ハイパリンクで結ばれた隣接ページの内容に基づくWebページのためのTF-IDF法の改良

杉山　一成[†]　　波多野賢治[†]　　吉川　正俊[††]　　植村　俊亮[†]

Improvement in TF-IDF Scheme for Web Pages Based on the Contents of Their Hyperlinked Neighboring Pages

Kazunari SUGIYAMA[†], Kenji HATANO[†], Masatoshi YOSHIKAWA[††], and Shunsuke UEMURA[†]

あらまし　ベクトル空間法に基づいた情報検索システムでは，文書を特徴づけるために，しばしばTF-IDF法が用いられる．しかし，Webページのようなハイパリンク構造を有する文書の場合には，ハイパリンクで結ばれた隣接ページの内容を利用することによって，より正確にWebページの内容を特徴づけられると考えられる．そこで本論文では，ハイパリンクで結ばれた隣接ページの内容を用いて，Webページ向けにTF-IDF法を改良するための手法を提案し，その手法の有効性を確認する．

キーワード　WWW，情報検索，TF-IDF法，ハイパリンク

1. まえがき

World Wide Web（WWW）は，その利用者にとって様々な情報を入手するための有用な情報源である．Web検索エンジンが格納しているWebページ数は，約30億ページ以上といわれており[1]，その数が今後増加し続けるのは明らかである．したがって，Web上の価値ある情報を見つけ出すことは，利用者にとって困難となる一方である．こうした状況の中で，Web検索エンジンは，価値ある情報を効率的に見つけるために，最もよく使われる方法の一つである．また，Web検索エンジンは，Webページの特徴付けの方法に基づいて二つの世代に分類される[2]．Webの初期段階に開発された第1世代の検索エンジンにおいては，木構造で表現可能な半構造化文書であるWebページの，根に近い部分に存在するtitleタグで囲まれた単語などが，そのWebページの索引として利用されているだけであった．したがって，こうした特徴付け手法では，利用者はその検索精度に満足することができなかった．このような問題に対処するために，第2世代の検索エンジンにおいては，Webページのハイパリンク構造が考慮されている．例えば，PageRank [3]は検索エンジンGoogle(注1) [4]に，HITS（Hypertext Induced Topic Search）[5]は，CLEVERプロジェクト[6]における検索エンジンとして，それぞれ適用されているアルゴリズムである．これらのアルゴリズムは，ハイパリンク構造を利用してWebページを重み付けすることにより，第1世代の検索エンジンに比べて，より高い検索精度を実現している．しかしながら，これらのアルゴリズムは，(1) Webページに対する重みが単に定義されているにすぎない，(2) ハイパリンクで結ばれたWebページ間の内容の関連性が考慮されているわけではない，という欠点があるため，利用者の検索語に適合しないWebページが，しばしば検索結果の上位に順位付けされるという問題が依然として残っている．

これらの問題点を考慮すると，検索語に適合したWebページを利用者に提供するためには，より正確にWebページの内容を表現する手法を開発することが

† 奈良先端科学技術大学院大学情報科学研究科，生駒市
Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan
†† 名古屋大学情報連携基盤センター，名古屋市
Information Technology Center, Nagoya University, Nagoya-shi, 464–8601 Japan

(注1)：http://www.google.com/

必要であると考えられる．そこで，対象 Web ページへリンクしている Web ページ（以下，in-link ページと呼ぶ），対象 Web ページからリンクしている Web ページ（以下，out-link ページと呼ぶ）の両方を考慮した上で特徴ベクトルを計算する必要があると考えられる．本論文では，あらかじめ TF-IDF 法 [7] に基づいて作成された Web ページの特徴ベクトルを，その Web ページの in-link ページ，out-link ページの特徴ベクトルに基づいて改良する手法を提案し，その有効性を確認するための評価実験の報告を行う．我々の手法は，第 2 世代の検索エンジンと比較して，ハイパリンクで結ばれた隣接する Web ページの内容を，対象とする Web ページの内容に反映させることによって，より正確に Web ページを特徴づけられる点に新規性がある．

本論文の構成は次のとおりである．**2.** では，WWW のハイパリンク構造を利用した関連研究について述べる．**3.** では，ハイパリンクで結ばれた隣接ページを用いることによって，Web ページの特徴ベクトルを改良する新たな手法を提案し，**4.** では，その提案手法を評価するための実験結果を示し，その結果について考察する．最後に **5.** では，本論文のまとめと今後の課題について述べる．

2. 関 連 研 究

ハイパリンク構造は WWW の特徴の一つであり，WWW の利用者はこのハイパリンク構造を通じて巨大な WWW 空間を容易にめぐることができる．したがって，Web 情報検索に関する多くの研究が，WWW のハイパリンク構造に着目している．本章では，WWW のハイパリンク構造を利用した情報検索システムの関連研究，特に，「最適な文書粒度」の概念に基づいた情報検索システム，及び Web ページに対する重み付けアルゴリズムである HITS と PageRank について振り返る．

2.1 「最適な文書粒度」の概念に基づいた情報検索システム

Tajima ら [8] は，Web 構造解析の結果である “cuts” という概念を使う技術を，WWW に対する検索単位として提案した．また，このシステムは，(1) すべての検索語を含む極小部分グラフを WWW ハイパリンク構造から見つけること，(2) 極小部分グラフ内のキーワードの局所性に基づいて，それぞれの極小部分グラフに関して検索語に対する類似度を計算すること，によって複数の検索語を含む検索結果を順位付けするように拡張されている [9]．これらの研究に続いて，Li ら [10] は複数の Web ページから構成される文書に対して，一つの極小検索単位となる “information unit” という概念を導入し，この単位による Web ページ検索のための新たな枠組みを提案した．しかし，これらの手法は，ハイパリンク構造を解析し，Web ページの意味を発見するために，相当な処理時間を必要とし，また，利用者が指定した検索語に適合しない検索単位を発見することもしばしば起こり得る．更に，これらのシステムが返す検索結果は，ハイパリンクで結ばれたいくつかの Web ページに，複数の検索語が分散しているため，利用者はその検索結果を直感的に理解しがたいという問題も存在する．

2.2 HITS アルゴリズム

HITS アルゴリズム [5] は，CLEVER プロジェクト [6] の検索エンジンに適用されている．このアルゴリズムは検索語に依存し，その検索結果となるページにリンクしている，あるいは検索結果となるページからリンクされているページの集合 S を考慮する．S において，多くの Web ページからリンクされているページは「オーソリティ」，また，多くの Web ページにリンクしているページは「ハブ」と呼ばれる．すなわち，より良いオーソリティは良いハブからリンクされ，より良いハブは良いオーソリティへリンクしている．$H(p)$，$A(p)$ をそれぞれ，Web ページ p のハブスコア，オーソリティスコアとすれば，これらの値はすべてのページ p に対して，次式を満たすように定義される．

$$H(p) = \sum_{u \in S | p \to u} A(u), \quad A(p) = \sum_{v \in S | v \to p} H(v),$$

ここで，$H(p)$ と $A(p)$ はすべてのページ p に対して正規化される．これらの値は繰り返し定められ，S のリンク行列の優固有ベクトルに収束する．しかし，このアルゴリズムは，in-link ページ数が少なく，かつ out-link ページ数が多いような Web ページ p の場合には，p のハブスコアが大きくなるため，繰り返し計算を行った場合に，p がリンクする Web ページに関して，ハブスコアの和として求められるオーソリティスコアが，極端に大きくなるという問題点がある．この問題点を解決するために，拡張された HITS アルゴリズムも，いくつか提案されている [11]〜[15].

