

アスペクト特徴ベクトルとリンク元ページ情報に基づく Web ページの特徴量計算

高橋 侑久[†] 大島 裕明^{††} 小山 聡^{†††} 田中 克己^{††}

[†] 京都大学工学部情報学科 〒606-8501 京都府京都市左京区吉田本町 36-1

^{††} 京都大学大学院情報学研究科 〒606-8501 京都府京都市左京区吉田本町 36-1

^{†††} 北海道大学大学院情報科学研究科 〒060-0808 北海道札幌市北区北 8 条西 5 丁目

E-mail: [†]{ytakahas,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp, ^{††}oyama@ist.hokudai.ac.jp

あらまし 本研究では、Web の持つグラフ的な構造を用いて Web ページの内容を表す特徴ベクトルを再帰的に分配することで、Web ページの内容とアスペクトを統合する方法を提案する。Web ページが一体どのようなページであるかは、その内容のみならず、そのページがどのようなページのどのような観点から参照されているかによって把握できる。こうした Web ページへの参照から測られる外面的な特徴を Web ページのアスペクトと呼ぶ。アスペクトを考慮して Web ページの特徴を考えることは、ある Web ページの Web 上における位置づけや評判を測るのに有用であると考えられる。

キーワード 特徴量, アスペクト, PageRank

1. ま え が き

近年、インターネットはますます一般へと広がり、Web は人々にとって重要な情報源の 1 つとなっている。Web 上には、様々なユーザによって作成された大量の Web ページが存在している。それらの Web ページはそれぞれに異なる内容を持っており、ユーザが求める内容の Web ページを膨大な数の Web ページ群の中から探しだし提示する検索サイトも広く用いられるようになっている。Web の特徴は、Web ページが相互にハイパーリンクによって相互に関係しあっていることであり、ハイパーリンクは検索エンジンのランキング手法に用いられている [1], [2]。これはハイパーリンクを支持と見なし、多くの Web ページに参照されている Web ページは有用である、という考えに基づいている。その一方で、ハイパーリンクは単なる支持ではなく、様々な異なる意味合いや観点が含まれていると考えることもできる。例えば、京都大学の Web ページでは、受験生、企業、在学生などへ向けての情報発信がされており、その内容は大学で行われた行事や研究成果などが主な内容となっている。外部の Web ページからは一風変わった校風を紹介する形で参照されたり、受験難関校としてやノーベル賞受賞者を多数輩出している大学として参照されていることがある。これらの中には京都大学の Web ページの中に発見できない記述もあり、リンク元の Web ページを考慮することで初めてそれを知ることができる。

このように、Web ページの特徴を記述する場合はその内容だけでなく、どのような Web ページからどのような意味合いで参照されているかも考慮されるべきである。本研究では、こうした Web ページが参照されている文脈のことを**アスペクト**と呼ぶ。Web ページのアスペクトは、その Web ページの Web

上における位置づけや評判を表すと考えられる。例えば、ある製品のページへのアスペクトは、その製品の Web 上における評判やイメージを分析するのに役立つと考えられる。

本論文では、ある Web ページの内容が持つ特徴と、他の Web ページからのアスペクトを考慮し、その Web ページの社会的な位置づけや評判を表す特徴量を抽出する方法を提案する。

2. 関 連 研 究

Web 上の有用なページを発見する手法として、Brin らによる PageRank がある [1]。この研究ではハイパーリンクによる参照を支持とみなし、多く参照されている Web ページは有用であり、そのような有用なページに参照されている Web ページはさらに有用である、という再帰的な考えによって Web ページのランキングを行う。本研究においても有用なページを発見する際の考え方として、PageRank の手法を用いている。

ユーザに非依存な PageRank に対し、ユーザ毎にパーソナライズされたランキングを作る研究がなされた [3]~[6]。その中の 1 つに Haveliwala の Topic-Sensitive PageRank がある [3]。これは、トピック毎に偏りを持つ PageRank ベクトルを複数個作り、それらをユーザの好みに合わせて配合を変えて和をとることで、ユーザ毎に異なる PageRank ベクトルを返すようにするものである。本論文で提案する手法は、単語を 1 つのトピックと考えた Topic-Sensitive PageRank と考えることもできるが、Topic-Sensitive PageRank が最終的にベクトルの和を取るのに対し、提案手法では得られたベクトルをそのまま扱う点異なる。

