A1-3

リンク先ページの内容を反映させた Webページの特徴ベクトル改良法

杉山　一成† 　波多野賢治† 　吉川　正俊†,†† 　植村　俊亮†

† 奈良先端科学技術大学院大学 情報科学研究科
〒 630–0101 奈良県生駒市高山町 8916–5
†† 国立情報学研究所 ソフトウェア研究系
〒 101–8430 東京都千代田区一ツ橋 2–1–2

E-mail: †{kazuna-s,hatano,yosikawa,uemura}@is.aist-nara.ac.jp

あらまし　文書検索を行う際の文書の特徴ベクトル生成法として，tf-idf 法が広く用いられている．しかし，tf-idf 法は一つの文書を単位として開発された手法である．したがって，Web 文書のようにハイパーリンクで結ばれ，そのアンカー文字列とリンク先ページの内容が関連している場合には，tf-idf 法を改良して利用した方が，より Web 文書の特徴を表現できると考えられる．我々もこの方針にしたがって，リンク先の Web ページの内容も含めて Web ページの特徴ベクトルを生成する手法を提案したが，ベクトル空間内におけるリンク元ページとリンク先ページの関係，およびリンク先ページの扱いに対する最適性と収束性の条件については検討不足であった．そこで本稿では，これらの不足点を補うために，Web ページのリンク先ページの内容を含めた特徴ベクトル改良法を提案し，どのような条件下で最適な検索精度が得られるかを実験によって確認する．

キーワード　WWW 情報検索 ハイパーリンク tf-idf 法

A Method of Improving Feature Vectors for Web Pages Reflecting the Contents of Its linked Pages

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

† Graduate School of Information Science, Nara Institute of Science and Technology
8916–5, Takayama, Ikoma, Nara, 630–0101 Japan
†† Software Research Division, National Institute of Informatics
2–1–2, Hitotsubashi, Chiyoda, Tokyo 104–8430 Japan

E-mail: †{kazuna-s,hatano,yosikawa,uemura}@is.aist-nara.ac.jp

Abstract　In the field of document retrieval, tf-idf schemes are popular for generating the feature vectors of documents. These schemes are developed for characterizing the content of a document, however, it is expected that improved tf-idf schemes can characterize the content of Web documents better than conventional tf-idf schemes because Web pages are linked each other with hyperlinks and their anchors relate to the content of linked documents. Based on these properties, we proposed a method to characterize the content of a Web page including that of its linking Web page, but the relation between a Web page and its linking pages in vector space and the optimum and convergent condition for its linking pages are not discussed. In this paper, we propose a method of improving feature vectors of Web documents including the content of their linking pages, and verify the condition to obtain the best retrieval accuracy by experiment.

Key words　WWW, information retrieval, hyperlink, tf-idf schemes

1. はじめに

World Wide Web は情報を入手するための源として，多くの人々に利用されている．この Web 上の情報は，コンピュータネットワークの発達と共に氾濫し続けており，膨大な量の情報の中から有用な情報を見つけることは，利用者にとって困難になるばかりである．こうした状況の中で検索エンジンは，Web 上の情報を探す上で欠かせない手段となっており，Yahoo! [1] に代表されるディレクトリ型検索エンジンと Google [2] に代表されるロボット型の検索エンジン，さらには複数の検索エンジンの結果をまとめて表示するメタ検索エンジンが挙げられる．ディレクトリ型の検索エンジンでは，利用者が望んでいる情報があると考えられるディレクトリまでたどることは可能でも，そのような情報が見つからないことがしばしば起こる．また，そのディレクトリの作成や Web コンテンツの分類は他者の基準によるものであり，検索を行おうとする側にはわかりにくい．さらにディレクトリ型の検索エンジンでは，利用者はキーワード検索を行うこともできるが，そのキーワードに該当するディレクトリが提示されるだけで，利用者はさらにそのディレクトリをたどる必要があり，情報が見つけにくい状態は変わらない．一方，ロボット型検索エンジンの場合は，検索語を入力すればその語に該当するページを見つけることは可能であるが，膨大な数の検索結果の中から利用者が本当に望んでいる情報を選別していくのは多大な労力を必要とする．さらに，Web ページ固有の特徴といえるハイパーリンク構造を情報検索に活用したロボット型検索エンジンに PageRank [3] や HITS(Hypertext Induced Topic Search) [4] などのアルゴリズムを適用した研究も存在する．しかし，これらの研究ではハイパーリンク構造を利用して Web ページに対して重み付けが行われているだけである．したがって，リンクで結ばれた Web ページ間の内容を考慮して Web ページの索引を生成しているわけではない．その結果，内容的に関係のない膨大な数のページが検索結果として表示されることになる．

そこで本稿では，より Web ページの内容を考慮した検索を行うために，特徴ベクトルを作成しようとする Web ページ p からリンクされているページを解析し，それらのリンク先ページの内容も考慮したハイパーリンク構造による情報を活用することで，あらかじめ tf-idf 法によって作成した p の特徴ベクトルを改良する手法を提案する．ここで，p からリンクされているページに着目するのは，p と類似した内容が，その近傍のリンク先ページにも存在するため，それらのリンク先ページの内容を反映する必要があるという考えに基づく．本稿の先行研究 [5] では，ベクトル空間において p とリンク先ページの関係が考慮されていないこと，また，リンクの段数を考慮したパラメータを導入していたので，p から離れたリンク先に p と内容の類似したページがある場合には，そのページの内容が反映されにくいこと，リンク先ページを扱う手法と p からたどるリンク数についての検討が不十分であった

