利用者の要求に応じたWebリンク自動生成手法

中谷　圭吾† 　鈴木　 優†† 　川越　恭二††

† 立命館大学大学院 理工学研究科　〒 525-8577 滋賀県草津市野路東 1-1-1
†† 立命館大学 情報理工学部　〒 525-8577 滋賀県草津市野路東 1-1-1
E-mail: †{nakatani,suzuki,kawagoe}@coms.ics.ritsumei.ac.jp

あらまし　本研究では，利用者が要求する Web ページを容易に閲覧することができるために，Web ページの適した部分から，その Web ページに関連した Web ページへのリンク自動生成手法を提案する．本稿では，1) Web ページを意味単位に分割する方法と，2) 分割した Web ページの適した部分からリンクを構築する方法を提案する．1) では，単語の出現密度分布を利用することによって， Web ページを意味単位に分割する．2) では，まず，Web ページ間の内容の類似度と，利用者がシステムへ入力したキーワードの二つの情報を用いることにより，キーワード検索による検索結果の最上位である Web ページと共に，その Web ページと関連する Web ページを決定する．そして，分割した検索結果の最上位である Web ページの各部分と関連する Web ページの内容の類似度を算出し，リンクを構築する．評価実験の結果，提案手法を利用することによって，関連リンクを構築することができ，利用者が容易に Web ページを検索することができることを確認できた．

キーワード　Web リンク，単語の出現密度分布，情報検索支援

Automatic and Adaptable Web Links Generation Mechanism for Individual Users

Keigo NAKATANI†, Yu SUZUKI††, and Kyoji KAWAGOE††

† Graduate School of Science and Engineering, Ritsumeikan Univ.
Nojihigashi 1-1-1, Kusatsu, Shiga, 525-8577 Japan
†† Faculty of Science and Engineering, Ritsumeikan Univ.
Nojihigashi 1-1-1, Kusatsu, Shiga, 525-8577 Japan
E-mail: †{nakatani,suzuki,kawagoe}@coms.ics.ritsumei.ac.jp

Abstract　In this paper, we propose a method for generating personalized Web links. These Web links are useful if users search Web pages related to the browsed Web pages. To generate these Web links, we proposed the following two processes, such as 1) the process of dividing the browsed Web pages into the meaningful blocks, and 2) the process of finding Web pages related to these blocks. We confirmed that, using our generated personalized Web links, users can find the Web pages related to the users' interests.

Key words　Hyperlink, Term Density Distributions, Assistance of Information Retrieval

1. はじめに

本稿では，利用者の検索要求に合致した Web ページへのリンクを Web サイト作成者ではなく，システムが自動的に構築するためのリンク構築法の提案を行う．

Web におけるリンクには，関連リンク，詳細リンク，戻るリンク，ブックマークリンクなどの意味を持つと考えられる．これらのうち，関連リンク呼ばれる，閲覧ページに関連の深いページへのリンクは，利用者の検索効率を向上させるためや，関係の深い情報の存在を利用者に提示するために重要なリンクであると考えられる．なぜなら，キーワード検索システムにおける検索結果は，入力されたキーワードを含むページに対するリンク群として表されるため，必ずしも利用者の検索要求に合致したページを発見できるわけではない．一方，検索されたページに類似したページが，利用者の検索要求に合致していることも考えられるため，検索されたページから類似ページへのリンクを構築することは有用であるといえる．これら関連ページへのリンクは，Web サイト構築者によって作成されることが多いが，Web サイト内に存在するページ数の増大により，Web サイト作成者によって全てのページを把握することが困難であ

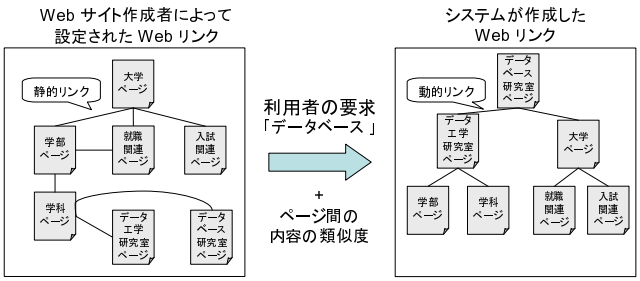

図 1　システムを利用したときの Web リンクの変化
Fig. 1　Change of the Web links when using a system.

る．そのため，関連リンクを構築することは，非常に手間のかかる作業であるといえる．

そこで本稿では，図 1 で示すように，企業・団体などの一つの Web サイト内において，Web サイト作成者が人手で構築していた関連リンクを自動生成する手法を提案する．提案手法では，ページ間の内容の類似度と，利用者がシステムへ入力したキーワードを用いることにより，キーワード検索による検索結果と共に，そのページと関連するページへのリンクを自動生成する．さらに提案手法では，リンク元のページにリンク先ページへのリンクを自動的に生成する際に，リンク先のページと関連が存在すると考えられるリンク元のページの部分を推定するための手法についても述べる．まず，リンク元のページに含まれる単語の出現密度分布を用いることによって，リンク元のページを意味単位に分割しておく．次に，分割したリンク元のページの各部分と，リンク先ページとの関連度を算出する．そして，最も関連したリンク元のページの部分にリンク先のページへのリンクを構築する．提案手法を利用することによって，ページ中の関連部分をリンク元とした，関連リンクを構築することができる．