スパムサイト発見手法として Gyongyi らによって TrustRank が提案されている [7]。TrustRank はスパムでないことが判明

しているページからスコアを伝播することでスパムサイトへスコアを割り当てにくくしている。本研究では、Web ページの特徴の項目毎に同様のスコアを考えているため、TrustRank の一種の多次元拡張と見ることができるが、要素間でスコアのやりとりをすることで特徴を際立たせようとしている点において、単なる多次元拡張ではなくなっている。

リンク元のページ情報から、対象 Web ページの特徴を抽出しようとする研究としては、Glover らによる研究がある [8]。この研究ではある Web ページへの概要を表す単語をリンク元 Web ページのアンカーテキスト周辺から取得することで Web ページの分類と特徴付け及びクラスタの名前付けを行っている。本研究とはリンク元 Web ページを集合として考え、それら情報と対象 Web ページの特徴を合成する点が異なる。

Web ページに対するアスペクトを発見しようとする研究としては、是津らによる研究がある [9]。この研究では、ある Web ページへのリンク元の Web ページ毎に、リンクアンカーやその周辺コンテンツからその Web ページへのコンテキストを求め、類似したコンテキストを要約することでアスペクトを求める。コンテキストを求める際には、リンク元ページをリンクしているページのリンクアンカーなども考慮される。それ故に、ある Web ページへのアスペクトは、その Web ページが参照している Web ページへのアスペクトに影響を与える。この再帰的にアスペクトが影響しあう構図が、本研究における再帰的な計算を着想するきっかけとなった。その一方で、この研究ではアスペクト自体を求めようとしているのに対し、本研究では Web ページの内容の持つ特徴とアスペクトを統合しようとしている点が異なる。

3. Web ページの特徴

あるオブジェクトの特徴を考える場合、着目点が異なれば、抽出される特徴も異なってくる。Web ページの特徴を考える場合、一般的方法としては HTML の構造や文書に注目する方法が挙げられる。本研究では、まず Web ページの文書に着目した。

その一方で、内容特徴のみに頼って Web ページの特徴を正しく抽出することは難しくなっている。例えば、企業の Web サイトのトップページからは特徴語として、その企業を端的に表すような単語が抽出されるべきであると考えられる。その一方で、トップページにはその企業からのお知らせやニュースなどの項目が並んでいることが多く、その内容は時とともに変わるので単純に内容特徴を利用した場合、抽出結果がそのページを取得した時刻に依存したものになってしまう。その結果、それらのニュースの中に繰り返し登場する単語があった場合、その単語が特徴語として抽出されてしまうことが考えられるが、これらは不適切な抽出結果であると言える。そこで、Web ページに対する外部からの視点を考える。

本節では、本研究において熱かった Web ページの 2 つの特徴、内容特徴と参照特徴がどのような観点に基づいて考えられるものであるか説明し、そして、これら 2 つを合わせた統合特

徴の考えを導入する。

3.1 内容特徴

Web ページの内容を考える場合、ページの文書やデザインなど様々な要素を用いることができる。Web ページの内容特徴とは、このようなそのページ自身から考えられる特徴量のことである。本研究では、簡単のために、内容特徴として文書特徴量だけを用いて内容特徴を表す。

3.2 参照特徴とアスペクト

Web ページの特徴を表す指標としてどのようなページのどのような観点から参照されているかというものが考えられる。Web ページのアスペクトとは、上記の内容特徴のような Web ページの内容に対する特徴ではなく、Web ページがどのように見られているかを表す意味的な外面のことである。一般に、1 つの Web ページは複数のアスペクトを持ち、1 つのアスペクトは複数のリンク元 Web ページによって参照される。例えば、京都大学の Web ページの場合、「受験難関校」や「ノーベル賞受賞者多数輩出」などのアスペクトが考えられ、複数の予備校関連の Web ページが「受験難関校」としてのアスペクトを参照していることが考えられる。

この場合、Web ページが持つ 1 つ 1 つのアスペクトではなく、どのようなアスペクト集合を持つか、という観点から Web ページの特徴を考えることができる。Web ページの参照特徴とは、このようにリンク元ページ全体から見て、どのように見られているかを表したものである。

3.3 統合特徴

3.1 節と 3.2 節で説明した内容特徴と参照特徴は、それぞれ Web ページの内面と外見だけに着目した特徴量であり、Web ページの統合特徴とは、このような内面と外見両方を考慮した特徴である。内容特徴と参照特徴を合わせたものとして統合特徴を求める場合、これら 2 つの要素間の重み付けを考える必要がある。

