

Web ページの表層的な特徴を用いた手法情報の発見

野中 諒志[†] 湯本 高行^{††} 新居 学^{††} 高橋 豊^{††}

[†] 兵庫県立大学工学部

〒 671-2201 兵庫県姫路市書写 2167

^{††} 兵庫県立大学大学院工学研究科

〒 671-2201 兵庫県姫路市書写 2167

E-mail: [†]eo05c086@steng.u-hyogo.ac.jp, ^{††}{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp

あらまし 近年, Web 上の情報量の増大により, ユーザが情報を探す作業の負担が増加している. このような負担は, ユーザが何らかの手法情報を取得する場合などにみられ, この負担を取り除くには専門化した検索エンジンが必要になる. そこで本研究では, 「1, ... 2, ...」といったような手順を表す表現に着目し, 料理のレシピや OS のインストールなどの手法を掲載している Web ページを発見するアルゴリズムを提案する. また, ユーザの負担を削減する意味でわかりやすいものを提示するため, わかりやすさの 1 つの指標として, 画像と文字の効果の多さを考慮に入れたランキングのアルゴリズムも提案する. また, 手法掲載ページ発見の評価を行い, 適合率 78%, 再現率 65% の精度で手法掲載ページが発見できるという結果が得られ, 同時に, 手法情報に特化した Web ページ分割アルゴリズムの開発が必要であることがわかった. また, リランキング方法の評価も行い, 現在の方法では, 一般的に画像や文字の効果を用いずに説明される手法掲載ページには有効でないことがわかった.

キーワード 情報抽出, Web 情報検索, 手法情報

Discovering How-to Information using Surface Feature of Web Page

Ryouji NONAKA[†], Takayuki YUMOTO^{††}, Manabu NII^{††}, and Yutaka TAKAHASHI^{††}

[†] School of Engineering, University of Hyogo

2167 Syosya, Himeji, Hyogo, 671-2201 Japan

^{††} Graduate School of Engineering, University of Hyogo

2167 Syosya, Himeji, Hyogo, 671-2201 Japan

E-mail: [†]eo05c086@steng.u-hyogo.ac.jp, ^{††}{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp

Abstract Recently, increasing information on the Web makes it difficult to find information that users want to know. For example, when the user wants to obtain some how-to information, it is difficult to find them efficiently. To solve this problem, we need specialized search engines. In this paper, we propose an algorithm to find how-to information, such as cooking recipes, installation of OS, and so on. This algorithm focuses on procedure expressions, such as “1,... 2,...”. Furthermore, we propose a ranking algorithm to find how-to information which can be easily understood. In this algorithm, we use the quantity of images and decorated characters in the Web pages. Then, we evaluated our algorithms. On the evaluation of our algorithm to find how-to information, the precision is 78% and the recall was 65%. As a result, we found that we need an algorithm segmenting Web pages to find how-to information. From the evaluation of our ranking algorithm, we found that our ranking algorithm was not effective when we want to obtain how-to information which images were not important for.

Key words information extraction, Web search, how-to information

1. はじめに

近年, Web 上の情報量の増大により, 検索エンジンの重要性が増している. 現在は Google の PageRank [1] などの汎用的な

検索用途のためのアルゴリズムが数多く開発され, それらを用いた検索エンジンが広く使われている. しかし, 特定の目的に特化した情報を探する場合, それらの検索エンジンでは望むページを発見することが困難な場合がある. その 1 つとして, 手法

情報の検索がある．例えば，ユーザがカレーのレシピについて知りたい場合，Google のテキスト検索エンジンで「カレー」を検索した場合，カレーの作り方を説明しているサイトより，カレー風味のお菓子を販売している会社のサイトやカレー料理専門店のサイトのトップページが上位になることが多く，レシピサイトが見たいのであればユーザが「カレー レシピ」といったように付属語を付け加えなくてはならない上に，適切な付属語を知っていなければならない．またこの場合だと，ユーザにとって手法掲載ページを探すことと手法掲載ページの中でも見やすいページを見つけることの 2 つの手間が生じる．このような手間は，料理のレシピ以外にも OS のインストールや旅行の手続きの仕方など何らかの手法を調べるときに多く見られる．

