九州大学学術情報リポジトリ
Kyushu University Institutional Repository

# WEB検索におけるキーワード関連語提案システム

大石，哲也
九州大学大学院システム情報科学府知能システム学専攻 ： 修士課程

長谷川，隆三
九州大学大学院システム情報科学研究院知能システム学部門

藤田，博
九州大学大学院システム情報科学研究院知能システム学部門

越村，三幸
九州大学大学院システム情報科学研究院知能システム学部門

https://doi.org/10.15017/1515921

出版情報：九州大学大学院システム情報科学紀要. 9 (1), pp.19-24, 2004-03-26. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

九州大学大学院
システム情報科学紀要
第9巻 第1号 平成16年3月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.9, No.1, March 2004

# WEB検索におけるキーワード関連語提案システム

大石哲也* ・ 長谷川隆三** ・ 藤田博** ・ 越村三幸**

# A System Proposing Words Relevant to Keywords in WEB Retrieval

Tetsuya OISHI, Ryuzo HASEGAWA, Hiroshi FUJITA and Miyuki KOSHIMURA

(Received December 12, 2003)

**Abstract:** Nowadays, we often use the Internet in order to get several information. Several search engines are available for retrieval. However, as the number of WEB pages increases, it becomes difficult to obtain useful information. Google is a reputable search engine with the accuracy of its retrieval result. It evaluates each WEB page with the PageRank system. This system has partially succeeded in presenting useful WEB pages. However, its retrieval result is still huge. If some appropriate keywords are added for retrieval, the number of obtained WEB pages will decrease. This paper presents a system proposing words relevant to given keywords in order to suppress the number of presented WEB pages.

**Keywords:** WEB retrieval, Search engine, Related word, TF/IDF, Co-occur word, Vector space model

## 1. はじめに

近年，事物を調べる際にインターネットを利用することは一般的であり，ユーザは主にサーチエンジンを用いて情報を得ている．しかしながら，現在世の中に存在しているWEBページはすべてを把握できないほど多くなっており，ユーザの必要とするページをなかなか見つけられないこともしばしばある．

検索結果の的確さで評判の良いサーチエンジンにGoogleがある．Googleは各ページのリンクに基づいたPageRankシステム[6),7)]を用いて各ページの評価を行う．この評価法は検索結果の上位にユーザの必要とするページを提示することに成功している．しかし，検索結果総数はやはり膨大であるので，ユーザが検索結果の上位に位置するページ以外のページを必要としているとき，ユーザは残りのページから必要なページを見つけなければならなくなって，ユーザにかなりの負担がかかるであろう．

そのような膨大なページを解析することでユーザの必要とするページを提示する研究も行われている．ところが，ページの量が多いと解析に時間がかかりすぎ，リアルタイムに結果を得たいユーザにとっては退屈なシステムとなってしまう．予めオフラインで任意の検索結果について解析しておいて，後から結果のみを利用するのであれば実用性もあるが，個々のWEBページは刻一刻と更新されており，過去の情報の解析だけでは満足しないユーザもいる．

検索結果の膨大さを解消するためには，サーチエンジンに与えるキーワードを増やすことが考えられる．一般にサーチエンジンに複数のキーワードを与えるとアンド検索を行うので，検索結果の量を減らすことができる．もしユーザが調べたい事柄に関連する複数のキーワードを適切に与えることができれば，サーチエンジンも適切なページをユーザに提示することができる．しかし，必ずしもすべてのユーザが調べたい事柄に関連するキーワードをサーチエンジンに与えることができるとは限らない．適切なキーワードを与えることができなければ，ユーザはいつまで経っても必要なページにたどり着かない．

そこで本論文では，ユーザが調べたい事柄に関連する語を提案するシステムを提示する．少量のキーワードをサーチエンジンに与えて得られた結果の上位に位置するページは，キーワードに関連するページではあるが，詳細まで一致しているページばかりではない．本システムは，それらのページをHTMLのタグ構造と単語の生起頻度に注目して解析して，ユーザが調べたい事柄に関連する語をリアルタイムに提案する．

