクトルと S_α のそれとは一致する。なお、I は単位行列を示す。

上記の定理を用いて、超行列 S に原始性を付加し、べき乗法により主固有ベクトルを求める。なお、べき乗法は、文献[5]において安定かつ高速な収束が報告されている 1 次加速のべき乗法を利用した。

3　評価実験

本提案手法の特性を検証するために、英語のテストコレクション、MEDLINE collection を用いた評価実験を実施した。MEDLINE collection は、1,033 件の文書と 30 個の query、それぞれの query に対する正解判定が用意されたテストコレクションである。実験データとして、上記テストコレクションの中の Query1 と Query29 において正解と判断されている文書を用いた。本提案手法では、事前にクラスタリングが行われた文書集合中において索引語の重要度ランキングを決定する状況を想定した。文書中に含まれる索引語は Porter のアルゴリズムによりステミングを行い、さらに、SMART の Stop Word List を用いて、Stop Word の除去を行った。それぞれの文書数と索引語数、Query の内容を表 1 に示す。

また、式(4)における定数 α は、0.5 とした。

表 1 実験データの詳細

Query	文書数	索引語数	Query のテーマ
1	37	805	脊椎動物の水晶体
29	37	910	新生児の遺伝的な黄疸

実験結果を表 2 に示す。表 2 は、全索引語のなかで、総合的評価の高い 25 単語を示している。また、比較手法として、文書頻度によって重みを決定した単語ランキングを示す。ここでは、文書頻度をその索引語が含まれている文書数とする。DF_rank が文書頻度、ANP_rank が提案手法の索引語ランキングである。

表 2 を見ると、文書頻度によって得られる索引語ランキングと提案手法のランキング手法は異なったランキングになっていることが分かる。これは、文書間の関係性が考慮されたためと考えられる。Query29 では、提案手法の上位ランキング語が DF のそれと類似しているが、Query1 では違いが顕著である。表 2 に示したランキング上位の索引語が含まれている文書を検討すると、DF_rank の索引語があまり含まれていない文書でも、ANP_rank で得られた索引語は多く含まれるといった状況があった。これは、ANP における評価基準に対するランキング、つまり文書に対するランキングの影響を受けているものと考えられる。

このように、従来の頻度情報だけでは表現されないような文書間の影響も考慮した潜在的な索引語の重要性を提案手法では検証できるものと考えられる。

4　まとめ

本研究では、ANP を用いて文書間の関係を考慮した新たな索引語の重要度を決定する手法を提案した。その結果、従来の頻度情報だけに基づいたものとは異なったランキングが得られることを確認した。今後は、より多くのデータを検証し、本ランキングの特性のより詳細な解明が課題である。また、ANP の超行列における零行列部分のチューニングによるランキング特性の改良を考えている。

参考文献

[1] 小熊淳一, 内海 彰：「語の共起情報を用いたクラスタリング」,人工知能学会全国大会,2005

[2] 中川 裕志, 森 辰則, 湯本紘彰：「出現頻度と連接頻度に基づく専門用語抽出」, 自然言語処理, Vol.10, No.1, pp.27-45, 2003.

[3] Saaty. T. L.: "The Analytic Network Process", RWS Publications, Pittsburgh, 1996.

[4] 高橋磐郎：「AHP から ANP の諸問題 5」, オペレーションズ・リサーチ, Vol.43, No.5, pp.289-293, 1998.

[5] 平博行, 金谷健一：「時間変化する大規模対象行列の固有値計算の速度比較」, 信学技報, PRMU2007-135, pp.1-6

表 2 索引語ランキングの比較

Query1	
ANP_rank	len, protein, crystalline, fract, active, cell, epithel, acid, specy, lens, bovin, rabbit, cataract, soluble, weight, molecul, study, rat, antigen, show, albuminoid, electrophoret, rna, amin, compon
DF_rank	compon, effect, epithel, amin, cataract, cell, compos, gel, molecul, norm, numb, rat, chromatgraph, concept, cortex, high, increase, isolate, occur, ph, rabbit, anal, complet, differ, eye
Query29	
ANP_rank	hepat, biliar, neonat, live, jaundice, atret, congenit, case, cell, develop, infant, study, duc, extrahep, active, typ, bilirubin, bile, discuss, anoma, diseas, enzyme, syndrome, operat, infanc
DF_rank	Live, study, biliar, hepat, atret, develop, jaundic, case, infant, congenit, neonat, cell, extrahep, increase, metabol, occur, report, active, bilirubin, clinic, hist, birth, charact, consd