# 文書間類似度とキーワードを用いた Web リンク自動生成手法

# Personalized Web Link Generation Method using Keywords and Document Similarities

中谷 圭吾♡ 鈴木 優♢ 川越 恭二♢

Keigo NAKATANI Yu SUZUKI
Kyoji KAWAGOE

**本論文では，利用者の閲覧している Web ページから，その内容が類似した Web ページへのリンク，つまり関連リンクを自動的に生成するための手法を提案する．本手法では，リンク元がリンク先と関連する部分となるように関連リンクを構築することを目的としている．そのため適切なリンク元ページ，リンク先ページを自動的に選択しなければならない．そこで，キーワード検索システムによる最上位の検索結果をリンク元とし，検索対象 Web ページ群を分類した後の各クラスタの任意の Web ページをリンク先ページとするリンクを自動的に構築する．提案手法によって構築されたリンクを用いることにより，キーワード検索システムによる検索結果が利用者の検索要求に適合していない場合であっても，利用者が適切なリンクをたどることによって，利用者の必要な Web ページを容易に検索することが可能となる．**

**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.**

## 1. はじめに

近年の Web 検索システムは，問合せとしてキーワードが用いられることが多い．そのため，多くの研究者によって Web 検索システムが開発されており，実際に利用されている．ところが，適切なキーワードが利用者によって入力されていない場合や，検索システムの性能の限界により，適切な検索結果を得ることができない場合がある．その結果，利用者が入力したキーワードだけを入力とした検索システムから，利用者の必要な情報を得ることが難しい．

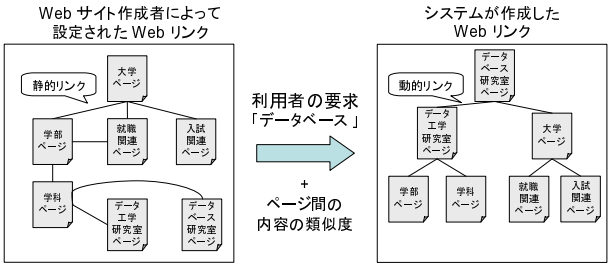


図 1: 提案システムによるリンク構築
Fig. 1 Generating Web links by our proposed system.

一方，Web サイト作成者は，利用者が必要な情報をサイトに含まれる Web ページ群から必要な情報を提供できるように，トップページから数多くのリンクが構築されている．ところが，サイトに含まれる Web ページの数，種類が多い場合，利用者が必要な情報を既に構築されているリンクだけから探索することは難しい．なぜなら，サイト作成者が作成するリンク構造は必ずしも利用者にとって容易に探索可能な構造であるとはいえないためである．

これらの問題を解決するための手法として，多くのサイトではサイト内検索と呼ばれる，ある一つのサイト内の Web ページ群だけを検索対象とした検索システムが付与されていることが多い．サイト内検索システムは，Web 全体を対象とした検索システムと比較して検索対象 Web ページ数が大幅に少ないため，利用者によるサイトの選択が適切である場合には必要な Web ページが得られる可能性が高い．ところが，サイト内検索では十分に大量の Web ページ群が存在しているとはいえないため，利用者が入力するキーワードに適合する Web ページが存在しないことが多い．そのため，利用者の必要な情報がサイト内の Web ページに存在しているにも関わらず，利用者がその情報を発見することができないことが考えられる．

そこで本研究では，キーワード検索システムによって検索された Web ページから自動的にリンクを構築する手法を提案する．利用者は，検索された Web ページに構築されたリンクをたどることによって，検索システムに利用しているアルゴリズムを変更することなく利用者にとって必要な情報が得ることができる．

提案手法によって構築されるリンクの特徴として，利用者が入力するキーワードが利用者の検索要求と完全に一致していなくても，利用者の必要な情報を含む Web ページを得ることができる点が挙げられる．一般的に，利用者の検索要求をキーワードとして表現することは難しいと考えられるため，不適切なキーワードをシステムに入力した場合に必要な情報を得られない場合がある．ところが，提案手法では利用者が入力したキーワードと共に，全ての検索対象 Web ページにおける出現単語の偏りを考慮している．そのため，検索システムによって最初に検索された Web ページが利用者にとって不要なものであった場合でも，利用者がその後提案システムによって構築されたリンクのうち適切なリンクを選択することにより，利用者とって必要な情報を得ることができると考えられる．