本論文の構成は以下の通りである．まず2章では，HTML文書からタグ部分を取り去った部分を高速形態素解析システム「MeCab」[2)]を用いて解析し，TF/IDFおよび共起頻度を元に解析結果を評価する方法について述べる[5)]．3章では，HTMLのタグ構造に注目したタグベクトルと各ページにおける単語の生起頻度に注目したページベクトルを定義し，それに基づいて解析結果を評価する方法について述べる[3)]．4章でこれらの方法を用いて行った実験について述べ，5章で今後の課題に触れ，本論文をまとめる．

平成15年12月12日受付
* 知能システム学専攻修士課程
** 知能システム学部門

## 2. 単語の評価

本章では，ユーザが実際に見ているページの文章部分の各単語を評価する一般的な方法であるTF/IDF法と，共起頻度という概念について説明する．

### 2.1 TF/IDF法

TF/IDF法とは文書に含まれる単語の重み付けをする手法である．TF(Term Frequency)とは，ある文書$d$における単語$t$の生起頻度であり，$tf(t,d)$と表記する．またIDF(Inverse Document Frequency)は文書の数$N$と，単語$t$が一回以上生起する文書数$df(t)$によって次のように定義される．

$$idf(t) = \log \frac{N}{df(t)} \tag{1}$$

単語$t$の文書$d$における重み$w(t,d)$として，これらの積を用いるのがTF/IDF法である．

$$w(t,d) = tf(t,d) \cdot idf(t) \tag{2}$$

本論文ではユーザがキーワードをサーチエンジンに与えることで得られた検索結果に対してこの手法を用いる．しかし，得られたページ各々に対して単語の評価をするよりは，得られたページ群に対して単語の評価をした方がキーワードに関連した評価をすることができる．そこで，キーワード$k$に対する検索結果のページ群$P(k)$における単語$t$の生起頻度$tf(t,P(k))$と$idf(t)$の積を，単語$t$の検索結果ページ群$P(k)$における重み$w(t,P(k))$として用いることにする．

$$w(t,P(k)) = tf(t,P(k)) \cdot idf(t) \tag{3}$$

### 2.2 共起頻度

上記のTF/IDF法は単に文書中に現れる単語の数に注目した重み付けをする手法であり，一般的によく使われる手法である．さらに，別の評価指標として，各単語がどこに現れたのかを評価の対象とするために共起語(Co-occur Word)という概念を定義する．

ユーザが与えたキーワードに対して得られた文書には当然そのキーワードを含んでいるはずである．その文書内でキーワードの近傍に現れる単語は，キーワードに関連している単語であると考えられる．そのような単語のことを共起語と呼ぶことにする．キーワード$k$に対する検索結果のページ群$P(k)$におけるキーワード$k$の共起語$t$の共起頻度を$co(t,P(k))$とし，共起語$t$の生起頻度$tf(t,P(k))$を用いて，共起語$t$の共起確率(Co-occur Probability)$cp(t,P(k))$を定義する．

$$cp(t,P(k)) = \frac{co(t,P(k))}{tf(t,P(k))} \tag{4}$$

本論文では，これとTF/IDF法による評価値の積をキーワード$k$に対する検索結果のページ群$P(k)$における単語$t$の評価値$v(t,P(k))$とする†1．

$$\begin{aligned} v(t,P(k)) &= w(t,P(k)) \cdot cp(t,P(k)) \\ &= idf(t) \cdot co(t,P(k)) \end{aligned} \tag{5}$$

## 3. 単語の分類

本章では，各単語をタグベクトルとページベクトルという2つのベクトルで表現し，ベクトル空間モデルを利用して単語の分類をすることを考える．まず，ベクトル空間モデルの概要について説明し，その後にタグベクトルとページベクトルを定義し，実際にベクトル空間モデルを利用する方法について述べる．

### 3.1 ベクトル空間モデル

ベクトル空間モデルとは，ある情報を多次元空間上のベクトルとして表現し，2つのベクトルを比較することにより類似度を調べるものである．このモデルでは，2つのベクトルが同じ方向を指すときに類似度が高いとする．