そこで本稿では，(1) ある Web ページに対して，その個々のリンク先ページの内容を反映させる，(2) リンク先ページから構成されるクラスタの重心ベクトルを反映させる，(3) リンクの段数ごとに存在する Web ページから構成されるクラスタの重心ベクトルを反映させる，の 3 種類の特徴ベクトル改良法を提案し，これらの改良した特徴ベクトルについて成分値の高い単語がキーワードと成り得るか，また，これらの特徴ベクトルを Web ページの索引とした場合に，既存の検索エンジンと比較してどれほどの検索精度が得られるかについての実験を行ない，最適な検索精度を得るための条件について検討した．

2. 関連研究

ハイパーリンク構造は Web の特徴の一つであり，Web の利用者はこの構造によって，膨大な Web 空間を容易に巡ることが可能になっている．したがって，Web 情報検索のコミュニティでは，この構造に着目して様々な研究が行われている．

Web 固有の特徴であるハイパーリンク構造に関する代表的な研究として、

（1） リンク構造で結びついた Web ページに「ページ群」という概念を導入した研究

（2） Web ページの質を判断するためにリンクが出入りする数を考慮した研究

などがある。(1) には，次に述べる研究が挙げられる．Tajima らは，文献 [6] において，Web 構造解析の結果である “cut” という概念を導入し，利用者が入力した問合せから HTML 文書の持つ意味構造と，それらを結びつけたリンク構造を考慮した検索結果を得ることができる検索システムの実装を行っている．また，文献 [7] においては、リンクの最小部分グラフとすべてのキーワードを含むページを発見し，部分グラフ内のキーワードの局所性に基づいて，それぞれの部分グラフのスコアを計算している．これらに

類似した研究として，Li ら [8] による “information unit” の概念があるが，これらの研究における検索結果は，全ての検索語を含んだページ群であり、これまでの研究のような Web ページ単位の検索ではない。しかし，一般に検索語が複数の場合にはそれらの意味的な結び付きは強いといえるため，複数の検索語すべてを含む Web ページを単位とした結果を出力する方が直感的に利用者が求めている情報であるといえる．つまり，これらの研究は検索結果の検索漏れを少なくする効果があると思われるが，逆に不要な情報を検索結果とする場合もあり，その点は大きな問題である．

また，(2) としては，次のような研究が挙げられる．Carriere ら [9] は，検索語に基づいて検索結果を順位付けするだけでなく，それぞれのページにおけるリンクの出入りの数を接続性と定義し，その値の降順に視覚化するシステムを作成している．しかし，視覚化する段階で有用な情報を隠してしまう可能性があり検索漏れの原因となりやすい．また Kleinberg [4] は，検索語に対する情報量の豊かなページである “authority” ページ，そして多くの “authority” ページへのリンクを持つ “hub” ページを見つける HITS アルゴリズムを開発した．このアルゴリズムは非常に優れており，他の多くの研究にも引用されているが，検索処理に十分な時間をかけてもよいという仮定のもとに開発された手法であり，実システムに適用するには慎重にパラメータを設定する必要がある．さらに，検索エンジン Google は，利用者が現在閲覧している Web ページからのリンクをたどる確率と全く無関係なページに遷移する確率を考慮した PageRank 法 [10] に基づいた検索エンジンである．しかし，あるページからリンク先ページに遷移する確率が，いずれのリンク先ページにおいても均一なものとして考えられており，リンク元・リンク先ページの内容によってこの確率を変化させなければ Web ページの内容を考慮することができない．また，特定の有名サイトやリンク集のようなページが検索結果の上位に順位付けされやすい．HITS や PageRank のアルゴリズムは，リンク構造に基づいて計算機によって Web ページの質を算出しているが，それが人間によって判断される質と一致しているか否かを確かめたものが Brian ら [11] による研究である．この文献では，リンク構造を利用して計算機によって求められた Web ページの質と，その話題の専門家によって判断された Web ページの質は，一致度が高いと述べられているが，実験で扱った話題は専門性の高いページであり，Web で扱っている多種多様の情報について有効か否かは不明である．これに対し Chakrabarti ら [12], [13] は，HTML タグによって決められた Web ページの部分構造とハイパーリンクとを統合利用して話題を抽出することで，リンク構造だけではなく Web ページの内容をも考慮した HITS を拡張したアルゴリズムの提案を行っている．しかし，文書の部分構造の選び方によって話題抽出の精度が変わりやすい点に問題がある．

我々の手法は，ある Web ページの内容と類似した内容を持つ Web ページが，その近傍のリンク先ページにも存在するという考えに基づき，そのリンク先ページの内容を反映させて，あらかじめ tf-idf 法によって作成した Web ページの特徴ベクトルを改良するというものである．Fujita [14] はリンク元ページのアンカーとして使われている単語を含めてリンク先ページの特徴ベクトルを作成しているが，参照元ページの内容が反映される程度は顕著でないと考えられる．しかし，本手法では，一つの Web ページに着目しただけではキーワード性の表れなかった単語を顕在化させる効果や，より Web ページの内容を考慮した特徴ベクトルの生成が期待される．