2.3 PageRank アルゴリズム

PageRank アルゴリズム [3] は，現在閲覧している Web ページに対して，利用者が全く無関係な Web ページへ確率 d で遷移する状態と，その Web ページからのリンクを確率 $1-d$ でたどる状態がモデル化されている．更に，既にたどったハイパリンクをさかのぼり，一度閲覧したページには決して戻らないということが仮定されている．この過程はマルコフ連鎖を使ってモデル化され，それぞれのページに滞在する定常確率が計算される．この確率の値は検索エンジン Google のランキング機構の一部として使われている．ここで，$C(a)$ を Web ページ a から外へ向かうリンク数とし，Web ページ p_1〜p_n が，Web ページ a にリンクしているものと仮定する．このとき Web ページ a の PageRank の値 $PR(a)$ は次式で定義される．

$$PR(a) = d + (1-d)\sum_{i=1}^{n}\frac{PR(p_i)}{C(p_i)}$$

ここで，d の値は経験的におよそ 0.15〜0.2 の間に定められる．他のページの重みは，そのページにおけるリンク数によって正規化される．PageRank は繰り返しアルゴリズムを使って計算することができ，正規化された Web のリンク行列の優固有ベクトルに対応する．このアルゴリズムの主な問題は，(1) Web ページの内容が解析されていないため，Web ページの重要度が検索語に関係なく定義される，(2) 特定の有名サイトが上位に順位付けされる傾向がある，という点にある．したがって，より正確な検索結果を得るために，本アルゴリズムを拡張したアルゴリズムが提案されている [16], [17].

3. 提 案 手 法

2.1 で述べたように,「最適な文書粒度」の概念に基づいた情報検索システムは，検索結果が利用者に理解しがたいという問題点があった．また，HITS アルゴリズム，及び PageRank アルゴリズムにおいても，(1) Web ページに対する重みが単に定義されているにすぎない，(2) ハイパリンクで結ばれた Web ページ間の内容の関連性が考慮されていない，という問題点があった．これらの問題点に基づけば，Web ページの内容を正確に表現するために，ハイパリンクで結ばれた隣接ページの内容を反映させて Web ページの特徴ベクトルを生成すべきであると考えられる．そこで，本論文では，あらかじめ TF-IDF 法に基づいて生成された Web ページの特徴ベクトルを，その隣接ページの特徴ベクトルに基づいて改良する手法を提案する．

以下，特徴ベクトルを生成する対象 Web ページを p_{tgt} と表す．また，p_{tgt} から別の Web ページへ最短でたどることのできる経路数を i と定義し，p_{tgt} から i 番目の階層には，$p_{i_1}, p_{i_2}, \cdots, p_{i_{N_i}}$ までの N_i 個の Web ページがあるものと仮定する．更に，p_{tgt} の特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ を次のように表す．

$$\boldsymbol{w}^{p_{tgt}} = (w_{t_1}^{p_{tgt}}, w_{t_2}^{p_{tgt}}, \cdots, w_{t_m}^{p_{tgt}}) \tag{1}$$

ここで，m は Web ページ集合中における単語の異なり数であり，$t_k (k = 1, 2, \cdots, m)$ はそれぞれの単語を表す．また，TF-IDF 法を用いて，$\boldsymbol{w}^{p_{tgt}}$ の各要素 $w_{t_k}^{p_{tgt}}$ を次のように定義する．

$$w_{t_k}^{p_{tgt}} = \frac{tf(t_k, p_{tgt})}{\sum_{s=1}^{m} tf(t_s, p_{tgt})} \cdot \log\frac{N_{web}}{df(t_k)} \quad (k = 1, 2, \cdots, m) \tag{2}$$

ここで，$tf(t_k, p_{tgt})$ は対象 Web ページ p_{tgt} における単語 t_k の頻度を，N_{web} は Web ページ集合中における Web ページの総数を，$df(t_k)$ は単語 t_k が出現する Web ページ数を表す．以下，$\boldsymbol{w}^{p_{tgt}}$ を「初期特徴ベクトル」と呼ぶことにする．更に，$\boldsymbol{w}^{p_{tgt}}$ を改良した特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を式 (3) のように表し，

$$\boldsymbol{w}'^{p_{tgt}} = (w_{t_1}'^{p_{tgt}}, w_{t_2}'^{p_{tgt}}, \cdots, w_{t_m}'^{p_{tgt}}) \tag{3}$$

$\boldsymbol{w}'^{p_{tgt}}$ を「改良特徴ベクトル」と呼ぶことにする．また，以下において，m 次元の二つのベクトル，

$$\boldsymbol{a} = (a_1, a_2, \cdots, a_m), \quad \boldsymbol{b} = (b_1, b_2, \cdots, b_m)$$

間の距離 $dis(\boldsymbol{a}, \boldsymbol{b})$ を，式 (4) のように定義する．

$$dis(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{k=1}^{m}(a_k - b_k)^2} \tag{4}$$

本論文では，式 (2) で定義される TF-IDF 法に基づいた，式 (1) の初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ を改良する手法として，以下で説明する三つの手法を提案する．なお，以下において，順方向とは p_{tgt} からリンクしているページ（out-link ページ，図 1 (a)，図 2 (a)，図 3 (a) において，p_{tgt} より下方向）へたどることを，逆方向とは p_{tgt} へリンクしているページ（in-link ページ，図 1 (a)，図 2 (a)，図 3 (a) において，p_{tgt} より上方向）へたどることを，それぞれ表す．

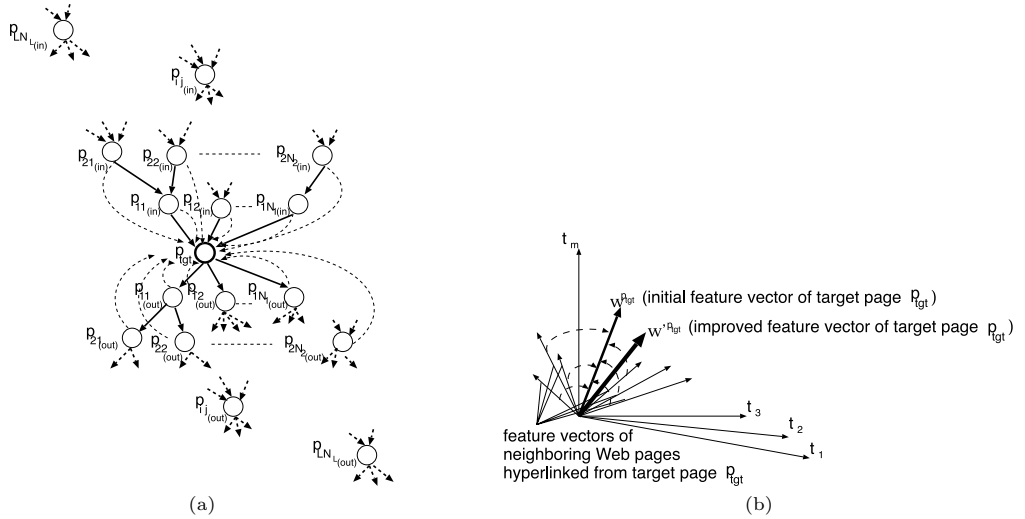

図 1　手法 I による特徴ベクトルの修正法
Fig. 1　The improvement of a feature vector as performed by Method I [(a) in the Web space, (b) in the vector space].