4. 特徴量抽出

Web ページの内容特徴ベクトルと参照特徴ベクトルの 2 つを考え、それらを用いて統合特徴ベクトルを記述する方法を説明する。まず、1 つの Web ページとそのリンク元 Web ページ集合に着目し、参照特徴ベクトルと統合特徴ベクトルの計算式を示す。次に、それら個々の Web ページに対して導入された計算式を行列を用いて Web グラフ上のページ集合に対して同時に扱えるように拡張する。

4.1 内容特徴ベクトル

文書の特徴は、現れている単語の本文中における出現回数と比較したい文書集合におけるその単語の出現頻度から測ることができる。このような本文中に出現する単語の重み付け手法として tf-idf 値 [10] が広く用いられるが、これは Web ページの

特徴を単語を基に記述する場合にも用いることができる。本研究では、df 値として Yahoo! Search^(注1)で単語を検索した時のヒット件数を用いて、各単語の tf-idf 値を求め、これらの合計が 1 となるように正規化したものを、各 Web ページの内容特徴ベクトルとして用いた。

4.2 参照特徴ベクトル

どのようなアスペクトのもとで参照されているかを測るには、自然言語処理を用いたアプローチがまず考えられるが、本研究ではリンク元ページがどのようなページであるかを見ることで、Web ページが持つ参照特徴を表すことを考える。そこで、Web ページ p_i の参照特徴ベクトル $\mathbf{v}_A(p_i)$ を、リンク元ページの統合特徴ベクトル $\mathbf{v}_I(p_j)$ を用いて、次のように表す：

$$\mathbf{v}_A(p_i) = \sum_{p_j \in B_{p_i}} w_{p_i, p_j} \cdot \mathbf{v}_I(p_j) \quad (1)$$

ここで、 B_{p_i} は p_i のリンク元の Web ページ集合であり、 w_{p_i, p_j} は p_j の p_i の参照特徴ベクトルへ対する有用度を表す。

リンク元ページ p_j の有用度を考えるにあたり、次のような仮説を立てた。

多くの Web ページに参照されているがその Web ページ自体はリンクをあまり持たないような Web ページからのアスペクトは、そうでない Web ページからのアスペクトより重要である。

この仮説を設定した理由を以下に示す。

多くの Web ページに参照されている Web ページは有用な Web ページであると考えられるため、そうでない Web ページに比べてトラフィックが大きい。従って、その Web ページからの参照のされ方は多くのユーザに影響する。また、出リンクをあまりもたない Web ページは、大量の Web ページへリンクしているページよりリンク先のことを正しく捉えていると考えられる。そのため、そのような Web ページからのアスペクトは有用である。

以上の仮定に基づき重み w_{p_i, p_j} を、次のようにおく。ここで、ページ p_j のリンク先ページ集合を D_{p_j} と表し、 $|D_{p_j}|$ で p_j の出リンク数を表すとする。

$$w_{p_i, p_j} = \frac{1}{|D_{p_j}|} \quad (2)$$

式 (1)、式 (2) より、参照特徴ベクトル $\mathbf{v}_A(p_i)$ は次のように表すことができる。

$$\mathbf{v}_A(p_i) = \sum_{p_j \in B_{p_i}} \frac{\mathbf{v}_I(p_j)}{|D_{p_j}|} \quad (3)$$

式 (3) では、リンク元ページ p_j の統合特徴ベクトル $\mathbf{v}(p_j)$ を p_j の出リンク数で割っている。これは

- リンク元のページがどのようなページであるかを知るにはそのページの統合特徴ベクトルを見ればよい

- 出リンク数で割ることで、出リンクが少ないほど影響力が大きくなる
- という考えを表している。仮説の中で設定した、多くの Web ページに参照されているページの影響力を大きくすべきだ、という考えの取り扱いとは 5.2 節の中で説明する。

4.3 統合特徴ベクトル

統合特徴ベクトル $\mathbf{v}_I$ を参照特徴ベクトル $\mathbf{v}_A$ と内容特徴ベクトル $\mathbf{v}_C$ を用いて次式のように表す。

$$\mathbf{v}_I = \alpha \cdot \mathbf{v}_A + (1 - \alpha) \cdot \mathbf{v}_C \quad (4)$$

ここで α は、内容特徴ベクトルに対する参照特徴ベクトルの重みを表す。 α は $0 \leq \alpha \leq 1$ を満たす定数とする。統合特徴ベクトルは、 $\alpha = 0$ の時内容特徴ベクトルと等しくなり、 $\alpha = 1$ の時参照特徴ベクトルと等しくなる。