具体的には2つのベクトルのなす角を$\theta$としたときの$\cos\theta$の値を類似度とする．つまり，2つのベクトルを$\vec{a}$と$\vec{b}$としたとき，類似度$sim(\vec{a},\vec{b})$は次式で求められる．

$$sim(\vec{a},\vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} \tag{6}$$

$\theta$の値が0に近いとき，すなわち$sim(\vec{a},\vec{b})$の値が1に近いとき，2つのベクトルは類似しているといえる．

### 3.2 単語のベクトル表現

本節では，ベクトル空間モデルを利用して単語を分類するための単語の2つのベクトル表現(タグベクトル，ページベクトル)について説明する．

#### 3.2.1 タグベクトル

我々が実際に見ているWEBページの実体はHTML(Hyper Text Markup Language)文書である．HTML文書は「<」と「>」で囲まれたタグを用いて記述される．タグによって特徴付けされた文章(単語群)は，WEBページの作者が強調したい部分である．

†1 この評価方法では，結果的に$tf(t,P(k))$の値が必要なくなってしまう．

**Table 1** 2-dimension classification.

| $v_{tag}(t,P(k))$ \ $v_{page}(t,P(k))$ | 1.0 | 0.0 |
|---|---|---|
| 1.0 | 1 | 2 |
| 0.0 | 3 | 4 |

ここで定義するタグベクトルは，あるキーワード$k$によって得られたページ群$P(k)$において，各単語がそれぞれのタグによって強調された頻度をベクトルで表現したものである．すなわちタグベクトルの次数は，タグの種類数に等しいことになる．実際は，めったに利用されないタグは評価対象としては相応しくないので，一般に利用されやすいタグにのみ注目してタグベクトルを生成することで次数を小さくすることができる．

#### 3.2.2 ページベクトル

ページベクトルとは，あるキーワード$k$によって得られたページ群$P(k)$において，各単語がそれぞれのページに現れた頻度をベクトルで表現したものである．すなわちページベクトルの次数は，$P(k)$に含まれるページ数に等しい．

### 3.3 ベクトル空間モデルの利用

タグベクトルとページベクトルにそれぞれベクトル空間モデルを利用することで，各単語をその使われ方によって分類することができる．

どちらのベクトル表現法においてもベクトル空間モデルを利用するには基準ベクトルを考える必要がある．ここでは基準ベクトルをキーワードとなっている単語$k$のベクトルとする．

キーワード$k$によって得られたページ群$P(k)$における単語$t$に対するタグベクトルとページベクトルのベクトル空間モデルによる評価値をそれぞれ$v_{tag}(t,P(k))$，$v_{page}(t,P(k))$とすると，それぞれの評価値によって大雑把に以下のように分けることができる(**Table 1**)．

1. $v_{tag}(t,P(k))$，$v_{page}(t,P(k))$ともに1に近い:
   単語$t$はキーワード$k$と同じようにタグ付けされ，同じページ内で使用されていることが多い．
2. $v_{tag}(t,P(k))$は1に近いが，$v_{page}(t,P(k))$は0に近い:
   単語$t$はキーワード$k$と同じようにタグ付けされているが，同じページ内ではあまり使用されていない．
3. $v_{tag}(t,P(k))$は0に近いが，$v_{page}(t,P(k))$は1に近い:
   単語$t$はキーワード$k$とは違ったタグ付けをされているが，同じページ内で使用されていることが多い．
4. $v_{tag}(t,P(k))$，$v_{page}(t,P(k))$ともに0に近い:
   単語$t$はキーワード$k$と違ったタグ付けをされ，同じページ内ではあまり使用されていない．