2. 従 来 手 法

本章では，一つの Web サイトから要求するページを閲覧する方法として，一般に用いられているキーワードだけによる検索システムを挙げ，これらの問題点を挙げる．

キーワード検索システムでは，利用者がシステムに入力したキーワードに対する検索結果のページから，キーワードを含むページに対してのリンクを動的に生成して，利用者に提示する方法である．ここで，利用者に提示するための検索結果のページの多くは，順位付きリストである．そのために大量の検索結果が表示される場合が多く，検索結果を一つ一つ調べる必要がある．また，キーワードによっては，利用者が要求するページが検索結果に出力されないために，再度システムに別のキーワードを入力し，検索し直す必要がある．これらのような問題点があるために，利用者は要求するページを閲覧することが困難であると考えられる．つまり，利用者がシステムに入力したキーワードに対する検索結果からのリンクでしか，利用者の欲しい情報を得ることが出来ない．そのために，利用者が要求するページへたどり着くためには，多くのリンクをたどる場合があるといえる．

そこで，このような単なる検索結果のページからのリンクではなく，提示中のページに対しての関連リンクを自動的に構築するための手法を提案する．

3. Web リンクの自動生成方式

本稿では，企業・団体などの一つの Web サイトを対象とした，利用者が必要なページを容易に探すための関連ページへのリンクの自動生成を行う．提案手法では，ページ間の内容の類似度と，利用者がシステムへ入力したキーワードを用いることにより，キーワード検索による検索結果と共に，そのページと関連するページへのリンクを自動生成する．さらに提案手法では，リンクを自動生成する際に，リンク先のページと関連が存在すると考えられるリンク元ページの部分を推定するための手法についても述べる．ここでは，提示中のページと関連したページ間にリンクを構築する場合，提示中のページ中の内容の類似した場所に関連リンクを埋め込むことによって，利用者にとって必要なページを容易に探し出すことが可能になると考えられる．そこで，ページ内の単語の出現密度分布を考慮することによって，ページを意味のあるまとまりに分割する．そして，リンク先のページと関連が存在すると考えられるリンク元ページの部分からのリンクを構築する．

本稿では大きく分けて，以下に示すように前処理部と Web リンクの自動生成部の 2 部からなる．本稿で提案する手法の概要図を図 2 に示す．

- **前処理部**

Web サイト内のすべてのページ間の内容の類似度を算出し，K-means 法を用いて Web サイト内のページ群をグループ化しておく．また，3.1 節 Step 6 に述べる手法を用いて，ページ内の単語の出現密度分布からページを意味単位に分割しておく．

- **Web リンクの自動生成部**

利用者がシステムに入力したキーワードに対して，検索結果の最上位である最重要ページを算出し，前処理部で算出した各グループ内のページ群から有用と考えられる関連ページを一つずつ抽出する．各グループ内から抽出した関連ページと，単語の出現密度分布により分割した最重要ページ内の各まとまりとの類似度を算出し，もっとも類似度の高い最重要ページ内のまとまりの後ろに関連ページへのリンクを埋め込みページ間を Web リンクでつなぐ．

以下では，それぞれの処理について述べる．

3.1 前 処 理 部

まず，以下に示す六つのステップを前処理として行う．

Step1:　ページの取得

Web サイト内からページ群を取得し，ページのタイトル，アドレス，本文のソースをコンテンツデータベースに格納する．ここでいう本文のソースとは，<BODY> タグから </BODY> タグまでのことを示す．

Step2:　単語の抽出

取得したページの本文のソースから HTML タグ，JavaScript，コメント等を取り除き，本文のみを抜き出す．そして抜き出した本文から，形態素解析システム「茶筌」[1] を用いてページの

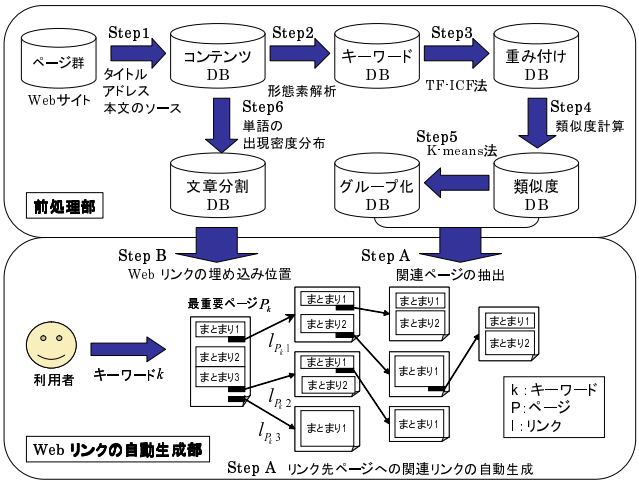

図 2　提案方式の概要図

Fig. 2　The overview of our proposed system.

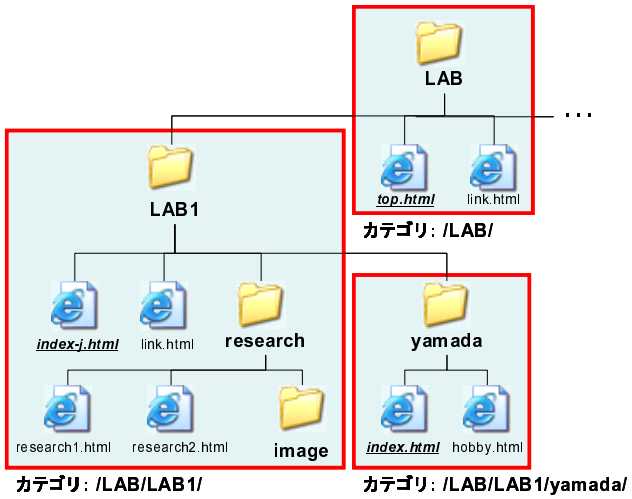

図 3　カテゴリの作成

Fig. 3　Creation of categories.