式 (4) と式 (3) より、Web ページ p_i の統合特徴ベクトル $\mathbf{v}_I(p_i)$ は、リンク元ページの統合特徴ベクトルを用いて次のように表される。

$$\mathbf{v}_I(p_i) = \alpha \sum_{p_j \in B_{p_i}} \frac{\mathbf{v}_I(p_j)}{|D_{p_j}|} + (1 - \alpha) \mathbf{v}_C(p_i) \quad (5)$$

5. 数値計算

5.1 反復法

式 (5) を計算するにあたり、本研究では反復法を用いた。 $\mathbf{v}_{k+1}(p_i)$ を $k+1$ 回目の繰り返しにおける Web ページ p_i のベクトルを表すとする。内容特徴ベクトルを初期値として、式 (5) の規則を全ての Web ページに対して順次適用していくと、次の式が成り立つ。

$$\mathbf{v}_0 = \mathbf{v}_C \quad (6)$$

$$\mathbf{v}_{k+1}(p_i) = \alpha \sum_{p_j \in B_{p_i}} \frac{\mathbf{v}_k(p_j)}{|D_{p_j}|} + (1 - \alpha) \mathbf{v}_C(p_i) \quad (7)$$

式 (7) に対して $\mathbf{v}_k(p_i)$ が十分に収束したとすると

$$\mathbf{v}_I(p_i) = \mathbf{v}_k(p_i) \quad (8)$$

式 (3) より、Web ページ p_i の参照特徴ベクトル $\mathbf{v}_A(p_i)$ は統合特徴ベクトル $\mathbf{v}_I(p_i)$ と内容特徴ベクトル $\mathbf{v}_C(p_i)$ を用いて

$$\mathbf{v}_A(p_i) = \frac{\mathbf{v}_I(p_i) - (1 - \alpha) \mathbf{v}_C(p_i)}{\alpha} \quad (9)$$

と表される。

5.2 計算式の行列表現

式 (7) を行列を用いて表現することを考える。これは、行列を用いることで、式 (7) において、1 つの Web ページごとに適用されていた手続きを、全てのページに対して一度に行うことができるようにするためである。そのために、 $\text{count}(\text{webpage}) \times \text{count}(\text{webpage})$ 行列 $\mathbf{H}$ と $\text{count}(\text{webpage}) \times \text{count}(\text{term})$ 行列 $\mathbf{M}_{k+1}$ を導入する。 $\text{count}(\text{webpage})$ と $\text{count}(\text{term})$ はそれぞれ Web ページの数と特徴ベクトルの次元数を表す。 $\mathbf{H}$ は列

(注1) : <http://search.yahoo.co.jp>

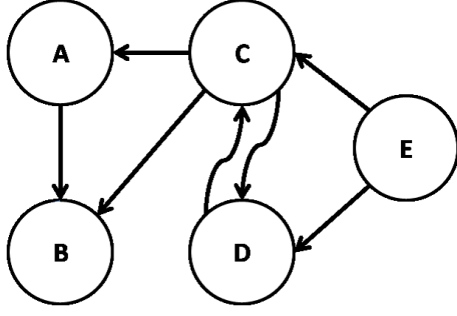


図 1 5つの Web ページからなる Web.

で正規化されたハイパーリンク行列であり

$$\mathbf{H}_{ij} = \begin{cases} \frac{1}{|p_i|} & p_i \in B_{p_j} \\ 0 & p_i \notin B_{p_j} \end{cases} \quad (10)$$

となる． $\mathbf{M}_{k+1}$ は $k+1$ 回目の繰り返しにおける各 Web ページの特徴ベクトルを行として縦に並べた行列である．図 1 のような Web があったとする．Web ページの特徴ベクトルの次元を表す単語として， t_1, t_2 の 2 つが存在し，Web ページ p_i の $k+1$ 回目の繰り返しにおける特徴ベクトルが $\mathbf{v}_{k+1}(p_i) = (f_1^{k+1}(p_i) \ f_2^{k+1}(p_i))$ と表されたとすると， $\mathbf{H}$ と $\mathbf{M}_{k+1}$ は