3. 提 案 手 法

本章では，我々が提案するハイパーリンク構造を利用した Web ページの特徴ベクトル改良法について説明する．ベクトル空間モデル [15], [16] を用い，あらかじめ tf-idf 法によって作成した Web ページの特徴ベクトルを，同じく tf-idf 法によって作成したリンク先ページの特徴ベクトルを用いて改良するのが基本的な方針である．ここでリンク先ページに着目するのは，リンク元ページで書かれている内容と類似した内容が，その近傍にリンクされたページにも存在するため，これらの内容を反映させる必要があるという考えに基づく．以下，特徴ベクトルの改良を行おうとするページを注目ページと呼び，p_f と表すことにする．また，注目ページ p_f から i 段目のリンク先にあるページは，$p_{i1}, \cdots, p_{iN_i}$ の N_i 個存在するものとする．この注目ページ p_f からたどるリンクの段数 i については複数の経路が存在し得るが，最短経路のリンク数を i の値と定義する．なお，tf-idf 法を用いてあらかじめ作成しておく Web ページ p_f の特徴ベクトル $\boldsymbol{w}^{p_f}$ を，

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

と表し，以下において $\boldsymbol{w}^{p_f}$ を初期特徴ベクトルと呼

ぶことにする．ここで，

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

であり，$tf(t_k, p_f)$ は Web ページ p_f における単語 t_k の頻度を，N_{web} は収集した 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．また，このとき Web ページの特徴ベクトルを構成する単語 t_k は，前処理として Zipf の法則 [17] を適用して低頻度語を取り除いておく．さらに，$\boldsymbol{w}^{p_f}$ を改良した特徴ベクトル $\boldsymbol{w}'^{p_f}$ を，

$$\boldsymbol{w}'^{p_f} = (w_{t_1}'^{p_f}, w_{t_2}'^{p_f}, \cdots, w_{t_m}'^{p_f})$$

と表し，以下において $\boldsymbol{w}'^{p_f}$ を改良特徴ベクトルと呼ぶことにする．我々は tf-idf 法によって作成した Web ページの初期特徴ベクトルを改良するために，次の 3 つの手法を提案する．

（1） 個々のリンク先ページを注目ページに反映させる手法 (以下，手法 1 とする)

（2） 注目ページから i 段目のリンク先までに存在するページから，Web ページ群を構成してクラスタリングを行い，それらの重心ベクトルを注目ページに反映させる手法 (以下，手法 2 とする)

（3） 注目ページからのリンクの段数ごとに Web ページ群を構成してクラスタリングを行い，それらの重心ベクトルを注目ページに反映させる手法 (以下，手法 3 とする)

以下，これらの 3 つの手法それぞれについて，詳細を説明する．

3.1 手法 1: 個々のリンク先ページの内容を注目ページに反映させる手法

注目ページ p_f に対して，そのリンク先ページ個々の内容を反映させる．本手法は，

- あるページからのリンク先ページには，内容の類似したページがある
- そのようなページは，すぐ近くのリンク先ページに存在する場合もあれば，遠くのリンク先ページに存在する場合もある

という考えに基づき，p_f とリンク先ページの距離を初期特徴ベクトルの各成分に反映させる．例えばリンクの段数 i に反比例する値を導入することも考えられるが，前述したように，p_f と内容の類似したページは必ずしも近くにあるとは限らない。したがって，注目ページ p_f の初期特徴ベクトルとそのリンク先ページの初期特徴ベクトルとの距離を反映させ，改良特徴ベクトルの作成を行う．

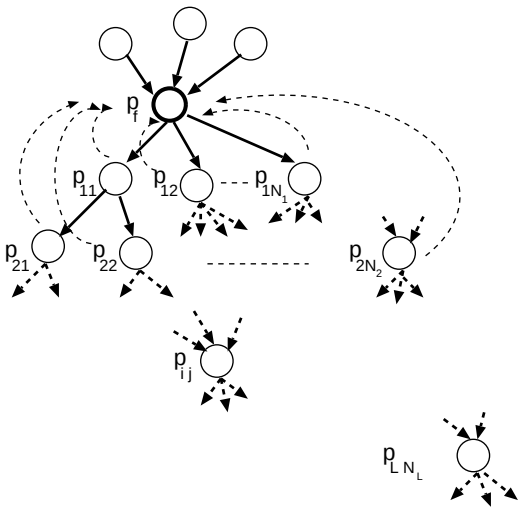

図 1　手法 1 による特徴ベクトルの改良法

例えば，図 1 の場合には，注目ページ p_f から 2 段目までのリンク先に存在するすべての Web ページの内容を，p_f に反映させることで改良特徴ベクトルを作成することを示している．本手法において，p_f の改良特徴ベクトルの各成分 $w_{t_k}'^{p_f}$ は，(3) 式のように表される．

$$w_{t_k}'^{p_f} = w_{t_k}^{p_f} + \sum_{i=1}^{L} \sum_{j=1}^{N_i} \frac{1}{dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{p_{ij}})} w_{t_k}^{p_{ij}} \tag{3}$$

(3) 式は，(2) 式の tf-idf 法によって求めた注目ページ p_f における単語 t_k の重み $w_{t_k}^{p_f}$ に，p_f からのリンク先ページ p_{ij} での単語 t_k の重み $w_{t_k}^{p_{ij}}$ と，ベクトル空間における $\boldsymbol{w}^{p_f}$ と $\boldsymbol{w}^{p_{ij}}$ の距離 $dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{p_{ij}})$ の逆数との積を，p_f から L 段目までのリンク先ページすべてについて加えていくことを示している．ただし，

$$dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{p_{ij}}) = \sqrt{\sum_{k=1}^{m} (w_{t_k}^{p_f} - w_{t_k}^{p_{ij}})}$$

