# 文書間の類似度を用いたランキング法

畠中　翔太†　　三浦　孝夫†

† 法政大学 工学部　情報電気電子学科　〒184–8584 東京都小金井市梶野町 3-7-2
E-mail: †07d3144@stu.hosei.ac.jp , ††miurat@k.hosei.ac.jp

**あらまし**　ベクトル空間モデルなどを用いた文書同士の類似度をリンクの重みとした仮想のリンクを生成することで, PageRank アルゴリズムにより文書間のリンク構造のみで文書の相対的な重要度を算出し，文書をランキングする．そこで本稿では，文書間の類似度を用いたランキング手法について提案する．また，本提案をベースにした応用について議論を行う．

**キーワード**　PageRank，類似度，ランキング

# Querying Documents using Similarity-based Ranks.

Syota HATAKENAKA† and Takao MIURA†

†　HOSEI University　3-7-2, KajinoCho, Koganei, Tokyo, 184–8584 Japan
E-mail: †07d3144@stu.hosei.ac.jp , ††miurat@k.hosei.ac.jp

## 1. 前　書　き

近年のインターネットの発展により, 電子化された文書が容易に, 大量に, しかも高速に入手できるようになった．電子文書・書籍は，手軽で高速に検索できることが利点があり，例えば情報爆発はこの検索技術なしにはあり得ない．実際，典型的な例として Web, Wiki, blog, twitters などでは，山のような情報を整理するために情報検索が欠かせない．

しかし, これらの情報から所望するものをどのように特定するのできるのか．従来, 情報検索に基づく技術と文書集合の特性を用いた手法が提案されている．

従来の情報検索技術では，キーワードの形で与えられた問い合わせ (query) に対して, そのキーワードにマッチする文書を選び, その存在有無 (term-matching) や出現頻度 (term frequency) の多いものを検索する．また，少ない文書に含まれる単語はそれらを特徴付けることを逆文書頻度 (inverse document frequency) と呼び，これに基づいた TF*IDF 法 が提案されている．

検索結果の適合度を改善するため, 文書の重みを算出する適合性フィードバックアルゴリズムが提案されている．しかし，従来の情報検索方法では，検索者が必要とする重要な文書を検索できているとはいえない．

文書特性に関する取扱いは，例えば検索エンジンを用いた情報検索ですでに利用されている．問合わせ質問を手掛かりに，確率や特殊なアルゴリズム, 経験的な知識により重み付けを行い，何らかの特別な評価により文書を一定の順序に並べる (ランキングする)．しかし，上位数個の文書しか閲覧しないことが多く，ランキング技術は大切である．

従来, Web ページの重要性を考えるときは，テーマ性 (theme)，正統性 (authoritative)，分配性 (distributive) の側面が考えられる．

通常, 予め決められた共通のテーマやトピックは存在しないため, 多くの人にとって興味ある共通のトピックやクラスタを抽出できればよい．テーマ検出・整理のための方法や特性化する方法は知られていない．テーマ抽出のためには, 頻出・相関するキーワードを用いて共通性を有する文書を求める方法や, 参照関係などからコミュニティの検出を行う方法が知られてるが, いずれも発見的である．

正統的な文書とは, 多くの影響力のあるトッピクを含み，他の重要な文書からその内容に言及されているものを言う．しかし，他の重要な文書が同じトッピクを有する文書とはいえず, 双方向的ではない．一方, 分配的な文書とは，必ずしも主要な内容を共有するわけではない文書であるが, 大量で多様な文書を言及する．本稿では, 正統的で分配的な文書を重要な文書と定義する．

検索結果の順位を決めるアルゴリズムをランキングアルゴリズムと呼ぶ．順位は多数のアルゴリズムの組合わせやチューニングパラメタにより決定される．多くのランキングアルゴリズムの主要なアイデアは, 重要度とキーワードとの関連性から決定される．重要度の算出のためには，Web 空間上では PageRank アルゴリズムや HITS アルゴリズムがある．本研究では, 文書の重要度の度合いに文書間の類似度を用いる．しかし Web ペー