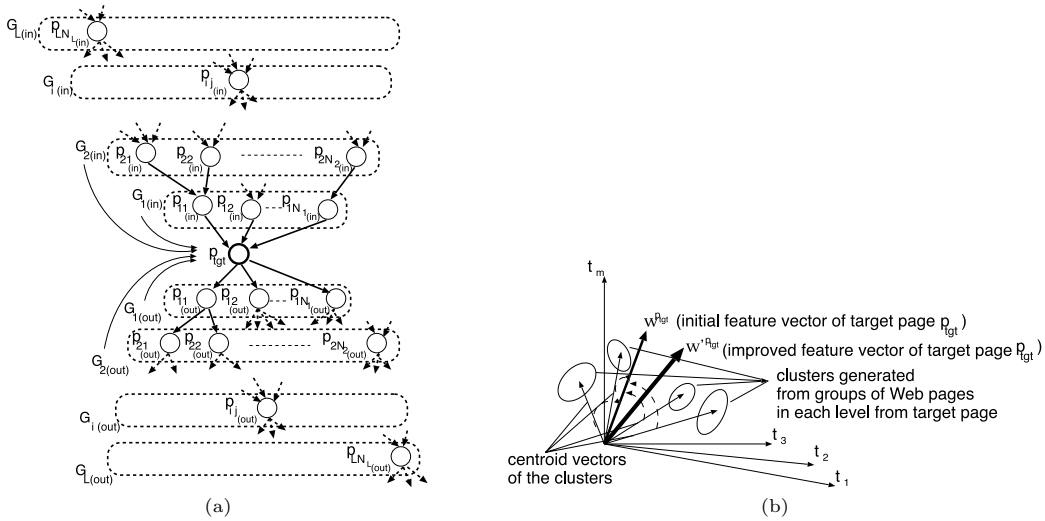

図 2　手法 II による特徴ベクトルの修正法
Fig. 2　The improvement of a feature vector as performed by Method II [(a) in the Web space, (b) in the vector space].

3.1 手　法　I

この手法では，対象ページ p_{tgt} から，逆方向に $L_{(in)}$ 階層まで，順方向に $L_{(out)}$ 階層までのすべての Web ページの特徴ベクトルを，対象ページ p_{tgt} の特徴ベクトルに反映する．これは，(1) p_{tgt} の近隣には，p_{tgt} の内容に類似した Web ページが存在する，(2) そうした Web ページは，ベクトル空間において，対象ページのすぐ近くに存在することもあれば，対象ページから離れて存在する場合もある，という考えに基づいており，初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ の各要素に，ベクトル空間における $\boldsymbol{w}^{p_{tgt}}$ と，p_{tgt} の in-link ページ，out-link ページの特徴ベクトル間の距離を反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を作成する．

例えば，図 1 (a) は，p_{tgt} から逆方向，順方向に 2

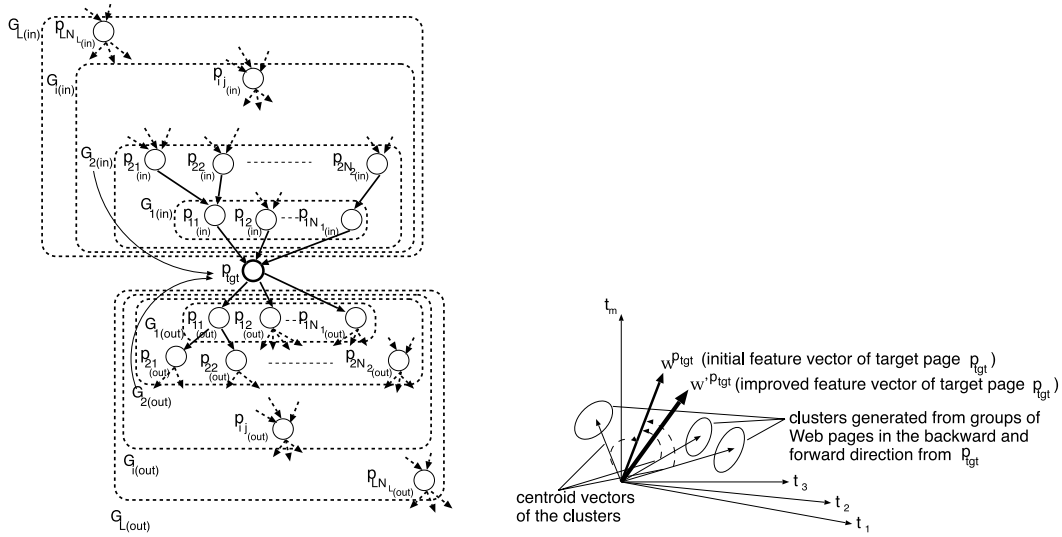

図 3　手法 III による特徴ベクトルの修正法
Fig. 3　The improvement of a feature vector as performed by Method III [(a) in the Web space, (b) in the vector space].

番目までの階層におけるすべての Web ページの特徴ベクトルを，$\boldsymbol{w}^{p_{tgt}}$ に反映することによって，$\boldsymbol{w}'^{p_{tgt}}$ が作成されることを示している．図 1 (a) において，$p_{ij_{(in)}}$ と $p_{ij_{(out)}}$ は，それぞれ p_{tgt} から逆方向，順方向に i 番目の階層における j 番目のページに対応する．更に，図 1 (b) は，p_{tgt} の in-link ページ，out-link ページのそれぞれの特徴ベクトルを，初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ が作成されることを示している．本手法において，$\boldsymbol{w}'^{p_{tgt}}$ の各要素 $w_{t_k}^{\prime p_{tgt}}$ は，予備実験によって式 (5) を用いる．

$$
\begin{aligned}
w_{t_k}^{\prime p_{tgt}} &= w_{t_k}^{p_{tgt}} \\
&+ \frac{1}{Dim}\left(\sum_{i=1}^{L_{(in)}}\sum_{j=1}^{N_{i_{(in)}}} \frac{w_{t_k}^{p_{ij_{(in)}}}}{N_{i_{(in)}} \cdot dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij_{(in)}}})}\right) \\
&+ \frac{1}{Dim}\left(\sum_{i=1}^{L_{(out)}}\sum_{j=1}^{N_{i_{(out)}}} \frac{w_{t_k}^{p_{ij_{(out)}}}}{N_{i_{(out)}} \cdot dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij_{(out)}}})}\right)
\end{aligned}
\tag{5}
$$

式 (5) は，式 (2) の TF-IDF 法によって計算される初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ における単語 t_k の重み $w_{t_k}^{p_{tgt}}$ に対して，対象ページ p_{tgt} の in-link ページ $p_{ij_{(in)}}$ における単語 t_k の重み $w_{t_k}^{p_{ij_{(in)}}}$ とベクトル空間における，$\boldsymbol{w}^{p_{tgt}}$ と $\boldsymbol{w}^{p_{ij_{(in)}}}$ 間の距離 $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij_{(in)}}})$ の逆数との積が，p_{tgt} から $L_{(in)}$ 階層までのすべての in-link ページに関して加えられることを示している．同様に，対象ページ p_{tgt} の out-link ページ $p_{ij_{(out)}}$ における単語 t_k の重み $w_{t_k}^{p_{ij_{(out)}}}$ とベクトル空間における，$\boldsymbol{w}^{p_{tgt}}$ と $\boldsymbol{w}^{p_{ij_{(out)}}}$ 間の距離 $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij_{(out)}}})$ の逆数との積が，p_{tgt} から $L_{(out)}$ 階層までのすべての out-link ページに関して加えられることを示している．ここで，Dim は索引語の数を表し，式 (5) の第 2 項と第 3 項がもとの索引語の重み $w_{t_k}^{p_{tgt}}$ の値よりも支配的になることを防ぐ目的で導入している．

3.2 手 法 II

本手法では，まず，対象ページ p_{tgt} から，逆方向に $L_{(in)}$ までの各階層において Web ページ群 $G_{i_{(in)}}$ を，また，順方向に $L_{(out)}$ までの各階層において Web ページ群 $G_{i_{(out)}}$ を構成する．次に，$G_{i_{(in)}}$ と $G_{i_{(out)}}$ から生成されるクラスタの重心ベクトルを，初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を作成する．本手法は，p_{tgt} から順方向，逆方向の各階層における Web ページは，その各階層においていくつかの話題に分類されるという考えに基づく．更に，ベクトル空間における $\boldsymbol{w}^{p_{tgt}}$ とクラスタの重心ベクトル間の距離を，初期特徴ベクトル