である．

3.2 手法 2: リンク先ページから構成されるクラスタの重心ベクトルを注目ページに反映させる手法

L リンク先まで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，その重心ベクトルを，注目ページ p_f の特徴ベクトルに反映させることで，改良特徴ベクトルを作成する．本手法は，「リンク先ページをまとめて見渡した場合に，いくつかの話題に分けることができる」という考えに基づき，p_f とリンク先ページから構成されるクラスタの

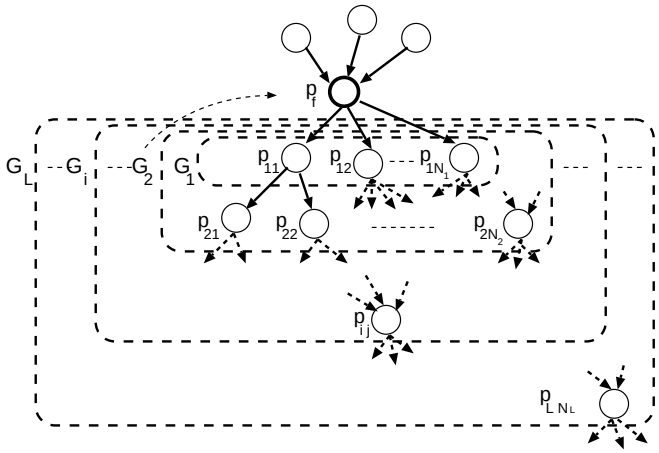

図 2　手法 2 によるベクトルの改良法

重心ベクトルとの距離を初期特徴ベクトルの各成分に反映させる．すなわち，p_f から L 段目のリンク先までに存在するすべての Web ページから，(4) 式で表される Web ページ群 G_i を構成する．

$$\begin{aligned} G_L = \{&p_{11}, p_{12}, \cdots, p_{1N_1}, \\ &p_{21}, p_{22}, \cdots, p_{2N_2}, \\ &p_{L1}, p_{L2}, \cdots, p_{LN_L}\} \end{aligned} \tag{4}$$

この G_L を $K-$ 平均アルゴリズム [18] によってクラスタリングし，K 個のクラスタを作成する．そのクラスタの各重心ベクトル $\boldsymbol{w}^{g_c}(c = 1, 2, \cdots, K)$ と注目 Web ページ p_f の初期特徴ベクトル $\boldsymbol{w}^{p_f}$ とのベクトル空間における距離を反映させ，改良特徴ベクトルを作成する．例えば，図 2 の場合には，2 段目のリンク先までに存在するすべての Web ページから，Web ページ群 G_2 を作成し，G_2 をクラスタリングすることで作成された K 個の重心ベクトルを $\boldsymbol{w}^{p_f}$ に反映させることで改良特徴ベクトルを作成することを示している．本手法において，p_f の改良特徴ベクトルの各成分 $w'^{p_f}_{t_k}$ は，(5) 式のように表される．(5) 式は，(2) 式の tf-idf 法によって求めた注目ページ p_f における単語 t_k の重み $w^{p_f}_{t_k}$ に，ページ群 G_L から作られるクラスタの重心ベクトルの成分 $w^{g_c}_{t_k}$ と，ベクトル空間における p_f と g_c の距離 $dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{g_c})$ の逆数との積を，生成するクラスタの数 K だけ加えていくことを示している．

$$w'^{p_f}_{t_k} = w^{p_f}_{t_k} + \sum_{c=1}^{K} \frac{1}{dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{g_c})} w^{g_c}_{t_k} \tag{5}$$

ただし，

$$dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{g_c}) = \sqrt{\sum_{k=1}^{m} (w^{p_f}_{t_k} - w^{g_c}_{t_k})}$$

である．

3.3　手法 3: リンクの段数ごとに存在する Web ページから構成されるクラスタの重心ベクトルを注目ページに反映させる手法

注目 Web ページ p_f からたどるリンクの段数ごとに存在する Web ページから Web ページ群 G_i を構成し，その各 G_i 内でクラスタリングを行う．そのクラスタの重心ベクトルを，対象 Web ページの特徴ベクトルに反映させることで，改良特徴ベクトルを作成する．本手法は，「たどったリンク先の Web ページは，いくつかの話題に分かれる」という考えに基づき，p_f から i 段目までのリンク先の段数ごとにクラスタリングを行い，$\boldsymbol{w}^{p_f}$ とその重心ベクトルの距離を初期特徴ベクトルの各成分に反映させる．すなわち，p_f からたどるリンクの段数 i ごとにその段階で存在する Web ページから，(6) 式で表される Web ページ群 G_i，

$$G_i = \{p_{i1}, p_{i2}, \cdots, p_{iN_i}\} \tag{6}$$

を構成し，この G_i を $K-$ 平均アルゴリズムによってクラスタリングすることで，K 個のクラスタを作成する．そのクラスタの各重心ベクトル $\boldsymbol{w}^{g_{ic}}$ $(c = 1, 2, \cdots, K)$ と注目 Web ページ p_f の初期特徴ベクトル $\boldsymbol{w}^{p_f}$ とのベクトル空間における距離を反映させ，(7) 式を用いて改良特徴ベクトルを作成する．例えば，図 3 の場合には，p_f から 2 段目までのリンク先の段数ごとに Web ページ群 G_1, G_2 を構成し，これらの各 Web ページ群でクラスタリングを行い，各群で作成された重心ベクトルを $\boldsymbol{w}^{p_f}$ に反映させることで改良特徴ベクトルを作成することを示している．本手法において，p_f の改良特徴ベクトルの各成分 $w'^{p_f}_{t_k}$ は，(7) 式のように表される．(7) 式は，(2) 式の tf-idf 法によって求めた注目ページ p_f における単語 t_k の重み $w^{p_f}_{t_k}$ に，i 段目までのリンク先の段数ごとに構成される Web ページ群から作られるクラスタの重心ベクトルの成分 $w^{g_{ic}}_{t_k}$ とベクトル空間における p_f と g_{ic} の距離 $dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{g_{ic}})$ の逆数との積を，p_f から L 段目までのそれぞれの段数で生成されるすべてのクラスタの重心ベクトルについて加えていくことを示している．