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & \frac{1}{3} & 0 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (11)$$

$$\mathbf{M}_{k+1} = \begin{pmatrix} \mathbf{v}_{k+1}(p_A) \\ \mathbf{v}_{k+1}(p_B) \\ \mathbf{v}_{k+1}(p_C) \\ \mathbf{v}_{k+1}(p_D) \\ \mathbf{v}_{k+1}(p_E) \end{pmatrix} = \begin{pmatrix} f_1^{k+1}(p_A) & f_2^{k+1}(p_A) \\ f_1^{k+1}(p_B) & f_2^{k+1}(p_B) \\ f_1^{k+1}(p_C) & f_2^{k+1}(p_C) \\ f_1^{k+1}(p_D) & f_2^{k+1}(p_D) \\ f_1^{k+1}(p_E) & f_2^{k+1}(p_E) \end{pmatrix} \quad (12)$$

となる．このような $\mathbf{H}$ と $\mathbf{M}_{k+1}$ を用いると式 (7) は次の様に表される．

$$\mathbf{M}_{k+1} = \alpha \mathbf{H} \mathbf{M}_k + (1 - \alpha) \mathbf{M}_0 \quad (13)$$

$\mathbf{H}$ を左から $\mathbf{M}_k$ に掛けることは，各 Web ページの k 回目の繰り返しにおける特徴ベクトルを，リンク先のページへ均等に分配する動作をシミュレートする．

5.3 アルゴリズム

以上をまとめると，アルゴリズムは以下のように表される．

各 Web ページの内容特徴ベクトルを並べた行列 $\mathbf{M}_0$ を作る
Web ページのグラフ構造からハイパーリンク行列 $\mathbf{H}$ を求める

$n \leftarrow 0$

表 1 テストセットのデータ数

ページ数	9942
インデックス登録数	1218
リンク数	23153
単語数	13825

repeat

$$\mathbf{M}_{n+1} = \alpha \mathbf{H} \mathbf{M}_n + (1 - \alpha) \mathbf{M}_0$$

$$n \leftarrow n + 1$$

until $\|\mathbf{M}_{n+1} - \mathbf{M}_n\|_2 \leq \tau$

return $\mathbf{M}_{n+1}$

ここで τ は $\mathbf{M}_k$ が収束したかを判定するための閾値を表す．

6. 実験

本節では，本研究において提案した Web ページの参照特徴ベクトルおよび統合特徴ベクトルの抽出手法に対する評価実験の結果を述べる．

実験の為，今回は京都大学ホームページ^(注2)を中心としてページ群を収集した．Yahoo! Search でリンク元ページの検索^(注3)を行うことで，出リンクと入リンクの偏りを出来る限り減らすよう工夫した．また，日本語のページのみを対象とし，非日本語ページへのリンクは全て無視している．これらのデータを簡単のため，テストセットと呼ぶことにする．テストセットに含まれる具体的なデータ数は表 1 のようになっている．

本節ではまず，提案手法によって得られる参照・統合特徴ベクトルを持つ全体的な傾向について考察する．次に，個々の Web ページの実際に得られたベクトルに注目し，その性質について考察する． $\alpha = 0.85$ の場合の実験結果を用いる．

6.1 Web ページ群に対する性質

ここでは，提案手法によって得られる参照特徴ベクトル $\mathbf{v}_A$ 及び統合特徴ベクトル $\mathbf{v}_I$ について，Web ページ群として捉えた時にどのような性質を持つか考察する．各ベクトル間の乖離を見るために，今回は $\cos$ 類似度を用いた．

6.1.1 内容特徴ベクトルと参照特徴ベクトルの乖離

内容特徴ベクトルと参照特徴ベクトルの類似度 $\cos(\mathbf{v}_C, \mathbf{v}_A)$ と Web ページ数及び平均 PageRank 値の関係は図 2 のようになった．図中の緑の線は，全ページの PageRank 値の平均を表し，値は 0.00019 となっている．

この図から，内容特徴と参照特徴が極端に似ている，もしくは似ていないページは数が少ないだけでなく，PageRank 値も低くなる傾向があることが分かる．一方，大多数の Web ページは中間くらいの類似度であることが分かる．また，平均 PageRank

(注2) : <http://www.kyoto-u.ac.jp/>

(注3) : クエリとして "link:<http://www.kyoto-u.ac.jp/>" を用いると，商用検索システムでは指定した URL のページへのリンクを含むページを検索することができる．この場合であれば，京都大学ホームページのリンク元ページが検索される．