$\boldsymbol{w}^{p_{tgt}}$ の各要素に反映する．すなわち，まず，式 (6), (7) によって定義される Web ページ群 $G_{i_{(in)}}$，$G_{i_{(out)}}$ を構成する．

$$G_{i_{(in)}} = \{p_{i1_{(in)}}, p_{i2_{(in)}}, \cdots, p_{iN_{i_{(in)}}}\} \quad (6)$$

$$G_{i_{(out)}} = \{p_{i1_{(out)}}, p_{i2_{(out)}}, \cdots, p_{iN_{i_{(out)}}}\} \quad (7)$$
$$(i = 1, 2, \cdots, L)$$

次に，K-平均アルゴリズム [18] を用いて，この $G_{i_{(in)}}$，$G_{i_{(out)}}$ において，K 個のクラスタを作成する．重心ベクトル $\boldsymbol{w}^{g_{ic_{(in)}}}$，$\boldsymbol{w}^{g_{ic_{(out)}}}(c = 1, 2, \cdots, K)$ が，$G_{i_{(in)}}$ と $G_{i_{(out)}}$ においてそれぞれ作成され，これらの各重心ベクトル $\boldsymbol{w}^{g_{ic_{(in)}}}, \boldsymbol{w}^{g_{ic_{(out)}}}$ と初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ 間の距離を $\boldsymbol{w}^{p_{tgt}}$ の各要素に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を作成する．

例えば，図 2 (a) は，p_{tgt} から，逆方向，順方向に 2 番目までの各階層において，Web ページ群 $G_{1_{(in)}}, G_{2_{(in)}}, G_{1_{(out)}}, G_{2_{(out)}}$ を作成し，これらの各 Web ページ群において作成された各クラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を作成することを示している．また，図 2 (b) は，それぞれのクラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ が作成されることを示している．本手法において，$\boldsymbol{w}'^{p_{tgt}}$ の各要素 $w'^{p_{tgt}}_{t_k}$ は，予備実験によって式 (8) を用いる．

$$w'^{p_{tgt}}_{t_k} = w^{p_{tgt}}_{t_k} + \frac{1}{Dim}\left(\sum_{i=1}^{L_{(in)}}\sum_{c=1}^{K}\frac{w^{g_{ic_{(in)}}}_{t_k}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(in)}}})}\right) + \frac{1}{Dim}\left(\sum_{i=1}^{L_{(out)}}\sum_{c=1}^{K}\frac{w^{g_{ic_{(out)}}}_{t_k}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(out)}}})}\right) \quad (8)$$

式 (8) は，式 (2) の TF-IDF 法によって計算される初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ における単語 t_k の重み $w^{p_{tgt}}_{t_k}$ に対して，Web ページ群 $G_{i_{(in)}}(i = 1, 2, \cdots, L_{(in)})$ から作成されるクラスタ c の重心ベクトル $\boldsymbol{w}^{g_{ic_{(in)}}}$ の単語 t_k の重み $w^{g_{ic_{(in)}}}_{t_k}$ と，ベクトル空間における，$\boldsymbol{w}^{p_{tgt}}$ と $\boldsymbol{w}^{g_{ic_{(in)}}}$ 間の距離 $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(in)}}})$ の逆数との積が，p_{tgt} から逆方向に $L_{(in)}$ までの各階層において生成されるすべてのクラスタの重心ベクトルに関して，同様に，Web ページ群 $G_{i_{(out)}}(i = 1, 2, \cdots, L_{(out)})$ から作成されるクラスタ c の重心ベクトル $\boldsymbol{w}^{g_{ic_{(out)}}}$ の単語 t_k の重み $w^{g_{ic_{(out)}}}_{t_k}$ と，ベクトル空間における，$\boldsymbol{w}^{p_{tgt}}$ と $\boldsymbol{w}^{g_{ic_{(out)}}}$ 間の距離 $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(out)}}})$ の逆数との積が，順方向に $L_{(out)}$ までの各階層において生成されるすべてのクラスタの重心ベクトルに関して加えられることを示す．また，Dim は索引語の数を表し，式 (8) の第 2 項と第 3 項がもとの索引語の重み $w^{p_{tgt}}_{t_k}$ の値よりも支配的になることを防ぐ目的で導入している．

3.3 手 法 III

本手法は，対象ページ p_{tgt} から，逆方向に $L_{(in)}$ まで，順方向に $L_{(out)}$ までの階層に存在する Web ページはいくつかの話題から構成されるという考えに基づく．この考えに従って，p_{tgt} から，逆方向に $L_{(in)}$ まで，順方向に $L_{(out)}$ までの階層に存在するすべての Web ページの集合に対してクラスタリングを行い，生成されたクラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を作成する．更に，ベクトル空間における $\boldsymbol{w}^{p_{tgt}}$ とクラスタの重心ベクトル間の距離を，$\boldsymbol{w}^{p_{tgt}}$ の各要素に反映する．すなわち，まず，式 (9), (10) によって定義される Web ページ群 $G_{i_{(in)}}, G_{i_{(out)}}$ を作成する．

$$\begin{aligned} G_{i_{(in)}} = \{&p_{11_{(in)}}, p_{12_{(in)}}, \cdots, p_{1N_{1_{(in)}}}, \\ &p_{21_{(in)}}, p_{22_{(in)}}, \cdots, p_{2N_{2_{(in)}}}, \\ &p_{i1_{(in)}}, p_{i2_{(in)}}, \cdots, p_{iN_{i_{(in)}}}\} \end{aligned} \quad (9)$$

$$\begin{aligned} G_{i_{(out)}} = \{&p_{11_{(out)}}, p_{12_{(out)}}, \cdots, p_{1N_{1_{(out)}}}, \\ &p_{21_{(out)}}, p_{22_{(out)}}, \cdots, p_{2N_{2_{(out)}}}, \\ &p_{i1_{(out)}}, p_{i2_{(out)}}, \cdots, p_{iN_{i_{(out)}}}\} \\ &(i = 1, 2, \cdots, L) \end{aligned} \quad (10)$$

次に，この $G_{i_{(in)}}$ と $G_{i_{(out)}}$ において，K-平均アルゴリズムを用いて K 個のクラスタを作成する．重心ベクトル $\boldsymbol{w}^{g_{c_{(in)}}}$ と $\boldsymbol{w}^{g_{c_{(out)}}}(c = 1, 2, \cdots, K)$ が，それぞれ $G_{i_{(in)}}$ と $G_{i_{(out)}}$ において作成され，これらの各重心ベクトル $\boldsymbol{w}^{g_{c_{(in)}}}$，$\boldsymbol{w}^{g_{c_{(out)}}}(c = 1, 2, \cdots, K)$ と初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ 間の距離を $\boldsymbol{w}^{p_{tgt}}$ の各要素に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を作成する．

例えば，図 3 (a) は，p_{tgt} から逆方向，順方向に 2 番目までの階層において，Web ページ群 $G_{2_{(in)}}$，$G_{2_{(out)}}$ を作成し，これらの Web ページ群において作成されたクラスタの重心ベクトルを，初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$