$$\begin{aligned} w'^{p_f}_{t_k} &= w^{p_f}_{t_k} \\ &\quad + \sum_{i=1}^{L} \sum_{c=1}^{K} \frac{1}{dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{g_{ic}})} w^{g_{ic}}_{t_k} \end{aligned} \tag{7}$$

ただし，

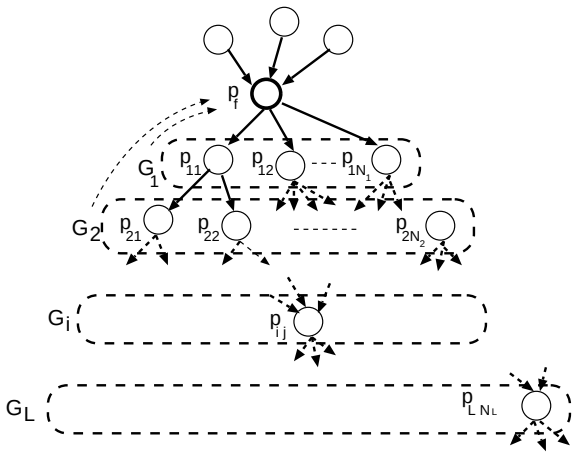

図 3　手法 3 によるベクトルの改良法

$$dis(\boldsymbol{w}^{p_f}, \boldsymbol{w}^{g_{ic}}) = \sqrt{\sum_{k=1}^{m}(w_{t_k}^{p_f} - w_{t_k}^{g_{ic}})}$$

である．

4. 評 価 実 験

3. で提案した各手法の有効性を確かめるために，次の 2 つの実験を行う．

（1） 改良した特徴ベクトルにおいて，成分値の高い単語がキーワードとなり得るか (以下，実験 1 とする)

（2） 改良した特徴ベクトルを Web ページの索引とし，既存の検索エンジンと比較してどれほどの検索精度が得られるか (以下，実験 2 とする)

(1) はリンク先ページの内容を考慮したことで，注目 Web ページのキーワードがより正確に抽出できるか否かを，(2) は改良特徴ベクトルが Web ページの索引として適切であるか否かを確かめることを目的とする実験である．

なお，3. で提案した手法は，デスクトップ PC(CPU: AMD Athron 1.4GHz，メモリ: 1GBytes，OS: Vine Linux2.1.5) 上に perl によって実装され，0.8GByte の Web ページ (250,000URL) を対象として実験を行った．各手法による Web ページの改良特徴ベクトル作成時間は，表 1 の通りである．

表 1　各手法による改良特徴ベクトル作成時間

	計算時間
手法 1	3.2 時間
手法 2	4.5 時間
手法 3	3.9 時間

4.1 実　験　1

注目 Web ページのキーワードが tf-idf 法に比べより正確に抽出できるか否かを確かめるために，改良特徴ベクトルにおいて，成分値の高い単語がキーワードとなり得るかについての実験を行った．

4.1.1 実 験 方 法

以下に述べる手順で実験を行った．

（1） 収集した Web ページ毎に，5 語の正解キーワードを人手によって作成する．

（2） 改良した特徴ベクトルにおいて，成分値の高い上位 10 語までに (1) で定めたキーワードがいくつ含まれるかを表す累積正解率を計算する．

4.1.2 実験結果・考察

3. で提案した各手法ごとの結果を図 4〜6 に示す．図 4 より，手法 1 では $L = 2$，すなわち注目ページから 2 リンク先までの個々の Web ページの内容を反映させることで，tf-idf 法に比べてキーワードをより正確に抽出できることがわかる．また，$L = 3$，すなわち注目ページから 3 リンク先までの個々の Web ページの内容を反映させた場合には，$L = 2$ の場合と比べてほとんど変わらないため，リンク先ページの個々の内容を反映したキーワード抽出においては，最低 2 リンク先のページまでたどれば効果があるものと考えられる．図 5 より，手法 2 では $L = 2, K = 3$，すなわちリンクを 2 つたどった Web ページから構成される 3 つのクラスタを用いて特徴ベクトルの改良を行った場合に，最も効果的にキーワードを抽出できることがわかった．さらに図 6 より，手法 3 では (5) 式，(7) 式からわかるように $L = 1$，すなわち 1 つ先のリンクに存在する Web ページの内容を反映させたときは手法 2 と同じ結果であるが，$L \geqq 2$，すなわち 2 リンク以上たどった場合には，$L = 1$，すなわち 1 つだけリンクをたどった場合と比較してキーワードを抽出する効果性は薄いことがわかった．これは，手法 3 では図 3 のようにリンクの段数毎に存在する Web ページをクラスタリングしているので，リンクによる Web ページ間の内容的なつながりが打ち消されてしまうことが原因であると考えられる．すなわち，図 5 と図 6 の比較から，ハイパーリンクを活用することで適切な特徴ベクトルが生成できることを示すことができたといえる．