このように分類することで，TF/IDF法と共起頻度による評価だけでは下位にある単語でも，キーワード$k$と関係のある単語を抽出することが可能になる．

## 4. 実験および評価

### 4.1 実験の概要

実験はTF/IDF法，共起頻度，タグベクトル，ページベクトルを基に単語の評価および分類を行う．具体的な流れは次の通りである．

1. 検索目的の単語$k$をキーワードとする．
2. キーワード$k$をサーチエンジン「YAHOO! JAPAN」に与え，検索結果上位20件($P(k)$)を得る．
3. 得られたページを高速形態素解析器「MeCab」を用いて解析し，単語群を得る．
4. TF/IDF法，共起頻度を用いて各単語$t$の評価値$v(t,P(k))$を求める．
5. タグベクトルとページベクトルにベクトル空間モデルを使用して単語を分類する．それぞれの評価値$v_{tag}(t,P(k))$，$v_{page}(t,P(k))$を0.2ずつで区切って分類する．すなわち，各評価値を5分割することになり，単語を25個に分類することになる．
6. 分類ごとに評価値$v(t,P(k))$の高い単語を5件ずつ表示する．最大で125件の単語が表示される．

### 4.2 実験結果の評価

以下の項目について実験結果の考察を行う．

- いくつかのキーワードに対する関連語の出力例．
- キーワード「java」を基にシステムが分類して出力した関連語と，その評価値$v(t,P(k))$との関係．

#### 4.2.1 関連語の出力例

**Table 2**にいくつかのキーワードについて本システムを適用した結果の一部を示す．実際は最大で125個の関連語が出力されるが，**Table 2**にはその代表のみを示した．

**Table 2** Experimental results for some keywords.

| keyword | related word |
|---|---|
| java | プログラミング，Applets，プラットフォーム，動物 |
| MGTP | 定理，証明，推論，問題，長谷川，隆三 |
| 六本松 | 九州大学，福岡，国体，道路，西鉄バス，広島，廿日市 |
| イラク | フセイン，アメリカ，石油，クウェート，侵攻，査察 |

**Table 3** The details of the result for a keyword "java".

| $v_{tag}(t,P(k))$ \ $v_{page}(t,P(k))$ | 1.0 | | | | 0.0 |
|---|---|---|---|---|---|
| 1.0 | **開発(25,7)**<br>**プログラミング (20.8)**<br>サーバ(17.4)<br>VM(16.4)<br>システム(16.3) | 紹介(8.7)<br>改訂(8.2)<br>製品(4.6)<br>new(3.5)<br>運用(2.7) | はじめ(6.6)<br>やすい(6.6)<br>モデル(5.5) | | |
| | **Solution(164.2)**<br>**サイド(21.9)**<br>**基礎(19.9)**<br>**基本(19.9)**<br>知識(17.3) | 解説(18.6)<br>米国(16.4)<br>問題(11.6)<br>アーキテクチャ(8.6)<br>いま(8.6) | 連載(10.0)<br>フォーラム(8.6)<br>インター(6.6)<br>フェイス(6.6)<br>世界(5.5) | | |
| | **研究(30.1)**<br>**Developer(21.2)**<br>**lang(19.9)**<br>Platform(16.4)<br>Sun(16.2) | nbsp(13.9)<br>Standard(11.6)<br>v(10.0)<br>セキュリティ(8.0)<br>実験(6.6) | プラットフォーム(16.3)<br>今後(6.6)<br>文法(6.6)<br>指定(6.6)<br>追加(4.0) | | コミュニティ(10.0)<br>世界中(3.3)<br>多く(3.3) |
| | **FAQ(25.5)**<br>**House(23.3)**<br>Connection(13.7)<br>仕様(11.6)<br>org(10.9) | **Labo(25.9)**<br>**リリース(20.0)**<br>**Edition(19.2)**<br>Applets(17.3)<br>Free(13.3) | 記載(13.7)<br>バージョン(10.9)<br>場合(8.2)<br>詳しい(6.6)<br>tm(5.5) | | 配備(13.3)<br>携帯(13.3)<br>Exception(13.0)<br>Applet(13.0)<br>電話(10.0) |
| 0.0 | **Vol(47.5)**<br>**Press(38.3)**<br>**都会(30.3)**<br>**JavaFAQ(21.6)**<br>Technology(16.3) | **BOF(30.3)**<br>gimlay(13.0)<br>道(13.0)<br>JavaWorld(8.6)<br>JavaNews(8.6) | 動物(17.3)<br>Programming(10.0)<br>悩み(8.6)<br>法人(8.6)<br>BEST(8.6) | 覚え書き(8.6) | **アップデート (21.6)**<br>指摘(8.6)<br>IAjapan(8.6)<br>Staff(8.6)<br>Engineer(8.6) |