に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ が作成されることを示したものである．また，図 3 (b) は，それぞれのクラスタの重心ベクトルを，初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映することによって，改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ が作成されることを示している．本手法において，$\boldsymbol{w}'^{p_{tgt}}$ の各要素 $w'^{p_{tgt}}_{t_k}$ は，予備実験によって式 (11) を用いる．

$$w'^{p_{tgt}}_{t_k} = w^{p_{tgt}}_{t_k} + \frac{1}{Dim}\left(\sum_{c=1}^{K}\frac{w^{g_{c_{(in)}}}_{t_k}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c_{(in)}}})}\right) + \frac{1}{Dim}\left(\sum_{c=1}^{K}\frac{w^{g_{c_{(out)}}}_{t_k}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c_{(out)}}})}\right) \tag{11}$$

式 (11) は，式 (2) の TF-IDF 法によって計算される初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ における単語 t_k の重み $w^{p_{tgt}}_{t_k}$ に対して，Web ページ群 $G_{i_{(in)}}$ から作成されるクラスタ c の重心ベクトル $\boldsymbol{w}^{g_{c_{(in)}}}$ における単語 t_k の重み $w^{g_{c_{(in)}}}_{t_k}$ と，ベクトル空間における $\boldsymbol{w}^{p_{tgt}}$ と $\boldsymbol{w}^{g_{c_{(in)}}}$ 間の距離 $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c_{(in)}}})$ の逆数との積が，同様に，Web ページ群 $G_{i_{(out)}}$ から作成されるクラスタ c の重心ベクトル $\boldsymbol{w}^{g_{c_{(out)}}}$ における単語 t_k の重み $w^{g_{c_{(out)}}}_{t_k}$ と，ベクトル空間における $\boldsymbol{w}^{p_{tgt}}$ と $\boldsymbol{w}^{g_{c_{(out)}}}$ 間の距離 $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c_{(out)}}})$ の逆数との積が，クラスタの数 K に関して加えられることを示す．手法 I, II で述べたように，もとの索引語の重み $w^{p_{tgt}}_{t_k}$ と比較して，式 (11) の第 2 項と第 3 項の値が支配的になることを防ぐために，索引語数を Dim を導入している．

4. 評価実験

4.1 実験環境

3. で述べた三つの手法は，ワークステーション（CPU: UltraSparc-II 480 MHz×4, Memory: 2 GByte, OS: Solaris8）上に Perl によって実装され，TREC（Text REtrieval Conference）WT10 g テストコレクション [19] を用いて，検索精度を確かめるための実験を行った．このテストコレクションは，オーストラリアの CSIRO（Commonwealth Scientific & Industrial Research Organization）(注2)が，1997 年に Internet Archive(注3)によって収集された Web ページの一部を利用して作成したものであり，約 169 万の Web ページ（10 GBytes）と，テストコレクション中の各 Web ページに対する in-link ページ，out-link ページの情報，検索課題集合，適合文書集合などから構成される． ここで，検索課題集合は Excite(注4)の検索語のログに基づいて 50 個の課題が作られ，図 4 のように，検索要求を記述した title フィールド，検索語を満たす文章を記述した description フィールド，適合文書の判断の基準を記述した narrative フィールドから構成される．

我々は，このテストコレクション中の Web ページに対して，不要語リスト(注5)に基づいて，不要語を取り除き，Porter Stemmer [20](注6)を用いて語幹処理を行った．また，図 4 を一例とする 50 個の検索課題集合それぞれに対して，title フィールドに含まれる単語を用いて，式 (12) で表される検索語ベクトル $\boldsymbol{Q}$ を作成した．

$$\boldsymbol{Q} = (q_{t_1}, q_{t_2}, \cdots, q_{t_m}) \tag{12}$$

式 (12) において，$t_k (k = 1, 2, \cdots, m)$ は索引語を表し，各要素 q_{t_k} は，式 (13) のように定義される．

```
<num> Number: 462
<title> real estate and new jersey

<desc> Description:
Find documents that contain residential
real estate listings within New Jersey.

<narr> Narrative:
Documents containing realtor data such
as point of contact, address, web site
or email address are considered as a
real estate listing are are relevant.
Listings of commercial real estate for
sale or auction are not relevant.
```

図 4 WT10g テストコレクションにおける topic の記述例
Fig. 4 An example of topic descriptions in WT10g test collection.

(注2)：http://www.csiro.au/
(注3)：http://www.archive.org/
(注4)：http://www.excite.com/
(注5)：ftp://ftp.cs.cornell.edu/pub/smart/english.stop
(注6)：http://www.tartarus.org/%7Emartin/PorterStemmer/

$$q_{t_k} = \left(0.5 + \frac{0.5 \cdot Qf(t_k)}{\sum_{k=1}^{m} Qf(t_k)}\right) \cdot \log \frac{N_{web}}{df(t_k)} \quad (13)$$
$$(k = 1, 2, \cdots, m)$$

ここで，$Qf(t_k), N_{web}, df(t_k)$ は，それぞれ，検索語ベクトル $\boldsymbol{Q}$ の中に含まれている索引語 t_k の数，Web ページの総数，単語 t_k が出現する Web ページ数を表す．文献 [21] で報告されているように，式 (13) は検索精度を最も良くする検索語ベクトルの成分である．更に，**3.** で述べた改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ と検索語ベクトル $\boldsymbol{Q}$ との間の類似度 $sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$ を式 (14) によって計算する．

$$sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q}) = \frac{\boldsymbol{w}'^{p_{tgt}} \cdot \boldsymbol{Q}}{|\boldsymbol{w}'^{p_{tgt}}| \cdot |\boldsymbol{Q}|} \quad (14)$$

4. 3 での実験結果の評価は，この $sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$ の値に基づいて，平均適合率による検索精度を用いて行った．この平均適合率 $\bar{P}$ は，検索語 q_i に対する適合文書数を $R_{q_i}(i = 1, 2, \cdots, N_q)$，上位 R_{q_i} までにシステムが出力した適合文書数を Rel_{q_i} とすれば，式 (15) のように表される．

$$\bar{P} = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{Rel_{q_i}}{R_{q_i}} \quad (15)$$

本論文では 50 個の検索課題集合 ($N_q = 50$) について，$sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$ の値に基づいて出力された上位 1000 件の Web ページに対して，式 (15) を適用して評価を行った．

4.2 実験手法

3. で述べた三つの手法，手法 I，II，III に関して，以下のように対象ページ p_{tgt} に対する改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を作成し，その検索精度を比較する実験を行った．

[手法 I]

(a) p_{tgt} から逆方向に $L_{(in)}$ までの階層におけるすべての Web ページの内容を初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合，

(b) p_{tgt} から順方向に $L_{(out)}$ までの階層におけるすべての Web ページの内容を初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合，

(c) p_{tgt} から逆方向に $L_{(in)}$ まで，順方向に $L_{(out)}$ までの階層におけるすべての Web ページの内容を初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合．

[手法 II]

(a) p_{tgt} から逆方向に $L_{(in)}$ までの各階層において作られる Web ページ群から生成されるクラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合，

(b) p_{tgt} から順方向に $L_{(out)}$ までの各階層において作られる Web ページ群から生成されるクラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合，

(c) p_{tgt} から逆方向に $L_{(in)}$ まで，順方向に $L_{(out)}$ までの各階層において作られる Web ページ群から生成されるクラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合．

[手法 III]

(a) p_{tgt} から逆方向に $L_{(in)}$ までの階層におけるすべての Web ページから構成されるグループによって生成されるクラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合，

(b) p_{tgt} から順方向に $L_{(out)}$ までの階層におけるすべての Web ページから構成されるグループによって生成されるクラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合，

(c) p_{tgt} から逆方向に $L_{(in)}$ まで，順方向に $L_{(out)}$ までの階層におけるすべての Web ページから構成されるグループによって生成されるクラスタの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合．