このように手法情報を取得したい場合において，ユーザの負担を極力減らすために，手法掲載ページの検索方法を提案する．この手法掲載ページの検索方法は手法掲載ページを抽出する方法と抽出したページをリランキングする方法の 2 つが必要になる．そこで本研究では，この考えを実現するために，手法掲載ページの発見方法およびページのリランキング方法を提案する．

2. 関連研究

小久保らは，検索隠し味を用いて専門検索エンジンを構築する手法を提案している [2]．例えばユーザが牛肉を使ったレシピについて知りたい場合「牛肉」と既存の検索エンジンに入力しても，牛肉を使ったレシピが得られることは少ない．しかしここで「塩」というキーワードを追加で入力するとレシピが多く得られる．検索隠し味とは，ここでいう「塩」である．小久保らの手法では，クエリ拡張により専門検索エンジンを構築しているが，これでは OS のインストールや折り紙の折り方といったような多方面での手法掲載ページに対応するとすると，それだけ多様な付属語が必要となり，クエリに対する適切な付属語を決定するために対象ドメイン毎に学習が必要になってくる．このため他方面での手法掲載ページに対応することは困難になる．そこで本研究では，クエリ拡張だけではなく HTML 文書を構造として理解する考えも入れることで，あらゆる手法掲載ページを発見する．

Deng らは，Web ページの構造やページ内の意味のつながり等から Web ページをブロック化し階層化する VIPS というアルゴリズムを提案している [3]．このアルゴリズムでは各ブロックに対して DoC(Degree of Coherence) という内容がいかに関係性があるかという尺度が用意されている．その DoC の値が大きくなるにつれて内容が密集したものを表し，それに応じて葉ノードが多くなり，階層も深くなる．また，ユーザは閾値 PDoC(Permitted Degree of Coherence) を設定することができ，ページ内のすべての PDoC $\geq$ DoC となるノードに対して PDoC $<$ DoC になるまで分割を行う．本研究では手法掲載部分のみの評価と手法掲載ページの抽出でこのアルゴリズムを利用している．このアルゴリズムにより，Web ページ内の手法掲載部分の開始地点と終了地点が正確に判断できる．

手法掲載部分の抽出に関連して，武智らは HTML 文書の手順に関する箇条書きを抽出する手法を提案している [4]．この

システムでは，HTML 構造の $\langle OL \rangle$ タグまたは $\langle UL \rangle$ タグで囲まれた部分を抽出するという方法を用いている．しかし，この表現を用いた場合， $\langle OL \rangle$ タグまたは $\langle UL \rangle$ タグ以外の手順表現，手法掲載部分を抽出することはできない．そこで本研究では，正規表現により「1、...、2、...。」といった記述を発見したり，タグの多さや HTML の書き方などの一定パターンを発見し，手法掲載部分を取り出すという方法を採用する．これによりタグに依存せずに手法掲載部分を取り出すことができる．

3. 手法情報の発見

本研究では，ユーザがクエリを入力したときに，Web からクエリに関する手法掲載ページのみを抽出し，見やすい順にリランキングしたものをユーザに提示することを目的としている．また，本稿では，手法掲載ページの定義を順序に沿った説明をしており，最終的な結果を掲載しているページとする．

ユーザにリランキング結果を提示するまでの処理の流れを以下に示すと共に，概念図を図 1 に示す．

- (1) ユーザがクエリを入力し既存の検索エンジンから検索結果を取得
- (2) 手法掲載ページのみ抽出
- (3) 見やすい順にリランキングし，リストをユーザに提示

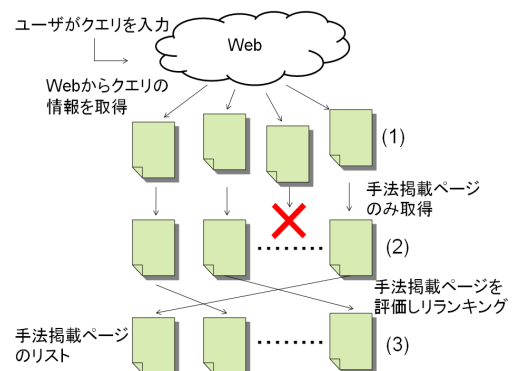


図 1 処理の概念図

3.1 手法掲載部分の抽出方法

3.1.1 手法掲載ページの分類

手法掲載ページには以下のように 3 つのパターンがある．

- パターン 1 $\langle LI \rangle$ タグや「1、...、2、...。」のように箇条書きや順序を付けて説明する場合
パターン 2 「まず～」や「次に～」のように言葉により順序づけがなされている場合
パターン 3 「～する。～する。」のように淡々と記述されて、説明がされている場合