特徴を表す単語を抽出する．抽出する単語は，一般名詞・固有名詞・未知語とした．未知語とは，茶筌では解析できない半角英数文字や辞書に登録されていない単語であるが，ページの特徴を表すキーワードが含まれているため，1 文字で構成される語，数字・記号のみで構成される語，あらかじめ設定しておいた不要語を除いたキーワードのみを抽出する．

Step3:　単語の重み付け

Step 2 で得たページごとの各単語に関して，TF-ICF(Term Frequency・Inverse Category Frequency) 法 [2] を用いて単語の重み付けを行う．TF-ICF とは，単語の重み付け手法の一つである TF-IDF(Term Frequency・Inverse Document Frequency) 法 [3] を発展させたものである．TF-IDF は，単語の出現頻度 (TF) と，ページ総数における単語の出現するページ数の逆数 (IDF) との積により求められる．TF は単語の網羅性を表し，IDF は単語の特定性を表しており，これらの積である TF-IDF は網羅性と特定性が共に高い単語の重みが高くなっている．しかし，Web サイト内のページ群は階層化されており，同じカテゴリ内のページ同士は関連が高いといえるため，ページ単位の重み付けである TF-IDF を用いるより，カテゴリ単位に重み付け計算を行った TF-ICF を用いた方が良いと考える．TF-ICF により重みを計算することによって，カテゴリの内容をより反映した重み付けを行うことができる．TF-ICF による単語の重みを式 (1) に示す．ここで，単語 k_j $(j = 1, 2, \cdots, M_i)$ のあるページ p_i $(i = 1, 2, \cdots, N)$ における重みを w_{ij}，TF による重みを T_{ij}，ICF による重みを C_j，p_i に含まれる k_j の出現頻度を f_{ij}，Web サイト中の総カテゴリ数を O，k_j が含まれるカテゴリ数を c_j とする．

$$\begin{aligned} w_{ij} &= T_{ij} \cdot C_j \\ &= f_{ij} \cdot (\log \frac{O}{c_j} + 1) \end{aligned} \tag{1}$$

[カテゴリの作成]

本稿が示すカテゴリとは，以下の 2 種類の条件を満たすディレクトリ同士を一つのカテゴリとして統合したものである．図 3 にカテゴリの作成例を示す．

(1)　ディレクトリ内の下位階層の複数のディレクトリ同士は同じカテゴリである．

(2)　(1) を満たす場合でも，下位階層のディレクトリにトップページを含む場合，その下位階層のディレクトリは，別カテゴリである．

ここでいうトップページとは，index*.* ，top*.* ，main*.* ，home*.* となっているファイル名のことを示す．

以下にカテゴリの作成手順を示す．ここで，作成した各ページごとのカテゴリ名を基に，総カテゴリ数 O とある単語 $k_u (u = 1, 2, \cdots, M_i)$ が含まれるカテゴリ数 c_u を算出する．O は重複無しのカテゴリ名をカウントしたものであり，c_u は k_u を含むページ重複無しのカテゴリ名をカウントしたものである．

(1)　Step 1 で格納された URL の後ろから一番初めのスラッシュ (/) までをページが格納されているディレクトリとし，同じディレクトリ内にトップページが存在するかを確認する．存在する場合 (3) へ進む．存在しない場合 (2) へ進む．

(2)　(1) で求めた URL の後ろから前のスラッシュまでを親階層のディレクトリとし，同じディレクトリ内にトップページが存在するかを確認する．存在する場合 (3) へ進む．存在しない場合 (2) を繰り返す．スラッシュがない場合 (3) へ進む．(2) を 3 回以上繰り返した場合 (4) へ進む．

(3)　そのディレクトリをページのカテゴリ名とし終了する．

(4)　(1) のディレクトリをページのカテゴリ名とし終了する．

Step4:　ページ間の内容の類似度算出

ページ間の内容の類似度を算出する方法として，一般に良く用いられているベクトル空間モデル [4] を用いる．すなわち，ページに含まれる単語を多次元空間上のベクトルとして表現し，類似度関数 S によって二つのベクトル間の類似度を求める．ページ p_i の持つ特徴ベクトル $\mathbf{P_i}$ を式 (2) で定義する．

$$\mathbf{P_i} = [w_{i1}, w_{i2}, \cdots, w_{iM_i}] \tag{2}$$

このとき，ページ p_x とページ p_y の類似度 $e(p_x, p_y)$ $(x, y = 1, 2, \cdots, N)$ を式 (3) で定義する．

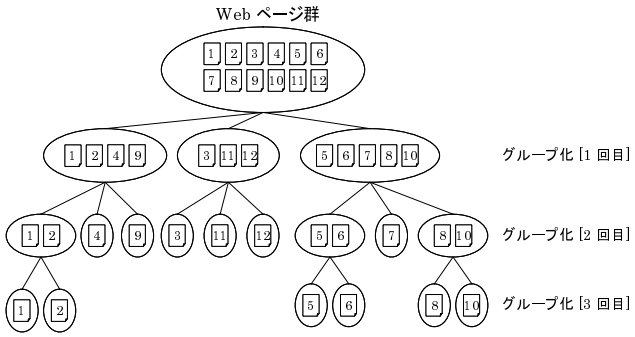

図 4　ページ群のグループ化
Fig. 4　Grouping of the pages in a Website.

$$e(p_x, p_y) = S(\mathbf{P_x}, \mathbf{P_y}) = \frac{\sum_{j=1}^{M_x+M_y}(w_{xj} \cdot w_{yj})}{\sqrt{\sum_{j=1}^{M_x+M_y} {w_{xj}}^2} \cdot \sqrt{\sum_{j=1}^{M_x+M_y} {w_{yj}}^2}} \quad (3)$$

Step5:　グループ化

Step 4 で算出したページ間の内容の類似度を基に，代表的なクラスタリング手法の一つである K-means 法 [5] を用いて，Web サイト内のページ群に対してクラスタリングを行う．本 Step では，図 4 に示すように，一つの Web サイトに含まれるページ群全体に対してグループ化を行うだけでなく，グループ化した後の各グループに対しても，同ように K-means 法を適用し，グループ内のページ数が一つになるまでグループ化を行うものとする．