なお，手法 I，II，III とも，(a) の場合には $1 \leqq L_{(in)} \leqq 5$，(b) の場合には $1 \leqq L_{(out)} \leqq 5$，(c) の場合には $1 \leqq L_{(in)}, L_{(out)} \leqq 5$ と変動させ，更に手法 II，III の場合には，クラスタ数 K を $1 \leqq K \leqq 5$ と変動させて改良特徴ベクトル $\boldsymbol{w}'^{p_{tgt}}$ を作成し，その検索精度を比較する実験を行った．

4.3 実験結果及び考察

図 5 は，手法 I(a)，(b)，(c) の各場合において，それぞれ $L_{(in)}$, $L_{(out)}$, $[L_{(in)}, L_{(out)}]$ の値を，変動させた場合の平均適合率の変化を，図 7〜図 9 は，それぞれ手法 II(a)，(b)，(c) の各場合において，クラスタ数 K の値を変動させた場合の平均適合率の変化を，図 10〜図 12 は，それぞれ手法 III(a)，(b)，(c) の各場合において，クラスタ数 K の値を変動させた場合の平均適合率の変化を示したグラフである．なお，提案手法との比較を容易にするため，各グラフに TF-IDF 法を用いた場合の平均適合率も示した．この TF-IDF 法による平均適合率は，$L_{(in)}$，$L_{(out)}$ の値，クラスタ数 K の値に依存しないため，これらの各値に対して

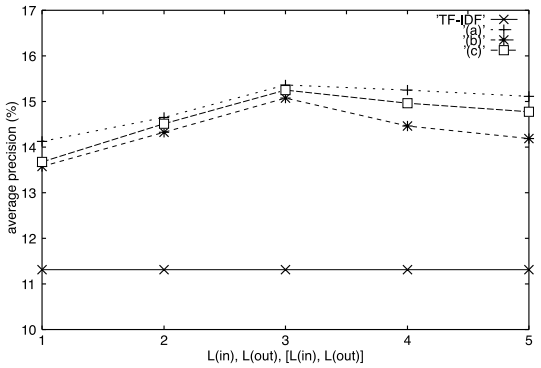

図 5　手法 I による平均適合率
Fig. 5　Average precision based on Method I.

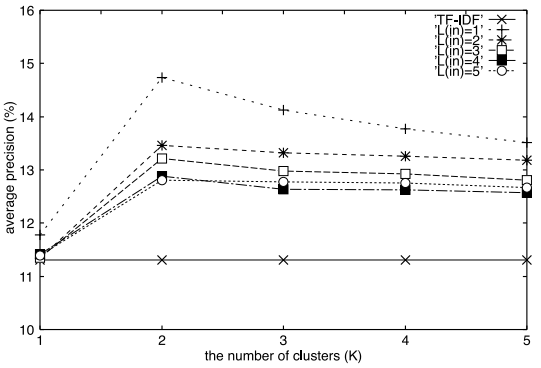

図 7　手法 II(a) による平均適合率
Fig. 7　Average precision based on Method II(a).

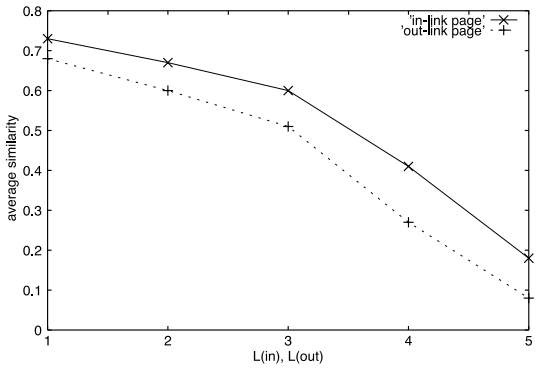

図 6　平均類似度の分布
Fig. 6　Distribution of average similarity.

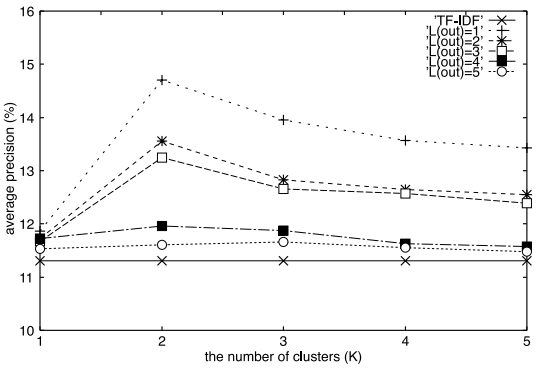

図 8　手法 II(b) による平均適合率
Fig. 8　Average precision based on Method II(b).

一定の値をとるものとして表示している．

各手法においては，次のような傾向が観察されている [22]．手法 I では，図 5 によれば，対象ページ p_{tgt} から，逆方向に 3 階層 ($L_{(in)} = 3$)，順方向に 3 階層 ($L_{(out)} = 3$) までに存在する Web ページは類似性が高く，p_{tgt} の特徴をより正確に表現することに寄与するが，p_{tgt} から逆方向に 4 階層以降 ($L_{(in)} \geqq 4$)，順方向に 4 階層以降 ($L_{(out)} \geqq 4$) のページは，対象ページとの類似性が小さいため，p_{tgt} の特徴をより正確に表現する効果が現れなかった．ここで，図 6 に，WT10g テストコレクションにおける各 Web ページと，そのページから逆方向，順方向の各階層における Web ページとの平均類似度の分布を示す．逆方向に 3 階層 ($L_{(in)} = 3$)，順方向に 3 階層 ($L_{(out)} = 3$) までに存在する Web ページとの平均類似度は比較的高いが，逆方向に 4 階層以降 ($L_{(in)} \geqq 4$)，順方向に 4 階層以降 ($L_{(out)} \geqq 4$) に存在する Web ページとの平均類似度は，低くなっている．この事実が，平均適合率にも影響を与えていると考えられる．また，図 5 によれば，$L_{(in)} \geqq 4$，$L_{(out)} \geqq 4$ の場合には，平均適合率が低下する傾向がうかがえる．したがって，$1 \leqq L_{(in)} \leqq 5$，$1 \leqq L_{(out)} \leqq 5$ の範囲で，平均適合率を調べたことは，実験として妥当であったと考えられる．

手法 II では，図 7〜図 9 によれば，対象ページ p_{tgt} から離れるに従って，つまり，$L_{(in)}$，$L_{(out)}$ の値が大きくなるにつれ，提案手法によって得られた各平均適合率のグラフと TF-IDF 法によって得られた平均適合率のグラフの間隔が小さくなっている．すなわち，TF-IDF 法と比較して検索精度の改善の割合が小さくなっていることから，p_{tgt} に対して，逆方向に 1 階層 ($L_{(in)} = 1$) まで，順方向に 1 階層 ($L_{(out)} = 1$) までに存在する Web ページをグループ化して作成した重心ベクトルは，p_{tgt} との内容的な関連性が強いが，p_{tgt} から離れるにつれ，その各階層をグループ化して

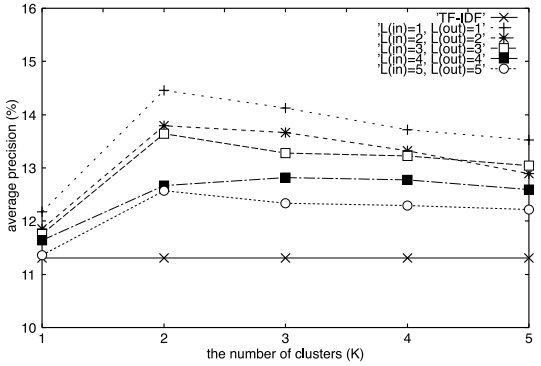

図 9　手法 II(c) による平均適合率
Fig. 9　Average precision based on Method II(c).