**Table 2**に示した結果には一見関連語としては相応しくない単語が含まれている．以下にそれぞれのキーワードについての考察を述べる．

**java** 一般的にJAVA言語が有名であるので，「プログラミング，Applets，プラットフォーム」は関連語として適した単語であると思われる．「動物」は「動物実験廃止を求める会(JAVA)」という団体が存在しているために出力されており，関連語として適した単語を出力していると言える．

**MGTP** MGTPはモデル生成法に基づいた定理証明器のことであり，その研究を行っている教授の名前を出力していることから関連語として適した単語を出力していると言える．

**六本松** 福岡県福岡市と広島県廿日市市に六本松という場所が存在しており，福岡の六本松周辺に関係している単語が出力されていることから関連語として適した単語を出力していると言える．

**イラク** イラクやその周辺地域で石油が取れ，2003年12月現在のイラク情勢を考えると，関連語として適した単語を出力していると言える．

ここに示したキーワード以外にも同様に関連語として適した単語を出力することに成功している．

### 4.2.2 関連語の分類と評価値

本節では，「java」というキーワードを基にシステムが分類して出力した関連語とその評価値$v(t,P(k))$との関係に注目して，本システムの有効性について考察する．

キーワード「java」に対してシステムが出力した結果の詳細を**Table 3**に示す．**Table 3**の中の()内の数値は2章で示した評価法を用いた評価値$v(t,P(k))$である．各マスの中には$v(t,P(k))$の高い上位5個の単語が表示されるが，$v(t,P(k))=0$である単語は予め除外してあるので，5個以下の単語が表示されている場合がある．さらに**Table 4**に評価値$v(t,P(k))$の高い上位20個の単語を示す．

なお，**Table 4**に含まれる単語は**Table 3**において太字で示している．

**Table 4** Top 20 words relevant to "java" according to $v(t, P(k))$.

| rank | word | rank | word |
|---|---|---|---|
| 1 | Solution(164.2) | 11 | サイド(21.9) |
| 2 | Vol(47.5) | 12 | JavaFAQ(21.6) |
| 3 | Press(38.3) | 13 | アップデート(21.6) |
| 4 | 都会(30.3) | 14 | Developer(21.2) |
| | BOF(30.3) | 15 | プログラミング(20.8) |
| 6 | 研究(30.1) | 16 | リリース(20.0) |
| 7 | Labo(25.9) | 17 | 基礎(19.9) |
| 8 | 開発(25.7) | | 基本(19.9) |
| 9 | FAQ(25.5) | | lang(19.9) |
| 10 | House(23.3) | 20 | Edition(19.2) |

解析対象となったページはキーワード「java」をサーチエンジンに与えて得られた上位20ページである．**Table 5**にそれらのページのタイトルを示す．プログラミング言語の「Java」に関するページが多く存在し，「Java Script」や「動物実験の廃止」に関するページも存在していることがわかる．

**Table 3**と**Table 4**を比較する．**Table 4**に含まれている20個の単語は**Table 3**における各分類に均等に現れておらず，偏って現れていることがわかる．すなわち，評価値$v(t, P(k))$だけでは偏った単語しか出力できない．また，分類によって**Table 4**には含まれなかった「動物」という単語を出力することができた．これは「動物実験の廃止を求める会JAVA」に関する単語であり，適切な関連語を出力していると言える．

#### 4.2.3 実験結果の利用

ここでは実験によって得られた関連語を利用して，サーチエンジンから得られるWEBページがどのように変わるのかについて説明する．キーワードを「java」，関連語を「動物」，目的ページを「動物実験の廃止を求める会JAVAに関する記述のあるページ」としたときの関連語を追加する前後の比較を**Table 6**に示す．