ジのように，文書間にリンクがないので文書の類似度によって関連を生成し，PageRank アルゴリズムを拡張して適用する手法を論じる．

## 2. ベクトル空間モデルと重要な文書

### 2.1 ベクトル空間モデル

ベクトル空間モデルとは，大量の情報の中から利用者が与えた質問に対して利用者が必要と思われる文書の集合を提示するデータ表現方法である．

データ記述では文書を単語の多重集合で表現する．単語の並びを無視し，文書を多次元ベクトルとして表す．このモデルで重要な語を扱うには十分である．位置情報を失うため精密な表現ではないが，意味内容を表現しない語や記号などを無視できる．文書 $d$ をベクトル $d =< v_1, ..., v_n >$ で表すとき $i = 1, ..., n$ は予め定まった語 $w_i$ を表し，$v_i$ はその語に与える重みを意味する．重みとしては，2 値や出現頻度，また TF*IDF 値が用いられることが多い．

またベクトル空間モデルでは，情報検索操作を自然に表現することができる．実際，ベクトルの各要素は語の重みを表すため，当該部の要素を対応させればよい．データベース検索と異なり，情報検索では，完全一致解よりも，質問に多く関連する文書を探索することが要求される．質問に類似する度合いを数値で表現し，この度合い順に文書をランキングして提示する．

例えば，与えられた質問をベクトル化した質問ベクトル $q$ と文書の単語をベクトル化した多次元の文書ベクトル $d_i$ のなす角度により類似度を計算する余弦類似度がある．余弦類似度を $cos(d_i, q)$ とすると，$cos(d_i, q)$ は次式で表せる．

$$cos(d_i, q) = \frac{d_i * q}{|d_i||q|}$$

質問ベクトル $q$ と全ての文書ベクトル $d_1, d_2 \cdots d_n$ の余弦類似度を計算し，余弦類似度の大きいほど質問に似た文書であることがわかる．

### 2.2 重要な文書

質問者が必要とする文書情報を見つけるには，何らかのキーワードで与えられた検索指示に従って結果を質問者に与えるか，キーワードから想定されるトピックにおいて重要と思える文書を抽出し，文書を与えるかのいずれかである．

検索による結果には，キーワードを多く含むか，人工的に様々な評価方法，キーワードの確率モデルやスコアリングなどによる重み付けされた文書をランキング結果が解として与えられる．

確率モデルやスコアリングでは， 語の重み付けにより重要な文書と質問との類似度や適合度の度合いによりランキングするが，形態素解析の精度により順位が変動するので， あまり正確な順 位付けができない．

人工的に様々な評価方法では， アンケートや閲覧数などの人工的に文書の重要度を算出する場合には，評価者の主観性な評価による文書のランキングになってしまう．

これに対して，重要性を扱うためには，参照重要度を有する文書を算出する，つまり支持の高い文書の判定が必要となる．本研究では，文書集合から相対的に重要度を算出し，文書のランキングを行う．

## 3. 文書ランキング

### 3.1 PageRank アルゴリズム

PageRank アルゴリズムとは，Web 空間のハイパーリンク構造を利用して Web ページの相対的な重要度を算出し，Web ページをランキングするアルゴリズムである．また，PageRank アルゴリズムで算出される Web ページの重要度を示す数値を PageRank 値と呼ぶ．

例えば，ページ A からページ B への参照リンクをページ A によるページ B への支持投票とみなし，この投票数によりそのページの重要度を判断する．また，投票数（リンク数）を見るだけではなく，票を投じたページについても考慮される．重要度の高いページからの支持票（被参照リンク）は高く評価され，重要度の高いページに支持せれたページは「重要なページ」になる．こうした繰り返しによって高評価を得た重要なページには高い PageRank（ページ順位）が与えられ，検索結果内の順位も高くなる．しかし，リンクの構造のみでページの重要度を算出するので，ページの内容は考慮されない．