4.2 実　験　2

3. で提案した手法で作成される改良特徴ベクトルを Web ページの索引として，既存の検索エンジンとの検索精度の比較を行った．

4.2.1 実 験 方 法

図 7 に実験システムの全体図を示す．この図からわかるように，本システムは，以下に述べる機能を

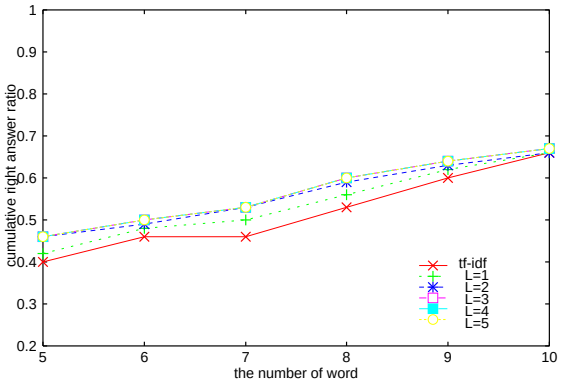

図 4　手法 1 による実験 1 の結果

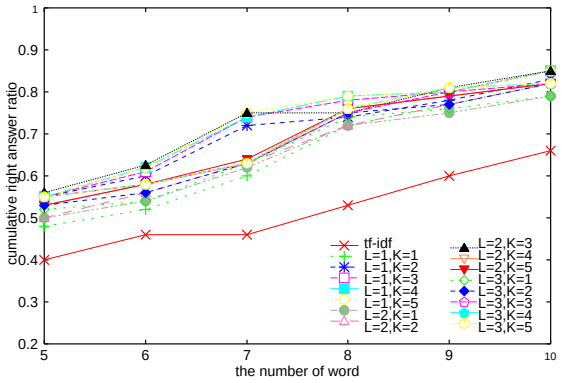

図 5　手法 2 による実験 1 の結果

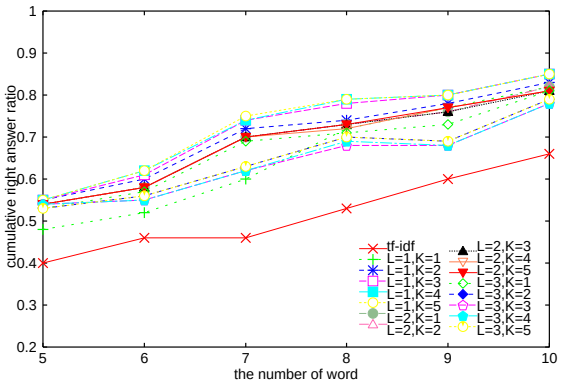

図 6　手法 3 による実験 1 の結果

持つ．

（1）　Web ページ収集部

Web ロボットを利用して，jp ドメイン内の日本語の Web ページを収集する．この制限は検索対象の質を向上させるために行っている．

（2）　リンク情報抽出部

収集した各 Web ページについて，リンク情報を抽出する．この際，前のページに戻るリンクは取り除き，そのページから先へたどることのできるリンクを抽出する．

（3）　Web ページ特徴ベクトル作成部

あらかじめ tf-idf 法によって作成した Web ページの初期特徴ベクトルを，3. で提案した手法により，改良特徴ベクトルを作成する部分である．

（4）　問合せベクトル作成部

問合せベクトル $\boldsymbol{Q}$ を

$$\boldsymbol{Q} = (q_{t_1}, q_{t_2}, \cdots, q_{t_m}) \tag{8}$$

と表す．t_k は索引語を表し，(8) 式の基底は初期特徴ベクトルの (1) 式と同じである．ただし，検索語ベクトル $\boldsymbol{Q}$ の各要素 q_{t_k} は，(9) 式のように表す．

$$q_{t_k} = \left(0.5 + \frac{0.5 \cdot Qf(t_k)}{\sum_{k=1}^{m} Qf(t_k)}\right) \times \log \frac{N_{web}}{df(t_k)} \quad (k = 1, 2, \cdots m) \tag{9}$$

(9) 式は，検索精度を最も良くする問合せベクトルの成分として、文献 [19] で報告されているものを利用している．ここで，$Qf(t_k)$ は問合せベクトル $\boldsymbol{Q}$ の中に含まれている索引語 t_k の数，N_{web} は収集した Web ページの総数，$df(t_k)$ は単語 t_k が出現する Web ページ数を表す．

（5）　検索インタフェース部

利用者からの問合せを受け付け，それに対する検索結果を提示する部分である．Web ページ p の特徴ベクトル $\boldsymbol{w}^p$ と問合せベクトル $\boldsymbol{Q}$ の類似度 $sim(\boldsymbol{w}^p, \boldsymbol{Q})$ を，(10) 式によって計算し，その値の降順に検索結果を提示する．

$$sim(\boldsymbol{w}^p, \boldsymbol{Q}) = \frac{\boldsymbol{w}^p \cdot \boldsymbol{Q}}{|\boldsymbol{w}^p| \cdot |\boldsymbol{Q}|} \tag{10}$$

なお，本研究では Web ページの内容が多様な分野にわたることを考慮して，Yahoo! Japan [20] のトップページのカテゴリを網羅するような以下に示す単語の組を検索語として用いた．また，各組について 10 個の適合文書を人手によって選定することで検索精度の評価を行なった．