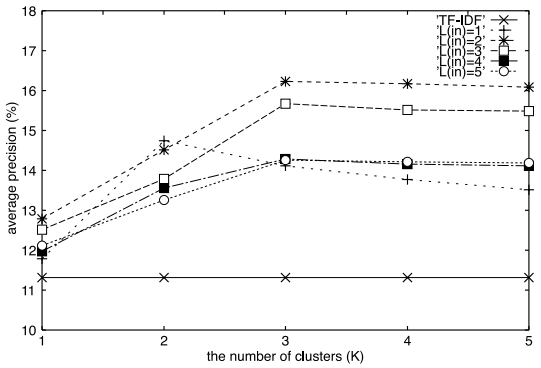

図 10　手法 III(a) による平均適合率
Fig. 10　Average precision based on Method III(a).

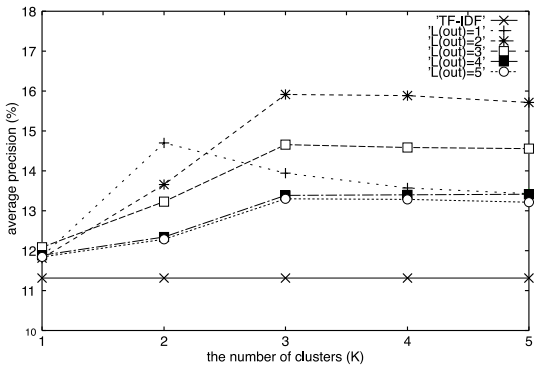

図 11　手法 III(b) による平均適合率
Fig. 11　Average precision based on Method III(b).

作成した重心ベクトルと対象ページ p_{tgt} との内容的な関連性は弱まるものと考えられる．

手法 III では，図 10〜図 12 によれば，p_{tgt} から

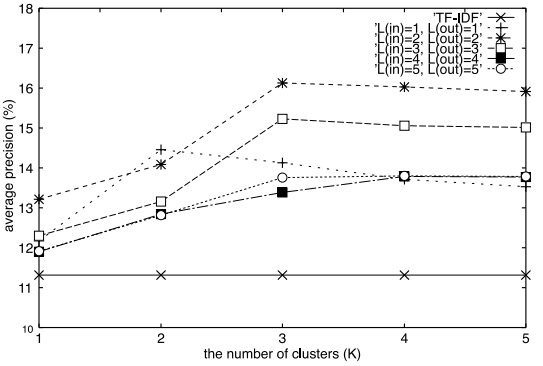

図 12　手法 III(c) による平均適合率
Fig. 12　Average precision based on Method III(c).

逆方向に 2 階層 ($L_{(in)} = 2$) まで，順方向に 2 階層 ($L_{(out)} = 2$) までに存在する Web ページをグループ化し，そのグループ内で生成される三つのクラスタ ($K = 3$) の重心ベクトルを用いて対象ページ p_{tgt} の特徴ベクトルを生成した場合に，最適な検索精度が得られている．したがって，p_{tgt} から逆方向に 2 階層 ($L_{(in)} = 2$) まで，順方向に 2 階層 ($L_{(out)} = 2$) までに存在する Web ページの話題は，クラスタ数が 3 ($K = 3$) の場合に最適な検索精度が得られていることから，およそ三つ程度で構成される傾向にあるものと考えられる．

また，手法 II，III におけるクラスタ数と平均適合率には，以下のような関係があることが分かる．まず，手法 II では，図 7〜図 9 によれば，$K \geqq 3$ の場合には平均適合率が低下する傾向が見られる．一方，手法 III では，$L_{(in)} = 1$，$L_{(out)} = 1$ の場合には，式 (8)，式 (11) から分かるように，手法 II と同じ結果となるが，$L_{(in)} \geqq 2$，$L_{(out)} \geqq 2$ の場合には，$K \geqq 4$ となると，平均適合率が緩やかに低下する傾向が見られる．したがって，手法 II，III において，クラスタ数 K に関して $1 \leqq K \leqq 5$ の範囲で平均適合率を調べたことは，実験として妥当であったと考えられる．

以上の結果を要約した表 1 は，Web ページの特徴ベクトルを TF-IDF 法を用いて生成した場合の平均適合率と，**3.** で述べた三つの手法それぞれにおいて，最適な検索精度が得られた場合の結果を示している．手法 I では，対象ページ p_{tgt} から逆方向に 3 階層 ($L_{(in)} = 3$) までにおけるすべての Web ページの内容を初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合に，手法 II では，対象ページ p_{tgt} から，逆方向に 1 階層 ($L_{(in)} = 1$) までの Web ページから作られる Web

表 1　手法 I, II, III を用いて得られた最適な検索精度の比較
Table 1　Comparison of the best search accuracy obtained using Method I, II, and III.

	% average precision	% improvement
TF-IDF	11.31	−
手法 I ($L_{(in)}=3$)	15.30	+3.99
手法 II ($L_{(in)}=1, K=2$)	14.74	+3.43
手法 III ($L_{(in)}=2, K=3$)	**16.23**	**+4.92**

ページ群から二つのクラスタ ($K=2$) を作成し，それらの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合に，また，手法 III では，対象ページ p_{tgt} から，逆方向に 2 階層 ($L_{(in)}=2$) までにおけるすべての Web ページから作られる Web ページ群から三つのクラスタ ($K=3$) を作成し，それらの重心ベクトルを初期特徴ベクトル $\boldsymbol{w}^{p_{tgt}}$ に反映する場合に，TF-IDF 法と比較して検索精度が最も良くなっている．また，表 1 によれば，手法 I，II，III のいずれの場合においても，(a) の実験，すなわち，対象ページ p_{tgt} の in-link ページを利用することで，最適な検索精度が得られている．これは，ある Web ページから逆方向へたどる場合には，そのページと内容の類似した Web ページに到達しやすいが，順方向へたどる場合には，対象ページの内容と異なる様々なページに到達するためであると考えられる．すなわち，Web ページの特性の一つとして，ある対象ページに対する in-link ページには，その対象ページの内容に適合するページが多く存在するという性質が挙げられると考えられる．**2.2** で述べたように，HITS では多くの Web ページにリンクしているページを「ハブ」と定義し，これを in-link ページと考えて Web ページの良質さを表す「オーソリティ」を定義している．また，in-link ページの重要性に着目して，あるページから逆方向に巡回するためのツールも開発されている [23]．表 1 の結果は，本研究においても in-link ページの有用性が示された一例であると考えることができる．

4.4　検索精度に関する考察

文献 [19] によれば，TREC-9 Web Track において，リンク情報を利用した手法によって得られた各参加チームが用いた手法の概略と平均適合率は表 2 のとおりである．なお，手法の詳細については，表中に示した文献を参照されたい．表 2 によれば，文献 [24] による HITS を基本とした手法で 4.88%，文献 [25] による HITS を改良した手法で，5.91%，6.37%という平均適合率しか得られていない．一方，表 1 によれば，我々の提案手法によって得られた結果のうち，最適な平均適合率は 16.23%であり，これは，文献 [26] や文献 [24] で得られた結果に相当し，Web ページの特徴付け手法として十分有効性があり，信頼できる結果であると考えられる．

表 2　リンク情報を用いた WT10g の平均適合率
Table 2　Average precision of WT10g using link information.