Step6:　ページ内の文章分割

関連したページ間にリンクを構築する場合，関連ページに類似したリンク元ページの部分から，リンクを構築することによって，利用者の検索要求に合致したページを容易に閲覧することができると考えええられる．しかし，ページの多くは，HTML 言語で記述されているために，デザイン重視・構造無視のページが多い．そのため，タグ情報だけでは，ページ内を意味単位に分割することができない場合が存在する．

そこで，利用者に対して有用であると考えられる複数の関連ページへのリンクを，リンク元のページ中の適切な場所に，適切なリンクを配置するために，各ページを意味単位に分割する．分割する方法として，ページ中に含まれる単語の出現密度分布により，ページを区切るための単語を抽出し，その単語を基にページを複数に分割する．ここで，抽出した単語でページを分割してしまうと，文章の途中で分割されてしまう．そこで，抽出した単語が出現した次の，特定の HTML タグまでをひとまとまりとする．このタグは，通常 HTML 言語において，文章を区切る意味を持つタグであり，表 1 に示す．

以下にページの分割方法を示す．

(1)　ページ $p_i (i = 1, 2, \cdots, N)$ から「茶筌」を用いてページの特徴を示す単語を出現順に抽出する．

(2)　ページ p_i 中に含まれる M_i 種類の単語 $k_j (j = 1, 2, \cdots, M_i)$ ごとに出現位置の平均と標準偏差を算出する．そして，抽出した同一の単語ごとに平均 ± 標準偏差を算出し，単語ごとの出現範囲とする．標準偏差を式 (4) に示す．ここで，ページ中の単語 k_j の標準偏差を D_j とし，B_j 個存在する k_j の出現位置を $R_j(v), (v = 1, 2, \cdots, B_j)$，$k_j$ の出現位置の平均を A_j とする．

$$D_j = \sqrt{\sum_{v=1}^{B_j} \frac{(R_j(v) - A_j)^2}{B_j}} \quad (4)$$

(3)　各単語ごとの出現範囲が，設定した閾値以上の文字間隔があいていれば，その平均値である出現位置の単語を区切り文字として抽出する．ただし，出現数が一つの単語は，区切り文字としては用いない．

(4)　抽出した単語の次の，表 1 に示された特定の HTML タグまでをひとまとまりとする．ただし，最後の文章の区切りから文章の最後までの間に，表 1 に示す特定のタグが存在する場合は，ひとつのまとまりとする．

あるページにおける分割例を図 5 に示す．図 5 の (a) はページの文章から出現順に抽出した単語，(b) はそれらの単語の出現密度分布と抽出した単語による文章間の区切り，(c) は実際の分割したページを示す．(b) の抽出した単語によるまとまり数が六つに対し，(c) の実際のまとまり数が二つになっているのは，(c) の各まとまりの抽出した単語の次の，表 1 に示されたタグが同じであるために，同じひとまとまりになったためである．

3.2　Web リンクの自動生成部

3.1 節で算出した結果を基に，次に示す二つのステップから Web リンクの自動生成を行う．

StepA:　リンク先ページへの関連リンクの自動生成

ページ間の内容の類似度と，利用者がシステムに入力したキーワードを基に，検索結果の最上位であるページを最重要ページとし，そのページに対する関連リンクを自動生成する．ここでは，図 6 に示すように，最重要ページに対する関連ページを，3.1 節 Step 5 であらかじめグループ化しておいた，Web サイト内のページ群の各グループから，一つずつ抽出する．そして，各グループの関連ページへのリンクを StepB で述べるように分割された最重要ページに埋め込む．次に，構築されたリンクの中から，利用者が選択したリンク先のページが含まれるグループを，さらにグループ化したグループ内から，先程と同様に関連ページを抽出する．そして，利用者が選択したリンク

表 1　文章を分割するために利用する HTML タグ
Table 1　HTML tag used in order to divide texts.

タグ名	タグの意味
<HR>	横線タグ
<H 数値 >	見出しタグ
<P>	段落タグ
<SPAN>	ひと塊を示すタグ（インライン要素）
<DIV>	ひと塊を示すタグ（ブロック要素）
  	連続して出現する改行タグ
<TABLE>	表作成のためのタグ