以上のことより手法掲載ページのみを取り出すためには，以下の 3 つの基準に注目する必要がある．

- (a) 「1、...、2、...。」といった順序に沿った説明があるか
- (b) 動作や物事の経過を表す語が存在するか
- (c) 文末に過去形の表現が存在していないか

上記での (b) に当たる動作や物事の経過を表す語とは、表 1 に示すような語である。

表 1 動作や物事の経過を表す語

語句が現れる位置	語句
文中	～たら、～
	～後、～
文頭	始めに～
	次に～
	最後に～

また、(c) の過去形表現が存在していないかは、淡々と記述されている場合に限らず、全ての手法掲載ページにおいて見られることである。

3.1.2 手法掲載部分抽出のアルゴリズム

3.1.1 で述べた判断を行う前に、少しでも手法掲載ページのみを取り出せるようにクエリ拡張の面でも処理を行う。クエリ拡張は、順序に沿った説明、つまり手順表現の存在するページが手法掲載ページであると定義しているので、ユーザが入力したクエリに「手順」という言葉を自動で付け加える。

このクエリ拡張の作業を経て得られた結果に以下の処理を行うことで手法掲載ページのみを取り出す。また処理の概念図を図 2 に示す。

(1) VIPS を用いて Web ページをいくつかの細かいブロックに分け、正規表現 “[^。]*[。]” に一致する文章を含む文字列のあるブロック b を抽出する。

ただし、 b は $b \in B$ で表されるブロックで B は VIPS により分けられたページ内のブロックで、文章化された全てのブロックを表す。

(2) 注目したブロック b に、3.1.1 の (a) で示した基準「手順表現への注目」を行い、この表現があれば 1、(b) で示した基準「動作を表す語への注目」を行い、この表現があれば 1 のスコアとし、 $b \in B$ に対してスコア $\text{score}(b)$ を計算する。ただし、 $\text{score}(b)$ はブロック b のスコアを示す。

(3) 一定値以上の閾値 θ を持つブロック b' を全て取得する。ただし、 b' は $b' \in B'$ で表されるブロックで B' は以下のように表現できる。

$$B' = \{b | \text{score}(b) \geq \theta, b \in B\} \quad (1)$$

また、式 (1) の閾値 θ は、今回は 0.47 で設定している。

(4) 取得した全てのブロックの共通の親ブロックを取得する。
(5) 取得した共通の親ブロックに対して、3.1.1 の (c) で示した基準「過去形表現の存在」の判断を行い、取得されたブロックに「...た。」という表現がないかを確認する。

また、(5) で示した判断を行うのは、一般的に手法掲載ページの手法掲載部分には、過去形に当たる「...た。」表現はないからである。この表現に注目し、取り出されたブロックに「...た。」といった表現が文中に閾値 θ_2 以上ある場合、手法掲載部

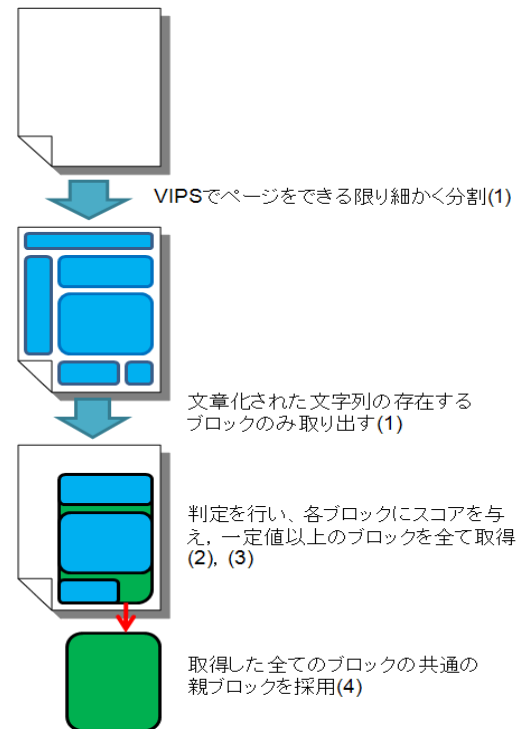


図 2 手法掲載部分の抽出法の概念図

分でない判断し、そのブロックのスコアを 0 にする。この場合もっともスコアの高いブロックは 0 となり、ページに手法掲載部分がないことになり、ページ自体を手法掲載ページでないと判断する。また、本稿では、閾値 θ_2 は、0.1 とする。