ページ $P_i$ の PageRank 値を $PR_i$ とし，入力リンクされているページ $P_j$ の PageRank 値を $PR_j$ とした時，次式で記述できる．

$$PR_i = \sum_{P_j \in B_{P_i}} \frac{PR_j}{|PR_j|}$$

$B_{P_i}$ は $P_i$ を指すページの集合であり，$|PR_j|$ はページ $P_j$ からの出力リンクの総数である．この繰り返しにより PageRank の値は，最終的には安定した値に収束すると期待され繰り返される．また，行列記法を用いることで繰り返しごとに 1 × n ベクトルを使い，すべての PageRank を計算することができる．そのために行列 H と 1 × n ベクトル $\pi^t$ を用い，次式で記述できる．

$$\pi^t = \pi^{t+1} H$$

行列 H は，ハイパーリンク構造に対する 2 値隣接行列を行で正規化した行列である．つまり，PageRank の計算は行列 H の固有ベクトル求めることである．しかし，現実の Web 空間はランクシンクや閉路などがあるために，固有ベクトルが収束しないので 2 回の調整をする．1 回目の調整は，出力リンクを 1 つも持たない Web ページで PageRank 値が集中しないように，出力リンクを 1 つも持たない Web ページはすべての Web ページに移動できる様に調整する．2 回目の調整は，固有ベクトルを急速に収束させるために，すべての Web ページがすべての Web ページに移動できる様に調整する．この 2 回の調整により行列 H は，次式の行列 H' のように書き換える．

$$H' = (1 - d)H + \frac{d}{N}$$

N は Web ページの総数であり，d は利用者がハイパーリンク構造に従わず移動する割合のパラメータである．

### 3.2 新ランキングアルゴリズム

本研究では，相対的に文書の重要度を算出させることで，文書ランキングをする．Web 空間で，Web ページに相対的に重要度を算出する PageRank アルゴリズムを用い，文書集合自体から文書の相対的な重要度を算出する．しかし，文書集合には，Web 空間でのリンクがないのでベクトル空間モデルを用い，余弦類似度による文書間の類似度により仮想リンクを生成する．文書の類似度によって関連を設定し，あたかもリンクが存在するかのようにし PageRank アルゴリズムを適用する．この仮想的な関連を仮想リンクと呼ぶ．Web 空間のような仮想リンク構造ができ，仮想リンクの重み付けに文書間の類似度を用いることで，文書の内容を考慮した文書の重要度を算出することができる．

記事 $d_i$ と記事 $d_j$ の類似度を求めるとき，次式から求める．

$$cos(d_i, d_j) = \frac{d_i * d_j}{|d_i||d_j|}$$

このとき，類似度が 0 以外は記事間に仮想リンクができる．また，仮想リンクの重み付けに文書間の類似度を用いることで，リンクが対称性になる．

従来の PageRank アルゴリズムでは，ハイパーリンク構造に対する 2 値隣接行列を行で正規化した行列を扱っているので，ページの内容関係なくリンクの重みを等価になる．記事 $d_0$ が記事 $d_1 \sim d_n$ の N 個にリンクされている (支持されている) とき，記事 $d_1$ から記事 $d_0$ への入力リンクの重みを $r_{01}$ としたとき

$$r_{01} = \frac{1}{N}$$

となり，$r_{01}$ は記事 $d_1 \sim d_n$ から記事 $d_0$ への入力リンクの重みになる．提案手法では，記事 $d_0$ から記事 $d_1 \sim d_n$ へのリンクがあるとき，記事 $d_1$ から記事 $d_0$ への入力リンクの重みを $rw_{01}$ としたとき

$$rw_{01} = \frac{w_{01}}{\sum_{i=1}^{n} w_{0i}}$$

$w_{01}$ は記事 $d_0$ と記事 $d_1$ の類似度とする．仮想リンクに文書間の類似度をリンクの重みに使用することで，類似した内容の記事間のリンクの重みは高くなる．