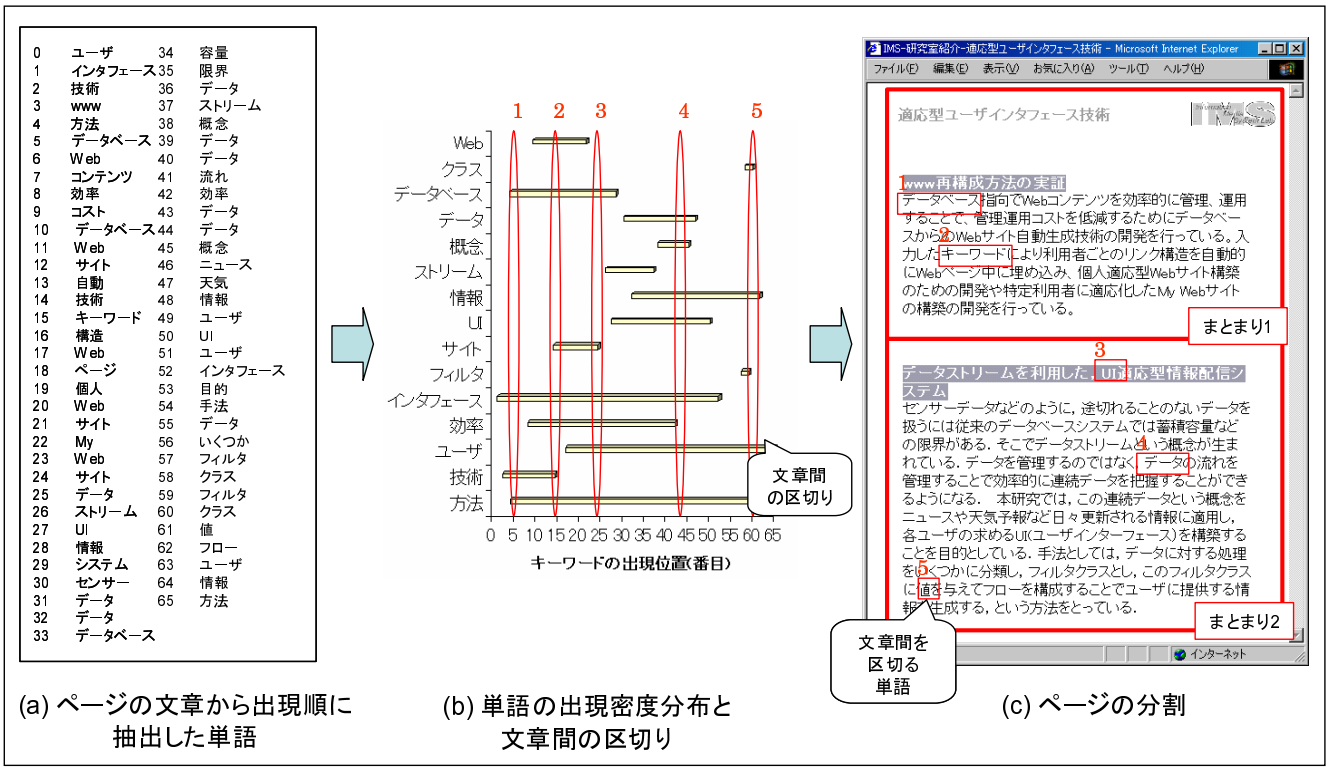

図 5　ページ分割の例

Fig. 5　Example for division of a page.

先のページに対して，関連ページへのリンクを自動的に生成する．これ以降は，グループ内のページがなくなるまで，同様にリンクを生成していく．

[関連ページの抽出方法]

リンク元のページに対して，関連するページへのリンクを構築するために，関連ページを抽出する．ここで，閲覧ページに類似度の高いページは，重要なページであるといえる．しかし，必ずしもリンク元のページが要求するページとはいえない．そのため，リンク元のページが，利用者の要求するページと異なる場合には，要求するページからかけ離れてしまう可能性がある．また，利用者がシステムに入力したキーワードは，利用者の要求を示す単語であるために重要なキーワードであるといえる．しかし，利用者がシステムに入力したキーワードに対する出現頻度が高いページは，リンク元のページに対して，類似した内容のページばかりであるとはいえない．

そこで，これらそれぞれの偏りを解消するために，次の方法を用いて，関連ページを抽出する．

利用者がシステムに入力したキーワードに対する出現頻度と，リンク元ページとの類似度を，それぞれ降順に並び替える．次に，二つのリストの順位をページごとに足し合わせた結果，最小値となるページを関連ページとして抽出する．ただし，最重要ページと選択したページは，既に閲覧したページであるため，関連ページとして抽出しない．

StepB:　Web リンクの埋め込み位置

リンク先のページと関連が存在すると考えられる部分へのリンクをページ中に埋め込むために，3.1 節で述べた式 (3) を用いて，リンク先のページに対する分割されたリンク元のページの各まとまりとの内容の類似度を算出する．そして，最も類似度の高いまとまりの右下の位置に抽出したページのリンクを埋め込む．リンク先のページへのリンクを，リンク元のページのどのまとまりに埋めるのかを式 (5) で算出する．ここで，リンク先のページを L，リンク元のページの各まとまりを $E_z(z = 1, 2, \cdots, D_i)$ とし，L と E_z の中から最も類似しているまとまりを E_L とする．

$$E_L = \max e(L, E_z) \tag{5}$$

上記のように関連ページへのリンクを埋め込むことによって，リンク元のページ中に自然な形で関連リンクを埋め込むことが可能となり，利用者が要求するページを容易に提示することができると考えられる．

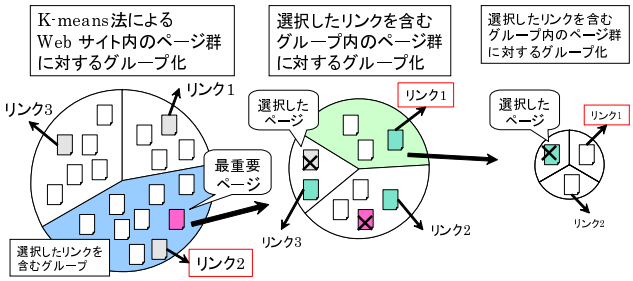

図 6　Web リンクの自動生成

Fig. 6　Automated generation of Web links.