提案手法の適用範囲として，10000 から 50000 ページ程度の Web ページを含む一つのサイトに適用することを考えている．提案手法は，検索対象となる全ての Web ページ相互の類似度をリンク構築

♡ 学生会員 立命館大学大学院理工学研究科
nakatani@coms.ics.ritsumei.ac.jp
♢ 正会員 立命館大学情報理工学部情報コミュニケーション学科
{yusuzuki, kawagoe}@is.ritsumei.ac.jp

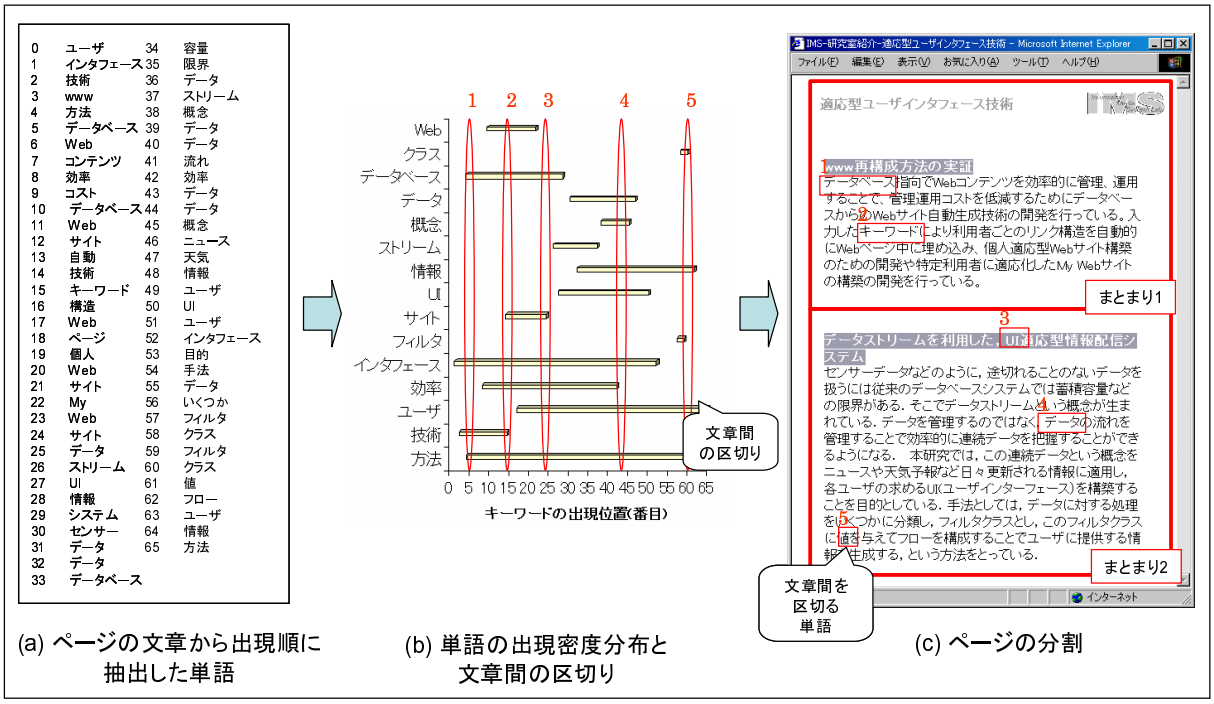


図 2: Web ページの分割例
Fig. 2 An example of dividing of Web pages.