また、過去形表現がない場合でも処理の (4) で、スコアがどのブロックも 0 であるページは手法掲載部分がない、つまり手法掲載ページではないと判断する。

3.2 手法掲載ページのリランキング方法

手法掲載ページの記述がわかりやすいか否かを、以下に示す 3 つの情報を基準として考える。

- 手法に関する画像の量
- 手法に関する箇所太字や斜字、見出し文などの文字の効果の量
- 手法についての詳細度

手法掲載ページの発見方法により抽出されたブロックに対して評価を行う。具体的には、手法掲載部分のブロック内の画像の多さと文字の効果がかけている文字列の密度、手法掲載部分の文字数の 3 つの要素を足し合わせたものをページのスコアとし降順に並び替える作業を行う。

また文字の効果については表 2 のようなタグに注目して行う。

次に、画像のスコアと文字の効果のスコア、手法掲載部分の詳細度のスコアについて詳しく述べていくが、以下で述べる式 (2)、式 (3)、式 (4)、式 (5) において、評価の段階では、手法掲載ページから抽出した手法掲載部分は 1 つであり、その手法掲載部分にのみ式を適用させるので、手法掲載部分の評価値が直接、ページの評価値になる。つまり以下の式では手法掲載部分であるブロックの評価値がページの評価値であるとして式を

表 2 文字の効果

効果	表現
太字	
	
斜字	
	<i>
	<cite>
	<var>
アンダーライン	<u>
見出し文	<h1>
	<h2>
	<h3>
	<h4>
	<h5>
	<h6>

記述している．

まず，画像に関する評価値には，画像の個数を正規化したものを用い，画像数を正規化した数値は，式 (2) に示すように各数値を最大値で割って取得する．

$$img^*(p) = \frac{img(p)}{\max_{q \in P} img(q)} \quad (2)$$

P はリランキング対象のページ集合， $img(p)$ は，現在評価中のページの手法掲載部分に含まれている画像の個数， $\max_{q \in P} img(q)$ はページ集合 P 内で最大の $img(p)$ である．

次に，手法掲載部分内の効果がかかけられている文字列の密度を式 (3) のように定義する．

$$effect_density(p) = \frac{\sum_{e \in effect(p)} |e|}{|p|} \quad (3)$$

$|p|$ は，手法掲載部分の文字数， $effect(p)$ は，手法掲載部分での文字効果（太字や斜字，見出し文など）を受けている文字の集合， $|e|$ は文字効果を受けている文字の文字数である．

また，文字効果の評価値には，式 (3) で得られた密度を正規化したものを用い，その数値は，式 (4) で示すように各数値を最大値で割って取得する．

$$effect_density^*(p) = \frac{effect_density(p)}{\max_{q \in P} effect_density(q)} \quad (4)$$

P はリランキング対象のページ集合， $effect_density(p)$ は，現在評価中のページの手法掲載部分の文字効果のかかった文字列の密度， $\max_{q \in P} effect_density(q)$ はページ集合 P 内で最大の $effect_density(p)$ である．

次に，手法掲載部分の詳細度については，手法掲載部分の文字数を見ることで判断でき，評価値としては，式 (5) で示すようにページ集合内の文字数を最大値で割って正規化した数値を用いている．

$$characters^*(p) = \frac{characters(p)}{\max_{q \in P} characters(q)} \quad (5)$$

P はリランキング対象のページ集合， $characters(p)$ は，現在評価中のページの手法掲載部分の文字数， $\max_{q \in P} characters(q)$ はページ集合 P 内で最大の $characters(p)$ である．