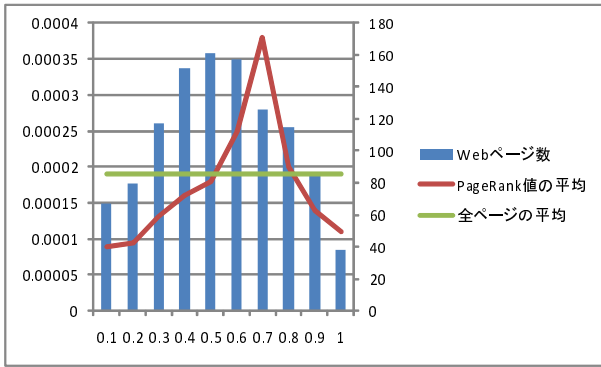


図 2 内容特徴ベクトルと参照特徴ベクトルの乖離と Web ページ数及び平均 PageRank 値の関係と全ページの平均 PageRank 値

値のピークにおける類似度 $0.6 < \cos(\mathbf{v}_C, \mathbf{v}_A) < 0.7$ が, Web ページ数のピークにおける類似度 $0.4 < \cos(\mathbf{v}_C, \mathbf{v}_A) < 0.5$ よりも高いことから, 平均的なページよりも, 似た内容のページによくリンクされているページは PageRank 値が高くなる傾向があることが分かる. 比較的似ているページに多くリンクされるということは, このようなページへのリンクは, 明確な理由があって貼られているとものであると考えられる. そのため, この結果は PageRank のロバスト性を示す 1 つの側面であると考えられる.

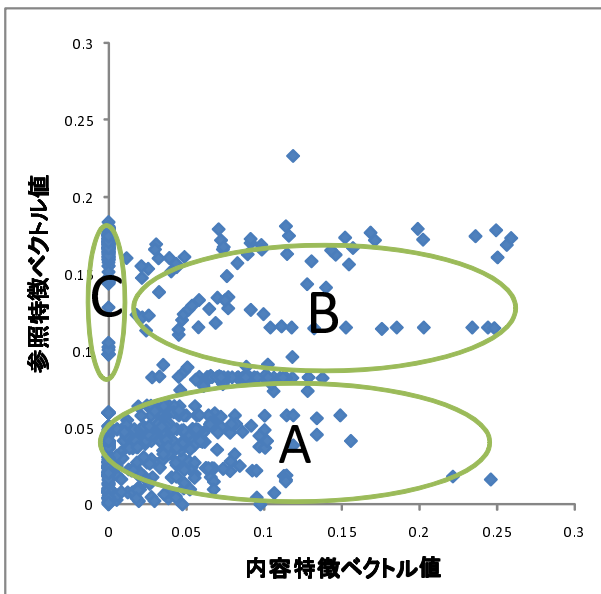


図 3 各 Web ページの内容特徴ベクトル及び参照特徴ベクトルにおける”京都大学”の値

各 Web ページを内容特徴ベクトル及び参照特徴ベクトルにおける”京都大学”の値を用いてプロットしたものが図 3 である. ここでは, 参照特徴ベクトルはベクトル値の合計が 1 に正規化されたものを用いている. 図中で, 右にあるページほど単語”京都大学”の内容特徴ベクトルにおける値が高く, 上にあるページほど参照特徴ベクトルにおける値が高いことを表している. また, 参照特徴軸上のページは, 内容に”京都大学”を含ま

ないが, リンク元ページには含まれているページを表している.

全体として上下方向に潰れたように散布しているのは, 通常, 内容特徴ベクトルよりも参照特徴ベクトルの方がベクトルの非零成分が多く, そのため正規化した際に, 個々の値が小さくなる傾向があるためだと考えられる. また, 図中の領域 A では内容特徴ベクトルの値が小さいページが圧倒的に多いのに対し, 領域 B では比較的均等に分布している. また, 本文に”京都大学”を含まないが, 参照特徴ベクトルの値が高いページを表す領域 C にも多くのページが存在することが分かる. そのため, 参照特徴ベクトルの値が小さいページは内容特徴ベクトルの値も小さいことが分かるが, 参照特徴ベクトルの値が大きいからと言って内容特徴ベクトルの値も大きいとは言えないことが分かる.

以上のことから, 参照特徴ベクトルの値は, 各ページにリンクしているページ群によって決定され, そのページの内容特徴ベクトルは大きく影響しないことが分かる. そのため, 提案手法における参照特徴ベクトルは, 参照特徴を表す 1 つの方法であると言える. 実際に, 参照特徴ベクトルがどの程度参照特徴を表すものとして有効であるかについて考察は 6.2 節において行う.