に利用するため，$n$ ページを検索対象とした場合の計算量はほぼ $n^2$ となる．この結果，検索対象とする Web ページの増加にともなって，提案システムのために必要な計算量が大きくなる．そのため，一つのサイトを対象とした小規模な Web ページを検索対象とする．

## 2. Web リンクの自動生成方式

リンクを自動的に構築する手順は次の通りである．

1. 利用者は検索システムにキーワードを入力する．
2. キーワード検索システムを用いて，利用者の検索要求に適合した Web ページ $D$ を検索する．
3. $D$ を $M$ 個の意味単位に分割し，各部分ページを $d_1, d_2, \cdots, d_M$ とする．
4. $D$ に関連する Web ページを，検索対象集合から $N$ 個選択し，それぞれ $P_1, P_2, \cdots, P_N$ とする．
5. $d_i$ と $P_1, P_2, \cdots, P_N$ との類似度を計算し，最も高い類似度である $P_j$ をリンク先，$d_i$ をリンク元としたリンク $L(D_i, P_j)$ を構築する．
6. 5. を $M$ 回繰り返す．

つまり，利用者の入力したキーワードによって，$D$ から $M$ 個のリンクを作成する．ここで，$M$ はサイト管理者によってあらかじめ検索システムに与えるパラメータである．

リンク構築の手順のうち，3. の部分における Web ページを意味単位に分割する方法と，4. の部分における $D$ に関連する Web ページの作成手法は，それぞれリンク元，リンク先の選択のための重要な手順である．そこで，まず 2.1 節ではリンク元となる Web 部分ページの算出方法について述べ，次に 2.2 節では，リンク先の選択手法について述べる．最後に 2.3 節では，2.1 節，2.2 節で求めたリンク元，リンク先双方の選択からリンクを構築する手順について述べる．

### 2.1 リンク元の選択

リンクを自動的に構築するためには，まずリンク元を利用者に提示しなければならない．そこでまず，既存のキーワード検索システムによって検索された一つの Web ページを利用者に提示する．この時，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部分をリンク元とした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 Web ページを容易に閲覧することができると考えられる．しかし，Web ページは HTML で記述されているため，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，Web 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と考えられる．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 Web ページへのリンクを，リンク元の Web ページ中の適切な場所に配置するために，Web ページを意味単位に分割する．

以下にページの分割手順を述べる．

1. Web ページから茶筌 [4] を用いて Web ページに含まれる単語を出現順に抽出する．ここで，抽出する単語として名詞，形容詞，および未定義語として抽出された単語を用いる．
2. ページ中に含まれる単語ごとに出現位置の平均と標準偏差を算出する．
3. 各単語の出現範囲を求める．単語の出現範囲は，出現位置の平均を中心とし，標準偏差の 2 倍の幅を持った範囲とする．
4. 各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

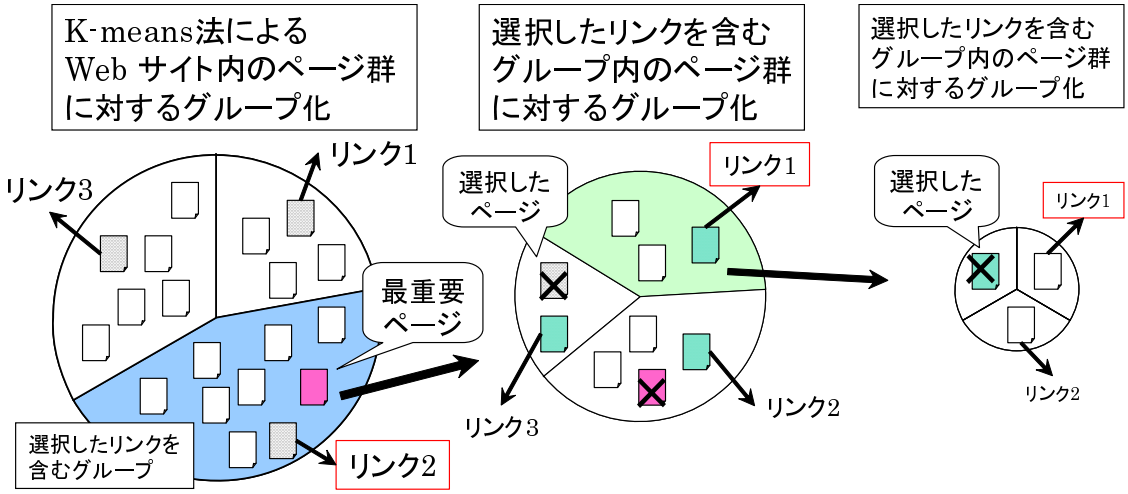


図 3: リンク先の選択手法
Fig. 3 A selection method of linked Web pages.