これにより，記事の内容を考慮したリンクの重み付けができる．ページの類似度からしきい値を設定し，しきい値以下の仮想リンクは削除することで，記事の内容に関連性がないリンクを削除することができる．

作成した行列は余弦類似度により，リンクの重み付けをすると記事 $d_0$ から記事 $d_0$ のリンクの重みは 1 になり，自分自身にリンクがあることになる．自分自身にリンクは存在しないので，作成した行列から類似度の 1 を取り除く．

次に，本と書籍などの同じ物だが名称の違いにより，類似度が 0 になってしまうのをカバーするために，すべてのリンクの重みにパラメータ α を加える．このときのパラメータ α は従来の PageRank の 2 回目の調整と同じ $1/N$ とする．N は文書数である．余弦類似度により求めた行列を M' としたとき，次式が記述できる．

$$M' = (1 - d)M + d\ \alpha$$

d には，同じ物だが名称の違う単語が全文書にあるとして 0.15 の割合与える．上の式の行列 M の固有ベクトルを求めることで，PageRank を求める．これにより，例えば文書 A と文書 B が類似し，文書 B と文書 C が類似していても，文書 A と文書 C が類似しているとはいえないことから，文書集合から重要な文書 (多くの話題性を含んだ文書) ランキングができる．

## 4. 実　　験

### 4.1 実験方法

実験では，新聞記事集合に対して提案手法を行う．新聞記事は，本文の第一段落のみで記事の大筋が要約することができる．毎日新聞の 2009 年の 1 月から 6 月の記事の内からランダムに抽出した 1 万記事の第 1 段落のみを使用する．記事の第 1 段落は形態素解析し，名詞のみを抽出したコーパスを使用する．

実験結果の評価方法として，PageRank のランキング上位 10 とすべての記事に対する余弦類似度の和のランキング上位 10 の重複度を用いる．PageRank のランキング上位 10 には，新聞記事集合の中で重要な記事が上位 10 にランキングされる．

提案手法では，記事間のリンクの重み付けに余弦類似度を用いたことにより，多くの話題性を含んだ記事が重要性な記事である．

重要な記事ランキングと話題性な記事ランキングの重複度が高いほど，重要な記事をランキングできたといえる．

二つのランキング $r_1$，$r_2$ の上位 10 の重複の割合は，次のように定義する．

$$Sim(r_1, r_2) = \frac{(A \cap B)}{k}$$

$A \cap B$ は二つのランキング $r_1$，$r_2$ の共通記事，$k$ はランキングされている記事数である．PageRank のランキング上位 10 とすべての記事に対する余弦類似度の和のランキング上位 10 の重複度が高いほど，多くの話題性を含んだ記事は，重要性な記事であるといえる．

### 4.2 実験結果

表 1 は，PageRank ランキングは PageRank，記事のタイトル，記事の第一段落を示す．表 2 は，余弦類似度ランキングはすべての記事に対する余弦類似度の和，記事のタイトル，記事の第一段落を示す．表 3 は，PageRank ランキングと余弦類似度ランキングの上位 10 の重複の割合を示す．表 1 のランキング上位 10 の記事と表 2 のランキング上位 10 の記事が 10 記事の内，9 記事が同記事になっている．表 3 よりランキング上位の重複の割合は 90 %であることがわかる．また，表 1 のラン