検索語

{ 工芸，木工 }，{ 雇用，賃金 }，
{ インターネット，セキュリティ }，
{ 中学校，英語 }，{ ミュージカル，音楽 }，
{ 日本，外交 }，{ ガン，食事 }，
{ 新聞，ジャーナリスト }，
{ サイクリング，公園 }，{ 奈良，博物館 }，
{ 京都，観光 }，{ 画像処理，パターン認識 }，
{ 人類，民族 }，{ 住まい，掃除 }

4.2.2　実験結果・考察

比較の対象とした既存の検索エンジンとしては

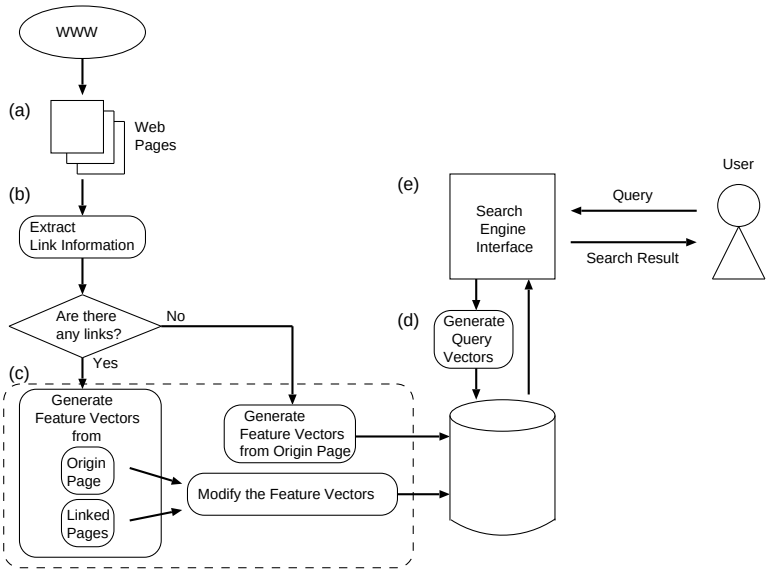

図 7　検索システムの全体図

Google [21] を用い，文献 [22], [23] の precision at 11 standard recall level に基づく再現率・適合率で評価を行った．なお，検索エンジンの出力結果全てを解析することは容易ではないため，上位 50 件を評価対象としている．3. で提案した各手法ごとの結果を図 8〜10 に示す．

図 8 より，手法 1 では $L = 1$，すなわち注目ページから 1 リンク先の個々のページを用いるよりも，$L = 2, 3$，すなわち 2 リンク先，3 リンク先の個々の Web ページの内容を反映させて改良特徴ベクトルを作成した方が検索精度が改善されること，$L = 2, 3$ で検索精度に大きな差がないこと，さらに $L \geqq 4$ では検索精度が $L = 2, 3$ と比較して悪くなることがわかる．これは，ある Web ページに着目した場合，その内容が 2〜3 のリンク先に集約できることを示しているものと考えられる．また手法 1 によって作成した改良特徴ベクトルを索引とした場合には，Google よりも良い検索精度が得られなかった．したがって，検索のための索引生成法としては，手法 1 は適切ではなかったといえる．

また，図 9 より，手法 2 では再現率の高い領域で Google よりも高い検索精度を実現することができた．また，$L = 1, K = 3$，すなわち 1 リンク先のすべての Web ページから構成される 3 つのクラスタを用いて改良特徴ベクトルを作成した場合，および $L = 2, K = 3$，すなわち 2 リンク先までに存在するすべての Web ページから構成される 3 つのクラスタを用いて改良特徴ベクトルを作成した場合に，より精度の良い検索が実現できた．特に $L = 2, K = 3$ は，前述したように，Web ページの内容は 2 リンク

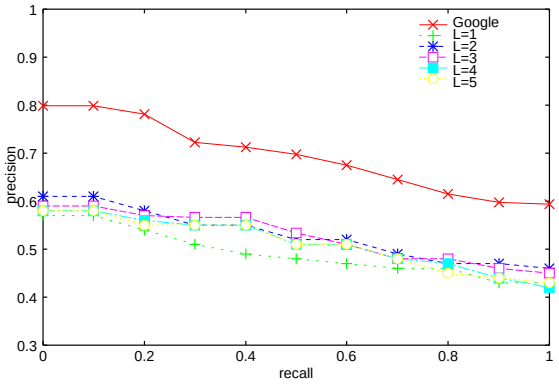

図 8　手法 1 と既存の検索エンジンの検索精度の比較

先までに集約できることを裏付ける事実として注目できる．また、リンク集のようなページが検索結果の上位に順位付けされやすい Google に比べ，より検索語の内容をとらえたページが上位に順位付けされるようになったことも特筆すべき点として挙げられる．

さらに，図 10 より，手法 3 では $L = 2$ 以降の検索精度が，手法 2 に比べて劣っているのがわかる．これは，リンクの段数ごとのクラスタリングを行うことで，ハイパーリンクによって構成されている Web ページ間の内容の継続性が途切れてしまうことによるものと考えられる．