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．

5. 抽出した単語の次の，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特定のタグが存在する場合は，ひとつのまとまりとする．

ここで，抽出した単語で Web 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 2 う場合がある．そこで，抽出した単語の出現位置以降に出現する HTML タグのうち，文章を区切る場合に用いられる HTML のタグの位置で Web ページを分割する．分割に用いられる HTML タグは次の 7 種類であり，それぞれ <HR>, <H 数値 >, <P>, <SPAN>, <DIV>, <BR>, <TABLE> である．

あるページの分割例を図 2 に示す．図 2 の (a) は Web 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) の実際の部分ページ数が二つになっているのは，(c) の各部分ページの抽出した単語の次のタグが同一であるために，同じ部分ページになったためである．

## 2.2 リンク先の選択

リンク先の選択の際には，リンク元の Web ページが利用者にとって不要な Web ページである場合でも適切なリンク先を選択しなければならない．このため，リンク元 Web ページに内容が類似した Web ページはリンク先として適切ではない．つまり，リンク先の Web ページは検索対象 Web ページ全体を代表する Web ページである必要があると考えられる．そこで提案手法では，検索対象 Web ページ全体に対して分類を行い，それぞれのクラスタのうち任意の一つの Web ページをリンク先として選択する．

まず，K-means 法 [3] を用いて分類された Web ページ群の中から，最重要ページに対する関連ページを，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 2.1 節で述べるように分割された最重要ページに埋め込む．以降は，分類後の各 Web ページ群に対しても同様に分類を行い，グループ内の Web ページ数が一つになるまで分類を行い，リンクを自動生成していく．

利用者がシステムに入力したキーワードの出現頻度と，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．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．

## 2.3 リンクの構築

最後に，リンク元 Web 部分ページ $d_i$ とリンク先ページ $P_j$ との内容の類似度を求め，リンクの構築を行う．Web ページ間の内容の類似度は，ベクトル空間モデルによるコサイン相関値を用いる [1].

まず，$d_i$ と $P_j$ の類似度を求めるために必要な特徴ベクトルを計算する．単語の出現頻度による $d_i$ , $P_j$ の特徴ベクトル $\vec{F}(d_i)$, $\vec{F}(P_j)$ は，それぞれ (1), (2) 式で表される．

$$\vec{F}(d_i) = [w_1(d_i, w_2(d_i), \cdots, w_t(d_i), \cdots] \quad (1)$$

$$\vec{F}(P_j) = [w_1(P_j, w_2(P_j), \cdots, w_t(P_j), \cdots] \quad (2)$$

ここで，$w_t(d_i)$, $w_t(P_j)$ はそれぞれ単語 $w_t$ が Web ページ $d_i$, $P_j$ に出現する回数を表している．

次に，二つの Web ページの類似度を求めるために，二つの特徴ベクトルのコサイン相関値を求める．$d_i$ と $P_j$ の類似度 $S(d_i, P_j)$ は，(3) 式で表される．

$$S(d_i, P_j) = \frac{\vec{F}(d_i) \cdot \vec{F}(P_j)}{|\vec{F}(d_i)| \cdot |\vec{F}(P_j)|} \quad (3)$$

最後に，(3) 式で計算された類似度を基に，リンクの構築を行う．

リンクの構築は，ある $d_i$ において全ての $P_j$ との類似度を計算し，最も類似度高い $P_j$ へリンクを構築する．

以上の手順によって，自動的にリンクを構築することが可能となった．

## 3. まとめ

本研究では，ページ間の内容の類似度と，利用者が必要と考えられる Web 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．キーワード検索システムによる検索結果が利用者の検索要求に適合していない場合であっても，提案手法によって自動作成されたリンクをたどることによって，利用者にとって必要な Web ページを検索することが可能となる．