Group	outline of each approach	% average precision
文献 [27]	anchor text	20.00
	anchor text + long query	18.38
文献 [26]	content-link	16.31
	4gram content-link	17.94
文献 [24]	cocitation top 10	16.30
	cocitation top 50	13.37
	HITS	4.88
文献 [28]	Okapi + probabilistic augmentation	17.36
文献 [29]	anchor text	12.50
	variant of anchor text	12.88
文献 [30]	back link frequency	10.62
文献 [25]	modified HITS	5.91
	modified HITS with weighted links	6.37

5. むすび

本論文では，Web ページの内容をより正確に表現するために，ハイパリンクで結ばれた隣接ページの特徴ベクトルを用いて Web ページ向けに TF-IDF 法を改良する手法を提案した．我々の手法は，これまで行われていなかったハイパリンクで結ばれた隣接する Web ページの内容を，Web ページの特徴ベクトル生成の際に反映する点に新規性がある．もちろん，これまでに提案されている HITS や PageRank と併用すれば，更なる検索精度の向上が期待できる．

本論文では，Web ページのより正確な特徴ベクトルを生成することを目的として，Web のハイパリンク構造に注目した．しかしながら，利用者の実際の検索要求を満足するためには，巨大な WWW 空間から適切な Web ページを発見することが重要である．したがって，より利用者個人に適合した情報を提供する技術についても，今後研究を進めていく予定である．

文　献

[1] R. Baeza-Yates, F. Saint-Jean, and C. Castillo. “Web structure, dynamics and page quality,” Proc. 9th International Symposium on String Processing and Information Retrieval (SPIRE 2002), pp.117–130, 2002.
[2] A. Broder and P. Raghavan, “Combining text- and link-based information retrieval on the web,” SI-

GIR'01 Pre-Conference Tutorials, Sept. 2001.
[3] L. Page, "The PageRank citation ranking: Bringing order to the web," http://google.stanford.edu/~backrub/pageranksub.ps, 1998.
[4] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," Proc. 7th International World Wide Web Conference (WWW7), pp.107–117, 1998.
[5] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," Proc. 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 1998), pp.668–677, 1998.
[6] IBM Almaden Research Center, Clever Searching, http://www.almaden.ibm.com/cs/k53/clever.html
[7] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1983.
[8] K. Tajima, Y. Mizuuchi, M. Kitagawa, and K. Tanaka, "Cut as a querying unit for WWW, netnews, e-mail," Proc. 9th ACM Conference on Hypertext and Hypermedia (HyperText '98), pp.235–244, 1998.
[9] K. Tajima, K. Hatano, T. Matsukura, R. Sano, and K. Tanaka, "Discovery and retrieval of logical information units in web," Proc. 1999 ACM Digital Libraries Workshop on Organizing Web Space (WOWS '99), pp.13–23, 1999.
[10] W-S. Li, K. Selçuk Candan, Q. Vu, and D. Agrawal, "Retrieving and organizing web pages by "Information Unit"," Proc. 10th International World Wide Web Conference (WWW10), pp.230–244, 2001.
[11] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," Proc. 7th International World Wide Web Conference (WWW7), pp.65–74, 1998.
[12] K. Bharat and M.R. Henzinger, "Improved algorithms for topic distillation in a hyperlinked environment," Proc. 21st Annual International ACM SIGIR Conference (SIGIR '98), pp.104–111, 1998.
[13] L. Li, Y. Shang, and W. Zhang, "Improvement of HITS-based algorithms on web documents," Proc. 11th International World Wide Web Conference (WWW2002), pp.527–535, 2002.
[14] S. Chakrabarti, "Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction," Proc. 10th International World Wide Web Conference (WWW10), pp.211–220, 2001.
[15] S. Chakrabarti, M. Joshi, and V. Tawde, "Enhanced topic distillation using text, markup tags, and hyperlinks," Proc. 23rd Annual International ACM SIGIR Conference (SIGIR 2001), pp.208–216, 2001.
[16] D. Rafiei and A.O. Mendelzon, "What is this page known for? Computing web page reputations," Proc. 9th International World Wide Web Conference (WWW9), pp.823–835, 2000.
[17] T.H. Haveliwala, "Topic-sensitive PageRank," Proc. 11th International World Wide Web Conference (WWW2002), pp.517–526, 2002.
[18] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symposium on Mathmatical Statistics and Probability, pp.281–297, 1967.
[19] D. Hawking, "Overview of the TREC-9 web track," NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9), pp.87–102, 2001.
[20] M.F. Porter, "An algorithm for suffix stripping," Program, vol.14, no.3, pp.130–137, 1988.
[21] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Inf. Process. & Manage., vol.24, no.5, pp.513–523, 1988.
[22] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura, "Refinement of TF-IDF schemes for web pages using their hyperlinked neighboring pages," Proc. 14th ACM Conference on HyperText and Hypermedia (HT '03), pp.198–207, 2003.
[23] S. Chakrabarti, D.A. Gibson, and K.S. McCurley, "Surfing the web backwards," Proc. 8th International World Wide Web Conference, pp.1679–1693, 1999.
[24] W. Kraaij and T. Westerveld, "TNO/UT at TREC-9: How different are Web documents?," http://trec.nist.gov/pubs/trec9/papers/tno-ut.pdf, 2000.
[25] F. Crivellari and M. Melucci, "Web document retrieval using passage retrieval, connectivity information, and automatic link weighting-TREC-9 report," http://trec.nist.gov/pubs/trec9/papers/jhuapl.pdf, 2000.
[26] C.L.A. Clake, G.V. Cormack, D.I.E. Kisman, and T.R. Lynam, "Question answering by passage selection (MultiText experiments for TREC-9)," http://trec.nist.gov/pubs/trec9/papers/mt9.pdf, 2000.
[27] S. Fujita, "Reflections on "Aboutness" TREC-9 evaluation experiments at Justsystem," http://trec.nist.gov/pubs/trec9/papers/jscbt9w_paper.pdf, 2000.
[28] J. Savoy and Y. Rasolofo, "Report on the TREC-9 experiment: Link-based retrieval and distributed collections," http://trec.nist.gov/pubs/trec9/papers/unine9.pdf, 2000.
[29] A. Singhal and M. Kaszkiel, "AT& T at TREC-9," http://trec.nist.gov/pubs/trec9/papers/att-trec9.pdf, 2000.
[30] P. McNamee, J. Mayfield, and C. Piatko, "The HAIRCUT system at TREC-9," http://trec.nist.gov/pubs/ trec9/papers/jhuapl.pdf, 2000.

（平成 15 年 6 月 3 日受付）

杉山　一成　（学生員）

1998 横浜国大･工･電子情報工学卒．2000 同大大学院電子情報工学専攻博士前期課程了．同年 KDD（現，KDDI）（株）入社，2001 退職．現在，奈良先端科学技術大学院大学情報科学研究科博士後期課程在学中．情報検索に関する研究に従事．情報処理学会，人工知能学会，IEEE, ACM, AAAI 各会員．

波多野賢治　（正員）

1995 神戸大･工･計測工学卒．1999 同大大学院自然科学研究科博士後期課程了．博士（工学）．同年奈良先端科学技術大学院大学情報科学研究科助手．XML データベース，情報検索に関する研究に従事．情報処理学会，ACM, IEEE Computer Society 各会員．

吉川　正俊　（正員）

1980 京大・工・情報工学卒．1985 同大大学院工学研究科博士後期課程了．工博．同年京都産業大学計算機科学研究所講師．同大学工学部情報通信工学科助教授，奈良先端科学技術大学院大学情報科学研究科助教授を経て，2002 名古屋大学情報連携基盤センター教授．XML データベース，多次元空間索引等の研究に従事．情報処理学会，ACM, IEEE Computer Society 各会員．

植村　俊亮　（正員：フェロー）

1964 京大・工・電子工学卒．1966 同大大学院工学研究科修士課程了．同年通産省工業技術院電気試験所（現，産業技術総合研究所）．1988 東京農工大学工学部数理情報工学科教授．1993 奈良先端科学技術大学院大学情報科学研究科教授．工博．1970〜1971 マサチューセッツ工科大学客員研究員．データベースシステム，自然言語処理，プログラム言語の研究に従事．IEEE Fellow，情報処理学会フェロー．ACM 等会員．