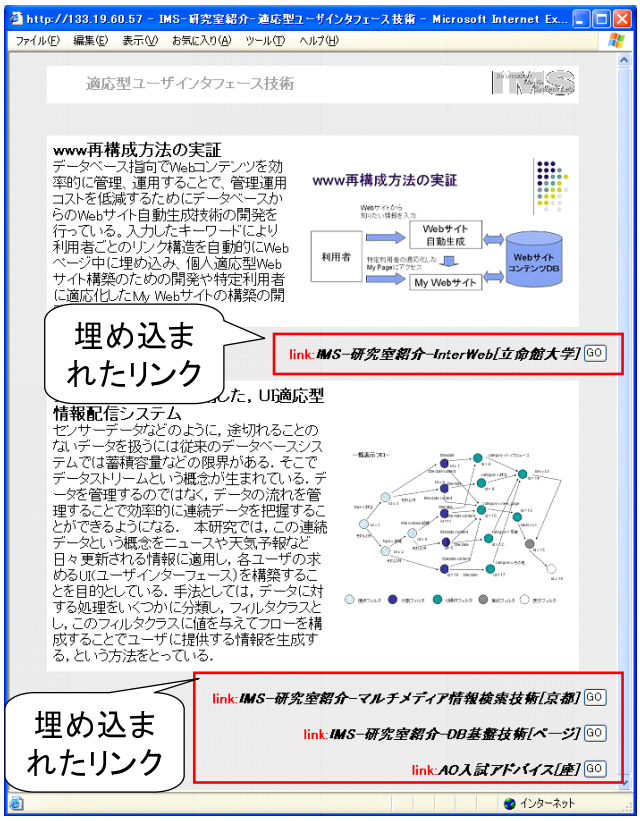

図 7　試作システムが生成したページ

Fig. 7　The page created by the system.

4.　試作システムの構築と評価

本提案が有効であることを確かめるために試作システムを構築した．実験データとして立命館大学のサイトから学部・学科・研究室・入試関連などから 2500 ページを抽出した．一つのページに対して生成されるリンク数は 4 として，3.1 節 Step 5 では，四つにグループ化を行う．試作システムが生成したページを図 7 に示す．

本章では，以下に示す 2 種類の実験を行なった．一つ目に単語の出現密度分布を用いて分割した場合とタグのみで分割した場合のページ分割の精度の比較に関する実験を行う．二つ目の提案システムの性能に関する実験では，あらかじめ筆者が決定した正解集合のページを閲覧するために必要なアクセス回数の測定を行った．ここでは，キーワード検索システムを用いた場合と，元々のリンク構造をたどった場合のアクセス回数との比較を行い，提案システムの有効性を示す．

4.1　ページ分割精度に関する評価

ランダムに選択した 7 ページに対して，筆者が理想とする分割箇所数と，システムが行った分割箇所数と，前者と後者の同じ分割箇所数に対して，閾値を変化させたときの再現率と適合率の平均と，表 1 に示したタグのみを用いて分割したときの再現率と適合率の平均を表 2 に示す．ここで，閾値とは，3.1 節 Step 6 において，抽出した文章を区切るための単語間の間隔のことを示す．本実験で用いた再現率と適合率を以下の式で求めた．ここで，再現率を α，適合率を β とし，理想とする分割箇所数を γ，システムが行った分割箇所数，表 1 で示したタグを用いて分割したときの分割箇所数を δ，両者の同じ分割箇所数を ϵ とする．

$$\alpha = \frac{\epsilon}{\gamma} \tag{6}$$

$$\beta = \frac{\epsilon}{\delta} \tag{7}$$

表 2 の結果から，表 1 に示されたタグを用いて分割した場合，それらのタグがページ中に多数存在するので，適合率は約 27% と低い．これに対し提案手法を用いて分割した場合の閾値をページ中の 単語数/30 としたとき，適合率は約 86% を実現しておりタグのみの分割より良くなっているといえる．しかし，そのときの提案手法の再現率は，約 64% しか実現しておらずまだ十分ではない．考えられる原因として，昨今のページの作成の変化により，<table> タグの役割が表としてではなくレイアウトとしての要素を持っていることである．また，ページは定まった書式ではないため，表 1 に示されたタグが多用してある一方で，見た目では一つのまとまりではないためである．

なお，表 2 において区切るための文字間隔が少なくなっているのに適合率が下がっているのは，分割箇所が増えているのに，それに伴って両者の同じ分割箇所数が増えていないためである．

4.2　提案システムの性能に関する評価

提案システムを用いて，リンクが有効に作成されているかを調べるために，以下に示す評価実験を行なった．

“ バリアフリー ”に関する正解集合をあらかじめ著者により 3 ページ決定しておき，正解集合を閲覧するために必要と考えられる 5 種類のキーワードを入力し，提案システムとキーワード検索を用いて，それぞれの平均アクセス回数を測定した．また Web サイトのトップページから Web サイト作成者によって構築された元々のリンクを用いてたどった場合のアクセス回数も測定した．それぞれの方法を用いたときのアクセス回数の結果を表 3 に示す．ここで，キーワード検索においては，1 位から順に閲覧していくものとする．図 8 では，提案システムに入力したキーワードを「バリアフリー」としたときの正解集合を閲覧するためのアクセス順序を示す．

以下に示すページは，正解集合に用いた 3 ページに対するタイトルである．

- バリアフリー研究室トップ
- 日本ロボット学会誌:「福祉社会と IT 技術」
- バリアフリー研究室紹介資料

提案システムを用いた場合の正解集合のページを閲覧する

表 2　ページ分割の精度

Table 2　Accuracy of Web page division.

閾値	再現率の平均	適合率の平均
単語数/10	0.119047619	0.214285714
単語数/20	0.441666667	0.821428571
単語数/30	0.638888889	0.861111111
単語数/40	0.659722222	0.785714286
単語数/50	0.659722222	0.785714286
タグによる分割箇所数	1	0.268522375

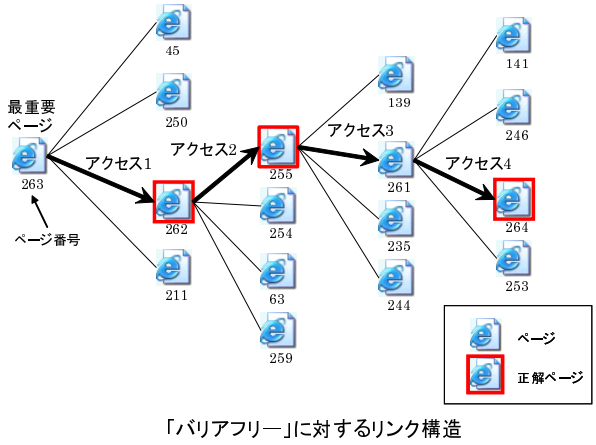

図 8　提案システムを用いたアクセス例

Fig. 8　Example of access using the proposal system.