6.1.2 内容特徴ベクトルと統合特徴ベクトルの乖離

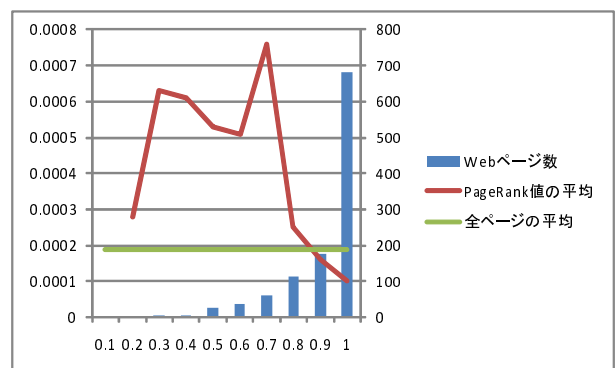


図 4 内容特徴ベクトルと統合特徴ベクトルの乖離と Web ページ数及び平均 PageRank 値の関係と全ページの平均 PageRank 値

内容特徴ベクトルと統合特徴ベクトルの類似度 $\cos(\mathbf{v}_C, \mathbf{v}_I)$ と Web ページ数及び平均 PageRank 値の関係は図 4 のようになった. $\cos(\mathbf{v}_C, \mathbf{v}_I) < 0.1$ を満たす Web ページが存在しなかったため, 該当区間における平均 PageRank 値は描写されていない.

この図から, 大多数のページが $0.9 < \cos(\mathbf{v}_C, \mathbf{v}_I)$ となっており, 内容特徴ベクトル $\mathbf{v}_C$ と統合特徴ベクトル $\mathbf{v}_I$ が非常に似ているページが多いことが分かる. また, 当該区間における平均 PageRank 値は全体平均を下回っている. その一方, ページの絶対数は少ないが, $0.1 < \cos(\mathbf{v}_C, \mathbf{v}_I) < 0.8$ において常に区間内での平均 PageRank 値が全体での平均を上回っていることから, 特に PageRank 値の低いページは $\cos(\mathbf{v}_C, \mathbf{v}_I)$ が大きくなる傾向があることが分かる. これは, PageRank 値の高いページは参照特徴ベクトルのベクトル値の合計が大きくなるこ

表 2 京都大学ホームページの内容特徴ベクトルにおける上位 10 単語

順位	内容特徴ベクトル
1	研究
2	支援
3	補佐
4	開催
5	職員
6	雇用
7	専攻
8	学生
9	更新
10	シンポジウム

表 3 参照特徴ベクトル：提案手法とベースラインにおける上位 10 単語

順位	提案手法	ベースライン
1	研究	研究
2	情報	財団
3	京都大学	助成
4	専攻	平成
5	京都	京都
6	平成	文化財
7	数理	センター
8	センター	募集
9	論文	情報
10	大学	京都大学

とが予想されるが、その一方で、内容特徴ベクトルは一定であるため、式 (4) における相対的な大きさの違いからこのような関係が生じたものと考えられる。このように、提案手法によって求められる統合特徴ベクトルは、各ページによって内容特徴ベクトルと参照特徴ベクトルの寄与する割合が大きく異なることが分かる。

6.2 個々の Web ページに対する性質

6.2.1 参照特徴ベクトル

本節では、個々の Web ページに対して得られた参照特徴ベクトルに対して考察する。ここでは、例として京都大学ホームページのトップページを用いる。

まず、京都大学ホームページの内容特徴ベクトルは表 2 のように得られた。「支援」「補佐」「開催」「シンポジウム」などが特徴語として抽出されたが、これらはトップページ中のニュースに繰り返し出現した単語である。一方、京都大学ホームページを表す単語として真っ先に思い浮かぶであろう「京都大学」という単語は 13 位に現れた。

次に、提案手法の参照特徴ベクトルを評価するために、比較対象としてベースラインを次のように定めた：

ベースライン：リンク元ページの内容特徴ベクトルを各ページの出リンク数の逆数倍し、それらの和を取ったものを参照特徴ベクトルとする。

提案手法とベースラインの上位 10 単語は表 3 のようになった。調べた結果、リンク元ページのほとんどのページに「京都

表 4 京都大学ホームページに対する $\mathbf{v}$ の上位 10 単語。

順位	$\beta = 0.25$	$\beta = 0.5$	$\beta = 0.75$
1	研究	研究	研究
2	京都大学	京都大学	京都大学
3	専攻	情報	情報
4	情報	専攻	専攻
5	開催	京都	京都
6	支援	大学	大学
7	補佐	開催	開催
8	大学	支援	科学
9	京都	補佐	教授
10	学生	学生	学生