これら 3 つの値を合わせたものを Web ページのスコアとし，スコアの降順でリランキングしたものをユーザに提供する．Web ページのスコアを式 (6) に示す．

$$score(p) = \alpha \cdot img^*(p) + \beta \cdot effect_density^*(p) + \gamma \cdot characters^*(p) \quad (6)$$

α, β, γ は $0 \leq \alpha, \beta, \gamma \leq 1$ の重みである．

4. 実 験

4.1 手法掲載ページ発見方法の評価

実験は，6 種類のクエリに対して行った．また，評価対象は，各クエリ毎に Yahoo! の検索エンジンの検索結果の上位 50 件に対して行う．この中の手法掲載ページの数，また提案手法により正確に手法掲載部分を抽出できたページの数と調べる．また実験方法としては，(a) が「1、...。2、...。」といった手順表現への注目，(b) が動作を表す語への注目，(c) が過去形表現の存在の確認として (b)，(c)，(a)+(b)，(b)+(c)，(a)+(c)，(a)+(b)+(c) のどの場合が有効であるかを確認するために各クエリに対して 7 つの場合において検証する．また本実験では VIPS の閾値 PDoC を 10 にして実験を行った．また，本実験において (c) のみの実験を行わない理由であるが，これは本来，提案手法 (c) が，提案手法 (a) ないし (b) を用いた後の手法掲載部分の候補となったブロックに対して適用するからであり，提案手法 (c) のみでは良い結果が得られないのは明らかだからである．

各クエリの実験結果の適合率を表 3 に，再現率を表 4 に示す．

表 3，表 4 より，提案手法の組み合わせにおいても，クエリ「ネクタイ 結び方」の結果が悪いことについては，各ページにおける手法掲載の仕方が現在の手法掲載ページの発見方法では，取得できないことが問題であることがわかった．具体的には，現在の提案手法は，HTML の文書に直接書かれた文字のみを判定材料としており，手法についての説明等が画像に書かれている場合には手法掲載ページとして判断できないということが問題となっている．このような画像に説明が書かれた手法掲載ページがクエリ「ネクタイ 結び方」の場合に多く見られた．これによりクエリ「ネクタイ 結び方」では手法掲載ページが，ほとんど取得できなかった．

また，適合率と再現率の両面が，提案手法を全て用いた場合（表 3，表 4 の (a)+(b)+(c) の場合）より過去形表現を注目する提案手法を除いた場合（表 3，表 4 の (a)+(b) の場合）の方が高い結果を得られているのは，VIPS による問題だということがわかった．具体的には，VIPS を用いて分割を行う場合，ページによっては細かく分割できない場合があり，大まかに分割される場合がある．ページが大まかに分割された場合，図 3 のように手法掲載ページ以外の部分も判定に加えることがある．もし

表 3 発見方法の実験結果 (適合率)

	使用した基準					
	(a)	(b)	(a)+(b)	(a)+(c)	(b)+(c)	(a)+(b)+(c)
クエリ						
肉じゃが	0.96	0.96	0.94	1.00	0.95	0.97
Fedora Core インストール	1.00	0.91	0.92	1.00	0.94	0.96
カレー	0.80	0.73	0.74	0.88	0.75	0.74
Cygwin インストール	0.83	0.88	0.86	0.81	0.87	0.84
留学 手続き	0.82	0.93	0.85	0.82	0.92	0.84
ネクタイ 結び方	0.42	0.23	0.35	0.40	0.20	0.38
適合率の平均	0.81	0.77	0.78	0.82	0.77	0.79

表 4 発見方法の実験結果 (再現率)

	使用した基準					
	(a)	(b)	(a)+(b)	(a)+(c)	(b)+(c)	(a)+(b)+(c)
クエリ						
肉じゃが	0.51	0.53	0.73	0.49	0.44	0.62
Fedora Core インストール	0.63	0.47	0.77	0.47	0.40	0.58
カレー	0.44	0.31	0.59	0.41	0.44	0.59
Cygwin インストール	0.56	0.49	0.70	0.51	0.47	0.63
留学 手続き	0.36	0.64	0.72	0.36	0.59	0.69
ネクタイ 結び方	0.26	0.16	0.37	0.21	0.11	0.37
再現率の平均	0.46	0.43	0.65	0.41	0.41	0.58