のに必要な平均アクセス回数は，5.8 回であった．これに対し，キーワードだけによる検索システムを用いた場合，「バリアフリー」，「高齢者」のキーワードについては，正解集合のページを平均 13.5 回で閲覧できた．その他のキーワードでは，正解集合のページをすべて閲覧することはできなかった．また，元々のリンクをたどった場合のアクセス回数は，10 回であった．つまり，提案システムでは，キーワードだけによる検索システムを用いることや，元々のリンクをたどることより要求するページを容易に閲覧することができるといえる．

提案手法を用いることにより，Web サイト作成者が人手で構築していた関連ページへのリンクを自動的に構築することが可能となった．

5. 関 連 研 究

本章では，Web リンクの自動生成に関する研究などを示し，提案手法との差異について述べる．

5.1 リンクやページの自動生成

利用する状況や環境に合わせて Web データを変化させる Web 適応に関して，清光ら [6] [7] が提案している．清光らは，端末の処理能力や通信環境に応じた Web コンテンツの動的な再構成や，利用者のアクセスパターンからページ作成者の意図に基づいて誘導できるリンク構造の動的生成，Web コンテンツの個別化などについて研究している．これらの研究は，あらかじめ存在するページにリンクを動的に生成し，提示している部分で提案手法と類似しているが，これらの研究ではページ作成者の意図するリンク巡行を動的に構築するのに対し，利用者の要求に対して自動的に関連するページへのリンクを自動生成する部分が本研究と異なる点である．

表 3　正解集合を閲覧するためのアクセス回数の比較

Table 3　Comparison of access frequencies for perusing correct answer pages.

入力キーワード	アクセス回数		
	試作システム	キーワード検索	元々のリンク
バリアフリー	4	13	
福祉	5	*	
高齢者	7	14	
手話	6	*	
介護	7	*	
平均アクセス回数	5.8 回	13.5 回	10 回

*　入力したキーワードでは，正解集合のページをすべて閲覧することができなかった．

また福村ら [8] は，個人化を考慮した Web サイトモデルとして，ページ単位ではなく，コンテナやコンテンツ，リレーションといった単位のコンポーネントからユーザの過去に辿った視聴傾向を調べ，それらを用いることにより利用者に適したコンポーネントの抽出手法と，個人化したページの動的な作成方法を提案している．この研究との類似点は作成したページ間がリンクでつながれている部分であるが，ページ中の適した場所に適したリンクを埋め込んでいる部分で本研究とは異なる．

また，我々の従来の研究で階層的クラスタリングを用いて，利用者がシステムへ入力したキーワードを用いることにより，Web サイト内の類似したページへのリンクを自動生成する方法を提案した [9]．従来の研究では，利用者がシステムに入力したキーワードに対して算出したページの階層構造の近いページから順にリンクとして表示したため，類似したページから順に閲覧していくことが可能であった．しかし，リンク先の内容が利用者の入力したキーワードによらず固定され，対象とするページが多くなれば，多くのリンクをたどる場合がある．そこで本稿では，非階層的クラスタリングを用いてグループ化した結果をさらにグループ化していく．これにより，利用者が要求するページを容易に探し出すことのできる関連リンクの自動生成手法を提案する．

5.2 クラスタ型検索モデル

検索実行時のファイル探索・処理の効率化や検索性能の向上を目的としたもので，クラスタ型検索モデルが存在する．これには，階層型と非階層型があり，検索質問に対する適合度が高いクラスタを利用者に提示する．代表的なシステムでは階層型として，Vivisimo [10] や GATA [11]，非階層型として，ナレッジクエリーサーチ [12] がある．これらは，利用者がシステムに入力したキーワードに対して類似したグループ内のページを提示している．これらと提案手法は，検索結果を提示する方法としてクラスタリング手法を用いる点で類似しているが，検索結果の提示方法として，利用者がシステムに入力したキーワードに対して，類似したグループ内のページに対してのリンクではなく，提示したページ中の適した場所に，適したリンクを構築している部分で大きく異なる．また，リンクをたどることですべてのページを閲覧できる点も異なる．これにより，利用者の要求するページを順に閲覧することができると考えられる．

5.3 ページ内の文章分割

Deng Cai ら [13] は，ページをブロック単位に分割するために，VIPS アルゴリズムを提案している．ここでは，ページのタグ構造 (DocumentObjectModel) と視覚的な情報 (背景色や文字サイズなど) であるレイアウトの特徴から意味的な水準になるようにページ内の文章を分割している．提案手法では，単語の出現密度分布とタグ情報から分割している．

5.4 単語の出現密度分布

佐野ら [14] は，ページを検索する際にハニング窓関数を使用してページ内の検索単語の出現密度分布を考慮したスコアリングを行っている．単語の出現密度分布を考慮する点で類似しているが，利用目的の点で異なる．佐野らは，このスコアリングを用いることによって，検索キーワードの重要度を検索結果に反映させるために用いており，提案手法ではページ内を分割するために用いている．

6. おわりに

本稿では，ページ間の内容の類似度と，利用者がシステムへ入力したキーワードを用いることにより，閲覧ページに関連するページへのリンクを自動生成するための手法を提案した．提案手法では，単語の出現密度分布を考慮することによりページを複数のまとまりに分割した．その結果，リンク先のページとの類似度が最も高い部分からのリンクを構築することができた．