表 1　PageRank ランキング

| TOP10 | PageRank | タイトル | 記事の第１段落 |
|---|---|---|---|
| 1 | 0.03287732 | 野球：ＷＢＣ　日本がカブスなどと練習試合 | ＷＢＣ日本代表が２次ラウンドに出場 … |
| 2 | 0.03139896 | 桜開花予想：早ければ3月２０日ごろ | 民間気象会社「ウェザーニューズ」… |
| 3 | 0.03073055 | 利益供与要求：容疑で元総会屋の男逮捕－－警視庁 | 東証１部上場の八千代銀行（本店・東京都新宿区）に利益供与を求めたとして … |
| 4 | 0.0307029 | 新型インフルエンザ：新たに静岡などで感染確認 | 静岡，千葉，徳島県で３日，男性１人，女性２人の新型インフルエンザ感染が … |
| 5 | 0.02942945 | 衆院：本会議開会，２１日午後で決定　「麻生降ろし」収束 | 政府は１７日，麻生太郎首相が２１日に衆院解散踏み切ることを踏まえ … |
| 6 | 0.02902681 | GW：全国で晴れ多く | 気象庁は２８日，ゴールデンウイークに … |
| 7 | 0.0288975 | 高速道路料金：トラック，バス計８日間半額に－－お盆 | 金子一義国土交通相は３０日の閣議後会見で休日（土日祝日）の … |
| 8 | 0.02887041 | インフルエンザ：新人警官ら５２人感染－－栃木県警察学校 | 栃木県警察学校（宇都宮市若草）に７日入校した新人警察官（初任科生）ら … |
| 9 | 0.02865223 | 梅雨明け：九州南部で | 鹿児島地方気象台は１２日，九州南部 … |
| 10 | 0.02849303 | 麻生首相：中国，欧州を訪問へ | 河村建夫官房長官は２４日午前の … |
| ～ | ～ | ～ | ～ |
| 9991 | 0.00148725 | おいしい !: Ｄｏやまびこ 讃岐うどんさくさく | 「香川名物・讃岐うどんを生かした製品を」と，昨年２月に … |
| 9992 | 0.00148256 | 経済観測：「本能寺」商法＝迪 | 安い．早い．うまい．牛丼の吉野家商法の … |
| 9993 | 0.00144688 | 寝ても覚めても：歴史を掘り起こした５９歳＝冨重圭以子 | 「すごかったねえ」「いや，大したもんだ」「惜しかった」… |
| 9994 | 0.00144405 | 経済観測：財務相必読の書！＝迪 | ひと昔前までの大蔵大臣はまばゆくて重いポストだった … |
| 9995 | 0.00143725 | がんを生きる：住みなれた家で／3 病院の無関心,「壁」に | 夕食のテーブルをはさんだ，何気ない … |
| 9996 | 0.00141954 | ガンマ線銀河：早稲田大や広島大などが発見 ブレーザーとは別の種類 | 非常に波長の短い高エネルギーのガンマ線を出す新しい種類の銀河を … |
| 9997 | 0.00141367 | １０年バンクーバー冬季五輪：先住民の道標 会場，人々を誘い | ＜ｖａｎｃｏｕｖｅｒ２０１０＞ |
| 9998 | 0.00141367 | １０年バンクーバー冬季五輪：開幕まで１年 環境保護し規模は小さく | ＜ｖａｎｃｏｕｖｅｒ２０１０＞ |
| 9999 | 0.00141367 | １０年バンクーバー冬季五輪：開幕まで１年 | ＜ｖａｎｃｏｕｖｅｒ２０１０＞ |
| 10000 | 0.00136372 | 水と緑の地球環境：ソーラータウンミーティング，各地で開催 太陽光，自宅で発電次々 | ＜マイＥＣＯ＞ |