表 5 リランキング方法の実験結果

	並び替え	順位										得点
		1	2	3	4	5	6	7	8	9	10	
クエリ												
肉じゃが	前	×				○	○	○	○	×	×	11
	後	○	○	○	×	×					○	12
Fedora Core インストール	前			○		○	○		○			14
	後	○	○	○	○	○	○	○				17
カレー	前	○	○	×	○			×			○	12
	後	○	○	×	○				○			13
Cygwin インストール	前	○	○		○	○	○	○	×	○	○	17
	後	○	○		○		○	○	○	○	○	18
留学 手続き	前	○					○		○	×		12
	後	○		○			×	○				12
ネクタイ 結び方	前	○	○	○	○	○	○	○	○	×	○	18
	後	×	○	○	×	○	×	×	×	×	×	6

手法掲載部分のみを取得できていれば、文末に過去形表現が来ることがないので問題はないが、手法掲載部分以外も取れてしまう場合、つまり前述で述べた場合であると手法掲載部分以外も取得し、過去形表現が入ってくる場合がある。こういった場合に手法掲載ページでないと誤認してしまい、過去形表現を注目する手法を加えることで逆に結果が悪くなることがわかった。

4.2 手法掲載ページのリランキング方法の評価

実験は、数種類のクエリに対して行う。実験方法は、既存の検索エンジンの検索結果の上位 10 件とリランキング結果の上位 10 件を対象とし、それぞれのページに対し、見やすい手法掲載ページであるなら「○」で 2 点、見やすくないが手法掲載ページなら「△」で 1 点、手法掲載ページでないなら「×」で 0 点として得点を与える。既存の検索エンジンは Yahoo を用い、また、リランキング対象は既存の検索エンジンにより得られた上位 100 件を手法掲載ページの発見方法を用いて抽出さ

れたページである。また式 (5) で示した重みはユーザが自由に変更でき、本実験では $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.3$ とした。

各クエリにおける実験結果を表 5 に示す。

表 5 において、クエリ「留学 手続き」ではリランキング前と後では得点が変わっていない。これは、現在のリランキング方法が画像の多さと文字の効果の多さ、文字数に依存しているためであると考えられる。このクエリで現れる手法掲載ページのほとんどは、画像や文字の効果を使って説明することがなく、また簡潔に述べられるようなことも、ほとんどない。そのため現在のアルゴリズムでは、リランキングによる影響はほとんど出ることがない。この問題を改善するためには、ページの構成等を把握する必要があり、現在のリランキング方法では十分でないことがわかった。

また、表 5 のクエリ「ネクタイ 結び方」において、リランキングの前と後では後のほうが悪くなっている問題がある。こ

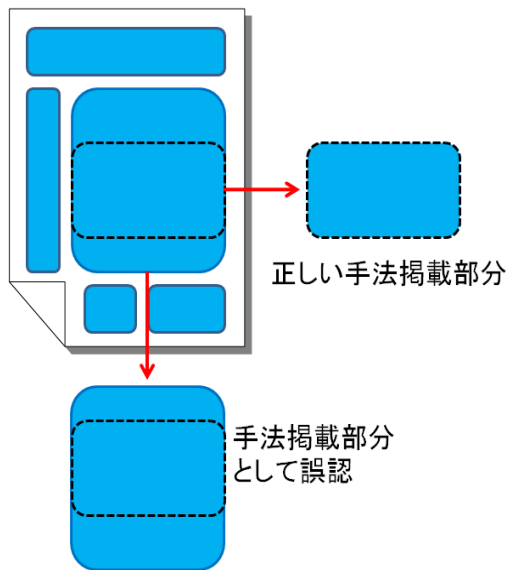
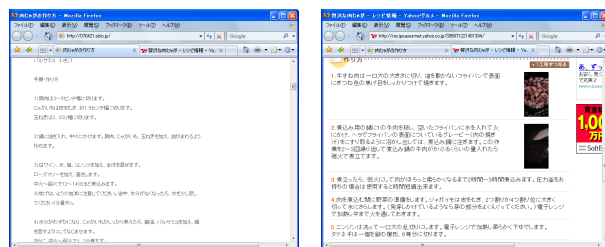