提案手法を用いることにより，Web サイト作成者が人手で構築していた，関連ページへのリンクを自動的に生成することが可能となった．

今後は以下の課題を解決する予定である．

- **大規模な Web サイトに対する提案手法の適用**

本稿では，提案システムの構築のために，立命館大学の Web サイトからページを抽出した．これらの多くのページの構成は統一されておらず不要なページが多い．今後は，HTML 言語で記述されたページではなく，構成が統一された別のコンテンツ記述を利用したニュースサイトの記事などに対する提案手法の適用と，動的 Web リンクの生成を可能にする提案手法の効率化を検討したい．

- **リンクの解析に関する検討**

本稿では，ページ中に含まれている元々のリンクは考慮されていない．つまり，ページ間はフラットな関係であり，詳細ページから概要ページへリンクが貼られている可能性がある．Web 作成者によって構築された元々のリンクの場合，概要ページから詳細ページへのリンクがページ作成者の意図されたリンクであり，その逆のリンクは戻るためのリンクであり，望ましくないと考えられる．そこで，カテゴリ内の階層構造や元々のリンクを考慮して，リンクを抽出することを検討したい．また，提案手法では，元々のリンクを選択した場合，提案システムが適用されない．元々のリンクを選択することで，要求するページを閲覧できることも考えられるので，元々のリンクを選択した場合でも，提案手法が適用できるように検討したい．

- **クラスタ数を自動決定するアルゴリズムの適用**

本稿では，クラスタリング手法として，K-means 法を用いた．これは，初期クラスタ集合からより良いクラスタを漸近的に求めるアルゴリズムであり，初期クラスタの取り方に依存する．また，クラスタ数がリンク数であるため，リンク数は常に固定であり，提案手法にとって大きな制約である．そこで，クラスタ数を自動決定する，ISODATA 法 [15] や X-means 法 [16] [17] の適用を考えている．

- **より信頼性の高い類似度算出方法の検討**

最近では，画像や flash など，テキスト以外のデータを含んだページが増加しており，ページの特徴を表す単語がページ中に出現しない事が多いため，テキストデータのみによる類似度算出では，信頼性が高いとはいえなくなってきている．そこで，マルチメディアデータを含めた類似度の算出や，元々存在するリンクを活用することにより，信頼性の高い類似度の算出方法が必要である．

- **精度の良いページ分割方法の検討**

本稿では，適切な場所に適切なリンクを埋め込むために，ページ中の単語の出現密度分布よりページ内を意味単位に分割した．しかし，評価によると理想の分割にはまだ十分ではない．そこで，単語の出現密度だけでなく，文字サイズや文字色，背景などの視覚的な情報やタグの構造を考慮することによって，より理想的な分割に近づけるようなページ分割の方法を検討したい．

文　献

[1] 松本裕治: 形態素解析システム「茶筌」, 情報処理, Vol. 41, No. 11, pp. 1208–1214 (2000).

[2] Ko, Y. and Seo, J.: Automatic Text Categorization by Unsupervised Learning, *Proceedings of the 17th conference on Computational linguistics (COLING-2000)*, pp. 453–459 (2000).

[3] Salton, G., Lesk, M. E.(編): *Introduction to Modern Information Retrieval*, McGrawHill Book Co. (1983).

[4] Salton, G., Wong, A. and Yang, C. S.: A vector space model for automatic indexing, *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620 (1975).

[5] Mac Queen, J.: Some Methods for Classification and Analysis of Multivariate Observations, *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, pp. 281–297 (1967).

[6] 清光英成, 田中克己: Web リンクの巡行に基づく動的なリンク活性化とアクセス管理, 情報処理学会論文誌:データベース, Vol. 42, No. 8, pp. 10–20 (2001).

[7] 清光英成, 竹内淳記: Web データの個別化と環境適応, 情報処理学会論文誌:データベース, Vol. 42, No. 8, pp. 185–194 (2001).

[8] 福村真哉, 中野賢, 春本要, 下條真司, 西尾章治郎: ユーザの視聴傾向に基づき個人化した Web ページの動的作成, 情報処理学会 研究報告, Vol. 132, No. 8, pp. 57–64 (2004).

[9] 中谷圭吾, 川越恭二: 適応型 Web サイト構築のためのリンク構造の自動生成手法, 第 2 回情報科学技術フォーラム (FIT2003), pp. 61–62 (2003).

[10] VivisimoInc.: . Vivisimo http://vivisimo.com/.

[11] 日立製作所, 東京工業大学, 北陸先端科学技術大学, 文部省国文学研究資料館: 汎用連想検索エンジンの開発と大規模文書分析への応用, 第 18 回 IPA 技術発表会 (1999).

[12] FujitsuBusinessSystemLTD.: . ナレッジクエリーサーチ http://kd.iws.ne.jp/kms/kqs/q-search.html.

[13] Cai, D., Yu, S., Wen, J.-R. and Ma, W.-Y.: Extracting Content Structure for Web Pages based on Visual Representation, *Proceedings of the 5th Asia Pacific Web Conference* (2003).

[14] 佐野稜一, 松倉健志, 波多野賢治, 田中克己: 部分グラフを基本単位とした Web 文書検索:単語の出現密度分布の適用, 情報処理学会研究報告：データベースシステム, Vol. 99, No. 61, pp. 79–84 (1999).

[15] 電子情報通信学会 (編): パターン認識, 電子情報通信学会 (1988).

[16] Pelleg, D. and Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)*, pp. 727–734 (2000).

[17] 石岡恒憲: クラスター数を自動決定する k-means アルゴリズムの拡張について, 応用統計学, Vol. 29, No. 3, pp. 141–149 (2000).