表 2　余弦類似度ランキング

| TOP10 | 余弦類似度の和 | タイトル | 記事の第１段落 |
|---|---|---|---|
| 1 | 1436.3933 | 野球：ＷＢＣ　日本がカブスなどと練習試合 | ＷＢＣ日本代表が２次ラウンドに出場 … |
| 2 | 1379.4339 | 桜開花予想：早ければ3月２０日ごろ | 民間気象会社「ウェザーニューズ」… |
| 3 | 1348.871501 | 利益供与要求：容疑で元総会屋の男逮捕－－警視庁 | 東証１部上場の八千代銀行（本店・東京都新宿区）に利益供与を求めたとして … |
| 4 | 1344.996201 | 新型インフルエンザ：新たに静岡などで感染確認 | 静岡，千葉，徳島県で３日，男性１人，女性２人の新型インフルエンザ感染が … |
| 5 | 1270.10094 | インフルエンザ：新人警官ら５２人感染－－栃木県警察学校 | 栃木県警察学校（宇都宮市若草）に７日入校した新人警察官（初任科生）ら … |
| 6 | 1258.451663 | 衆院：本会議開会，２１日午後で決定　「麻生降ろし」収束 | 政府は１７日，麻生太郎首相が２１日に衆院解散に踏み切ることを踏まえ … |
| 7 | 1248.997552 | GW：全国で晴れ多く | 気象庁は２８日，ゴールデンウイークに … |
| 8 | 1244.844031 | 高速道路料金：トラック，バス計８日間半額に－－お盆 | 金子一義国土交通相は３０日の閣議後会見で休日（土日祝日）の … |
| 9 | 1242.542849 | 梅雨明け：九州南部で | 鹿児島地方気象台は１２日，九州南部 … |
| 10 | 1238.969157 | 大相撲：夏巡業を追加 | 日本相撲協会は８日，夏巡業の日程を … |

キングの上位 4 記事と表 2 のランキングの上位 4 記事の順位が
同じである．

また，表 1 の上位 9991 から 10000 の記事はタイトルや第一

表 3　ランキングの重複度

| 比較するランキング | k | Sim |
|---|---|---|
| (PageRank ランキング, 余弦類似度ランキング) | 10 | 0.9 |

段落から，コラムであることがわかる．

### 4.3 考　　察

実験結果から，提案手法について考察を行う．

表 1 と表 2 のランキング上位 10 の記事は，重複の割合は 90 %であるので，重要な記事ができた．また，表 1 と表 2 の上位 4 から上位 9 まで順位は違うが，上位 4 から上位 9 まで同じ記事が重複している．このことから話題性のある記事ではなく，多くの話題性を含んだ記事がランキングすることができている．

表 1 のランキング下位には，コラム類がランキングされているのは，コラムは新聞記事の集合では話題性な記事ではないので，多くの話題性を含んだ記事は重要性な記事なのでコラム類が重要な記事ではないのといえるので，コラム類がランキング下位付けされている．また，上位 9997 から上位 10000 に関しては第 1 段落が＜ｖａｎｃｏｕｖｅｒ２０１０＞などからランキング下位の原因である．

提案手法での余弦類似度による仮想リンクの生成により，文書間のリンクは対称性になる．リンク関係を行列記法すると対称行列になるので，固有値は実数になり，べき乗より固有値と固有ベクトル (PageRank) を求めることができる．

また，本研究の提案手法は情報検索への適応することができる．本研究で使用したコーパスの記事の集合の代わりに Web ページの集合を使用することで，検索結果の上位に質問に対して重要なページに，類似したページの集合が検索結果の上位に表示される．これにより，従来の PageRank アルゴリズムとよりも良い検索結果が上位に表示されると考えられる．

## 5. 結　　論

本研究では，文書間の類似度を用いたランキング手法を提案した．本手法により，文書集合から重要な記事を抽出することができた．そして，本稿では新聞記事集合に対してランキング手法を示したが，Web 空間に対して適応することができると考えられる．

## 文　　献


[1] S. Brin and L. Page, The anatomy of a large scale hypertextual Web search engine, Computer Networks and ISDN Systems, 30, 107-117, 1998.

[2] T. H. Haveliwala, Topic-sensitive PageRank, Proc. 11th Int′1 World Wide Web Conf. (CD-ROM), 2002.

[3] 岡村寛之, 宮内聡, 土肥正, マルコフ決定過程による Web ページランキングアルゴリズムの提案, 数理解析研究所講究録 1383 巻 2004 年 81-86

[4] エイミー・N・ラングヴィル , カール・D・マイヤー, 岩野 和生 , 黒川 利明 , 黒川 洋 :Google PageRank の数理, 共立出版発行 ,2009