図3 VIPS による問題

これは、手法掲載ページを抽出する段階で問題が生じており正確に抽出できていないことが問題であり、手法掲載ページの発見方法に見直しが必要であることがわかった。また、手法掲載ページの発見方法で起きている問題については、手法掲載ページ発見方法の評価での考察にて説明したとおりである。

次に、手法掲載ページのリランキング前の第1位とリランキング後の第1位を比較する。図4にクエリ「肉じゃがの作り方」の場合、図5にクエリ「カレーの作り方」の場合、図6にクエリ「留学の手続き」の場合の出力例を示す。



(a) リランキング前 (b) リランキング後

図4 リンキング方法の実行例(肉じゃがの作り方の場合)

図4のリランキング前では、第1位にはブログで紹介されている肉じゃがのレシピページ^(注1)になっており、このページでは肉じゃがのレシピについては画像も使われず簡潔にしか述べられていない。しかし図4のリランキング後では、第1位は画像も多く使われ、わかりやすく解説された肉じゃがのYahoo!のレシピサイト^(注2)になっている。

また図5についても同様で、リランキング前の第1位には文字だけの解説しかないページ^(注3)がリランキング後の第1位

(注1): <http://070421.sb1o.jp/>

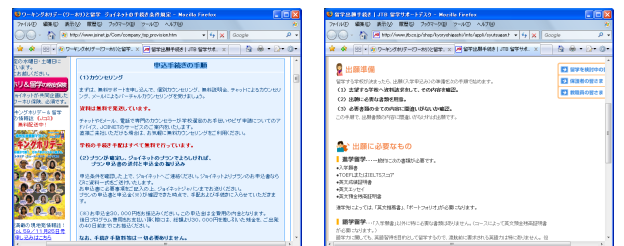
(注2): <http://recipe.gourmet.yahoo.co.jp/335971221481394/>

(注3): <http://yaplog.jp/mi9yu18ki/archive/400>



(a) リランキング前 (b) リランキング後

図5 リンキング方法の実行例(カレーの作り方の場合)



(a) リランキング前 (b) リランキング後

図6 リンキング方法の実行例(留学の手続きの場合)

では画像が使われた、わかりやすいレシピサイト^(注4)になっている。

表5の結果の考察でも述べたように、この図4と図5の出力例から、本稿での手法掲載ページのリランキング方法が有効である場合もあるが、図6のように基本的に画像や文字効果が多く用いられずに説明される手法については有効でないことがわかった。(注5)(注6)

5. おわりに

本稿では、見やすい手法掲載ページのリストをユーザに提供するために、Web上から手法掲載ページを発見する方法と手法掲載ページを見やすい順にリランキングする方法を述べた。また提案したアルゴリズムを用いて有効性を確認するために実験を行い、発見方法にて適合率78%、再現率65%という結果が得られ、VIPSの分割方法に問題があることがわかった。またリランキングにおいても実験を行い、多くのクエリにおいて現在のリランキング方法が有用であることがわかったが、同時に現在のリランキング方法では不十分なことがわかった。今後の課題は、手法情報に特化したWebページ分割アルゴリズムの開発とリランキング方法の見直しである。

文献

- [1] Sergey Brin, Lawrence Page, "The anatomy of a large-scale

(注4): <http://ameblo.jp/bikke3105/entry-10152413253.html>

(注5): http://www.joinet.jp/Com/company_top_provision.htm

(注6): <http://www.jtb.co.jp/shop/kyoryohigashi/info/appli/syutsugan.html>

- hypertextual Web search engine” ,Computer Networks and ISDN Systems , Vol.30 , Issue1-7 , pp.107-117 , (1998).
- [2] 小久保 卓, 小山 聡, 山田 晃弘, 北村 泰彦, 石田 亨, “検索隠し味を用いた専門検索エンジンの構築” , 情報処理学会論文誌 , Vol.43 , No.6 , pp.1804-1813 , (2002).
- [3] Deng Cai , Shipeng Yu , Ji-Rong Wen , Wei-Ying Ma , “VIPS: A VIsion based Page Segmentation Algorithm” , Microsoft Technical Report , (2003).
- [4] 武智 峰樹, 徳永 健伸, 松本 裕治, 田中 穂積, “WWW ページからの手順に関する箇条書きの抽出” , 情報処理学会論文誌 , Vol.44 , No.SIG 5 , pp.1-13 , (2003).