本研究の目的を達成するため，我々は以下二つの提案を行った．

- リンク元の選択手法として，単語の出現密度分布と HTML のタグ情報を用いた Web ページの分割を行った．その結果，検索結果ページが利用者の検索要求に適合していない場合に，利用者がどのようなリンクをたどれば良いかを適切に提示することができた．
- リンク先の選択のために Web ページ群全体の分類を行い，代表的なページへのリンクを行った．その結果，最初に利用者にとって必要ではない Web ページが検索された場合であっても，必要な Web ページまでのリンクを構築することができた．

本研究の今後の課題は次の通りである．

- 提案手法では，検索対象を分類するために K-means 法を用いるため，リンク数は常に固定である．そのために，利用者が全く要求しないページも同じグループに存在してしまう場合がある．それを解決するために，X-means 法 [2] など，クラスタ数を自動的に決定することが可能な分類アルゴリズムを利用する必要があると考えられる．
- 本論文における提案手法では，検索対象 Web ページ全ての組み合わせに対して類似度を計算する必要がある．そのため，計算量が増大し，大規模な Web ページ群を対象とすることができないという問題点がある．そこで，ある Web ページに対して類似度が上位 $k$ 件の Web ページを高速に検索する手法などを適用することによって，計算量を削減する必要があると考えられる．

### [文献]


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

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

[3] J. M. Queen. Some Methods for Classification and Analysis of Multivariate Observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281 – 297, 1967.

[4] 松本. 形態素解析システム「茶筌」. 情報処理, 41(11):1208 – 1214, 2000.


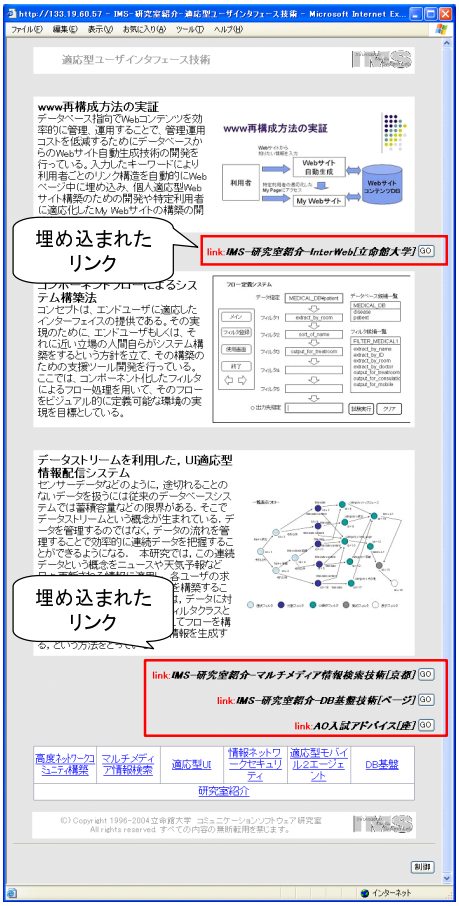


図 4: 提案手法によって作成されたリンクの例
Fig. 4 An example of generated Web links.


**中谷 圭吾　Keigo NAKATANI**

立命館大学大学院理工学研究科博士前期課程修了．現在，サントリー株式会社勤務．Web サービス，情報検索に関する研究に従事．日本データベース学会会員．

**鈴木 優　Yu SUZUKI**

立命館大学情報理工学部情報コミュニケーション学科講師．マルチメディア電子文書検索に関する研究に従事．ACM SIGMOD, IEEE Computer Society，情報処理学会，日本データベース学会各会員．

**川越 恭二　Kyoji KAWAGOE**

立命館大学情報理工学部情報コミュニケーション学科教授．情報システム，ネットワークサービス，データベースに関する研究に従事．情報処理学会，IEEE，ACM SIGMOD，日本データベース学会各会員．