関連語を追加する前は総ページ数が874,000ページであったのが関連語を追加した後は17,800ページになり約2%までページ数が減少した．一方，検索結果の上位20ページ内の目的ページの数は，関連語を追加する前が1ページであったのに対して，関連語を追加した後は9ページになり効率良い検索が可能になった．

**Table 5** The title of top 20 pages obtained by the search engine with the keyword "java".

| rank | title |
|---|---|
| 1 | Java覚え書き |
| 2 | Java(tm) House Mailing List Homepage (in Japanese) |
| 3 | Sun Microsystems - Java Developer Connection |
| 4 | Sun Microsystems - Java テクノロジ |
| 5 | Nakagawa Masami Home Page with Java |
| 6 | Free Java Applets |
| 7 | @IT : Java Solution |
| 8 | Java FAQ |
| 9 | ページ移動のお知らせ |
| 10 | IAjapan Java(TM) 研究部会 |
| 11 | Java:What's New (Javaに関する最新ニュース) |
| 12 | Java 2 Platform, Standard Edition v 1.4 Overview |
| 13 | Platform, Standard Edition 1.3 |
| 14 | The Source for Java Technology |
| 15 | Javaなお子様的HP |
| 16 | JAVA Developer.jp |
| 17 | JAVA 動物実験の廃止を求める会 |
| 18 | Java-Da! |
| 19 | 1番役立つWWW Java Script |
| 20 | Microsoft Java 仮想マシンからの移行 |

**Table 6** Effect of adding a related word to the keyword.

| | before | after |
|---|---|---|
| The total number of all pages | 874,000 pages | 17,800 pages |
| The number of goal pages in top 20 pages obtained by the search engine | 1 page | 9 pages |

## 5. おわりに

本論文では，単語の使用頻度を基にした単語の評価と，単語の使われ方をタグベクトルとページベクトルという2つのベクトルで表現し分類することで，ユーザが調べたい事柄に関連する語を提案するシステムを提示した．

単なる単語の評価のランキングだけでは得ることのできない重要な単語を得ることができており，システムの有効性を確認できた．

しかし，現段階では，キーワードに関連する単語は出力できているものの，まったく関係のない単語が出力されているのも事実である．キーワードとして与える単語によっては，まったく意味のない結果が出力されてしまうこともある．また，出力された結果が必要であるか必

要でないかは，ユーザによって違ってくる．各ユーザが満足するために，ユーザに特化した結果を出力できるシステムを構築することが今後の課題である．

本システムはリアルタイムに結果を出力していたが，今後は結果を蓄えて学習させることも視野に入れている．学習機能を搭載することで，より精巧なシステムを構築できるであろう．

## 参 考 文 献


1) 砂山渡，野村勇治，大澤幸生，谷内田正彦，“Webページ検索におけるユーザの興味表現支援システム”，**電子情報通信学会論文誌**，D-I，Vol.J82-D-I，No.12，pp.1394-1402，1999.
2) “高速形態素解析システム「MeCab」version 0.77”，**奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座**，http://cl.aist-nara.ac.jp/~taku-ku/software/mecab/，2004.
3) 松田 能成，“重要単語の抽出とユーザの行動履歴を利用したメタ検索システムのレイティング手法”，**早稲田大学大学院理工学研究科修士論文**，2001.
4) 椰野憲克，山田誠二，“マルチWebロボットによるユーザの興味を反映した情報収集”，**電子情報通信学会論文誌**，D-I，Vol.J83-D-I，No.7，pp.780-788，2000.
5) 大澤幸生，ネルス．E. ベンソン，谷内田正彦，“KeyGraph：語の共起グラフの分割・統合によるキーワード抽出”，**電子情報通信学会論文誌**，D-I，Vol.J82-D-I，No.2，pp.391-400，1999.
6) 馬場肇，“Google の秘密 - PageRank 徹底解説”，http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html
7) Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd, “The PageRank Citation Ranking: Bringing Order to the Web”, *Technical Report, Stanford University*,1998