5. おわりに

tf-idf 法によって作成した Web ページの特徴ベクトルを，そのリンク先ページを用いて改良する手法を提案した．また，その改良した特徴ベクトルの成

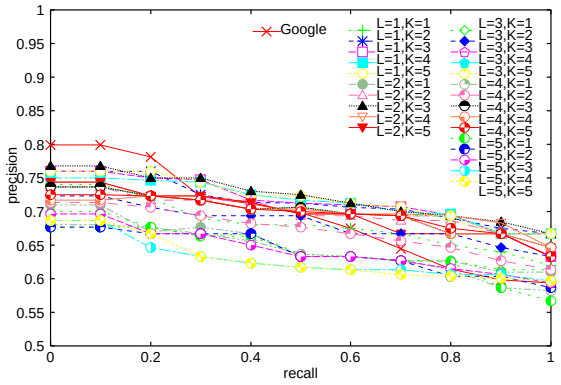

図 9　手法 2 と既存の検索エンジンの検索精度の比較

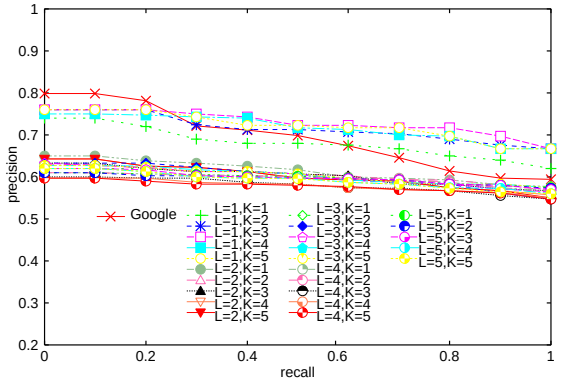

図 10　手法 3 と既存の検索エンジンの検索精度の比較

分値の高い単語が，キーワードとして適切であるか否か，および改良した特徴ベクトルが Web ページの索引として適切であるか否かを確かめる実験を行った．本研究によって，

- リンク先ページに着目して Web ページの特徴ベクトルを構成することにより，キーワード抽出や検索のための有効な索引生成を行うことができる
- Web ページの内容は，そのページから 2〜3 リンク先の Web ページに集約される

といった知見が得られた．また，本稿においては，各 Web ページのリンク先ページから生成されるクラスタ数は，$K = 1$〜5 に固定していた．しかし，個々の Web ページによってそのリンク先ページの状況は様々である．したがって，各々の Web ページのリンク環境に応じた特徴ベクトルの作成し，検索精度を検証することが，今後の課題として挙げられる．

謝辞　本研究の一部は，文部省科学研究費基盤研究 (B)(2)「言語横断型知識発掘システムに関する研究」(課題番号: 11480088)，基盤研究 (C)(2)「開放型高機能 XML サーチエンジンに関する研究」(課題番号: 12680417) ならびに奨励研究 (A)「XML で表現されるマルチメディアデータの効果的検索法に関する研究」(課題番号: 12780309) による．ここに記して謝意を表す．

文　　献

[1] *http://www.yahoo.com/.*
[2] *http://www.google.com/.*
[3] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. of the 7th International World Wide Web Conference*, April 1998.
[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *In proc. of ACM-SIAM Symposium on Discrete Algorithms*, 1998.
[5] 杉山一成, 波多野賢治, 吉川正俊, 植村俊亮. 順リンクを活用した web ページの主題抽出法. Proc. of DBWeb2001 情報処理学会シンポジウムシリーズ Vol.2001, No.17, pp. 209–216, 2000.
[6] K. Tajima, Y. Mizuuchi, M. Kitagawa, and K. Tanaka. Cut as a querying unit for www, netnews, e-mail. In *Proc. of the 1998 ACM Hypertext Conference*, pp. 235–244, 1998.
[7] K. Tajima, K. Hatano, T. Matsukura, R. Sano, and K. Tanaka. Discovery and retrieval of logical information units in web. In *Proc. of the 1999 ACM Digital Libraries Workshop on Organizing Web Space*, 1999.
[8] W. Li, K. Selçuk Candan, Q. Vu, and D. Agrawal. Retrieving and organizing web pages by 'information unit'. In *Proc. of the 10th International World Wide Web Conference*, pp. 230–244, 2001.
[9] J. Carriere and R. Kazman. Webquery: searching and visualizing the web through connectivity. In *Proc. of the 6th International World Wide Web Conference*, 1997.
[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *In proc. of the 7th International World Wide Web Conference*, pp. 107–117, 1998.
[11] B. Amento, L. Terveen, and W. Hill. Does 'authority' mean quality? predicting expert quality ratings of web documents. In *Proc. of the 22nd annual international ACM SIGIR Conference(SIGIR2000)*, pp. 296–303, 2000.
[12] S. Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *proc. of the 10th International World Wide Web Conference*, pp. 211–220, 2001.
[13] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks. In *Proc. of the 23nd annual international ACM SIGIR Conference(SIGIR2001)*, pp. 208–216, 2001.
[14] S. Fujita. Reflections on "aboutness" trec-9 evaluation experiments at justsystem. In *Proc. of TREC-9*, pp. 281–288, 2001.
[15] G. Salton and M.J. McGill. Introduction to modern information retrieval. *McGraw-Hill*, 1983.
[16] G. Salton. Automatic text processing: The transformation, analysis, and retrieval of information by computer. *Addison-Wesley*, 1989.
[17] G. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley, 1949.
[18] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc.*

of the 5th Berkeley Symposium on Mathmatical Statistics and Probability, pp. 281–297, 1967.

[19] G. Salton and C. Buckley. *Term-weighting approaches in automatic retrieval*. Information Processing & Management, 24(5):513-523, 1988.

[20] *http://www.yahoo.co.jp/*.

[21] *http://www.google.com/intl/ja/*.

[22] I. H. Witten and A. Moffatand T. C. Bell. *Managing Gigabytes*. Van Nostrand Reinhold, 149-150, 1994.

[23] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.