大学」という単語が出現していた。そのため、提案手法とベースラインではともに「京都大学」が参照特徴を表す語として抽出されているが、提案手法では 3 位であるのに対しベースラインでは 10 位となっている。このような差が生じたのは、ベースラインではリンク元ページに出現する単語のみを考えるが、提案手法ではより多くのページを考慮するため、それらに共通して現れた「京都大学」はより高いスコアを獲得し、リンク元ページだけに出現した「財団」などの単語はスコアが下がったのだと考えられる。このような結果から、提案手法はベースラインと比較してロバスト性が高いと考えられる。

6.2.2 統合特徴ベクトル

6.1.2 節で述べたように、提案手法で求められる統合特徴ベクトル $\mathbf{v}_I$ は、内容特徴ベクトルと参照特徴ベクトルの割合がページによって異なるという問題がある。京都大学ホームページの場合、比較的 PageRank 値が高く、上位 10 単語において $\mathbf{v}_I$ は $\mathbf{v}_A$ とほぼ一致した。そこで、参照特徴ベクトル $\mathbf{v}_A$ のベクトル値の合計が 1 となるように正規化した $\hat{\mathbf{v}}_A$ を用いて、次のようなベクトルを考える。

$$\mathbf{v}(p_i) = \beta \hat{\mathbf{v}}_A(p_i) + (1 - \beta) \mathbf{v}_C \quad (14)$$

京都大学ホームページに対して、 $\beta = 0.25, 0.5, 0.75$ と動かした時、 $\mathbf{v}$ の上位 10 単語は表 4 のようになった。表 2 より、内容特徴ベクトルにおいて、「支援」は 2 位「学生」は 8 位であったが、 β を大きくしていくにつれてこれらの単語の順位差は縮まり、 $\beta = 0.75$ では「支援」はランキング圏外まで落ちている。このような結果が生じたのは、「学生」が両特徴ベクトルに共通して高いスコアを持っていたのに対し、「支援」が内容特徴ベクトルでしか高いスコアを持っていなかったためである。

このように、内容特徴ベクトルと正規化された参照特徴ベクトルを用いることで、対象ページのより本質的に重要な単語を抽出できるため、 $\mathbf{v}$ を統合特徴の一つのベクトル表現として用いることが可能であると考えられる。

7. まとめと今後の課題

本論文では、Web ページの参照特徴、統合特徴の考えを示し、Web のリンク構造を用いて参照特徴ベクトル及び統合特徴

ベクトルを抽出する手法を提案した。今後は次のような課題に取り組む:

- 式 (14) における, 最適なパラメータ β を求める.
- 得られた参照特徴ベクトル及び統合特徴ベクトルの客観的な評価を行う.

謝辞 本研究の一部は, 京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」, および, 文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」, 計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」(研究代表者: 田中克己, A01-00-02, 課題番号: 18049041) によるものです. ここに記して謝意を表します.

文 献

- [1] Sergey Brin, L. P.: The anatomy of a large-scale hypertextual Web search engine, *Seventh International World Wide Web Conference* (1998).
- [2] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, pp. 604–632 (1999).
- [3] Haveliwala, T. H.: Topic-sensitive PageRank, *Eleventh International World Wide Web Conference* (2002).
- [4] Taher Haveliwala, Sepandar Kamvar, G. J.: An Analytical Comparison of Approaches to Personalizing PageRank, *Technical Report, Stanford University* (2003).
- [5] Glen Jeh, J. W.: Scaling personalized web search, *Twentieth International World Wide Web Conference*, pp. 271–279 (2003).
- [6] Michelangelo Diligenti, Marco Gori, M. M.: Web page scoring systems for horizontal and vertical search, *Eleventh International World Wide Web Conference*, pp. 508–516 (2002).
- [7] Zoltan Gyongyi, Hector Garcia-Molina, J. P.: Combating web spam with trustrank, *Thirtieth International Conference on Very Large Data Bases* (2004).
- [8] Eric J. Glover, Kostas Tsuioutsoulouklis, D. M. P. and Flake, G. W.: Using Web Structure for Classifying and Describing Web Pages, *Eleventh International World Wide Web Conference*, pp. 562–569 (2002).
- [9] 田中 克己是津 耕司: Web ページのアスペクトの発見, *DBSJ Letters*, Vol. 2, No. 4, pp. 15–18 (2004).
- [10] Gerard Salton, C. B.: Term-weighting approaches in automatic text retrieval, *Information Processing and Management*, Vol. 24, pp. 513–523 (1988).