リンク解析に基づくWebページの理解容易性評価

赤松　弘一[†]　ニミットパッタナスリ[††]　アダムヤトフト[††]　田中　克己[††]

† 京都大学工学部情報学科　〒606–8501 京都市左京区吉田本町
†† 京都大学大学院情報学研究科社会情報学専攻　〒606–8501 京都市左京区吉田本町
E-mail: †{akamatsu,nimit,adam,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　調べたい事柄に対して分かりやすく解説されたWebページを検索したいというユーザの欲求を満たすためには，Webページの理解容易性を考慮したWeb検索エンジンが必要である．本稿では，Webページの理解容易性とWebのリンク構造との関係に着目し，実際にどのような関係が成り立っているかについて調査するために行った実験について記述する．さらに，そこで得た，Webページの理解容易性とリンクとの関係に関する知見を元に，リンク解析に基づくWebページの理解容易性の評価手法を提案する．

キーワード　理解容易性，リンク解析

1. はじめに

近年のWebの発展に伴い，人々がWebを利用して自分の知りたい物事について調べる機会が増えてきた．Web検索エンジンは，ユーザがWebを通じて効率的に情報を得るために役立つ．Web検索エンジンは与えたクエリに関連するWebページのリストを出力してくれるので，適切なクエリを入力すれば，調べたい物事についての解説が記載されたWebページを発見することは比較的容易になる．検索結果として出力するWebページのリストを決定する際に，従来のWeb検索エンジンでは，クエリとの適合度や，リンク解析に基づいて計算されるWebページの重要度などが考慮され，Webページの内容が理解しやすいかどうかは考慮されていない．そのため，Web検索エンジンが出力する検索結果のランキング上位に，理解するのが難しいWebページが多く現れることがあるが，そうした検索結果は分かりやすい解説を求めるユーザにとっては望ましくない．そこで，Web検索エンジンがユーザの望む理解容易性をもつWebページのみを検索結果として出力できるようにすることで，情報検索がより効率的，効果的なものになると考えられる．このような理解容易性を考慮したWeb検索を実現するためには，検索結果の候補であるWebページに対して，あらかじめ理解容易性の度合いに関する情報を付与しておかなければならない．そこで，Web上に存在する膨大な数のWebページについて，理解容易性を自動的に評価する手法が必要となり，本研究の目的は，そのようなWebページの理解容易性の評価手法を確立することである．

Webページの理解容易性を正しく評価できる手法が実現し，Web上に存在する大量のWebページに対して，理解容易性に関する情報を付与することができるようになれば，ユーザは自分の知識レベルに合った情報にアクセスすることが可能となる．このことを利用できるシステムは，理解容易性を考慮したWeb検索だけにとどまらない．例えば，Webを通じて自分の知識レベルからは少し高い理解容易性をもつ情報を取得することを，段階的に行うことができるようになる．つまり，Webを，広大な範囲の情報をカバーする学習システムとして利用することが可能になるかもしれない．

Webページの理解容易性の評価手法の候補として，文章を解析し，その分かりやすさを推定する手法が挙げられる．文章の分かりやすさを推定する手法はこれまでにいくつか提案されてきており，第2章において関連研究としてそれらを紹介する．

Webページの理解容易性を評価するために，Webページから文章を抽出し，抽出された文章についてその分かりやすさを推定するという方法が考えられる．しかしながら，Webページには文章だけでなく，表，画像，動画，音声などが含まれる場合があり，これらもWebページの理解容易性に影響を与える．また，デザインやレイアウトを工夫することによってWebページが見やすくなり，ユーザにとってWebページの内容がより理解しやすくなることもあり得る．したがって，Webページの理解容易性を評価するためには，Webページに含まれる文章のみを解析するだけでは不十分な場合があると考えられる．とはいえ，Webページ内の画像，動画やレイアウトなども含めた総合的な内容を解析するのは困難である．

本研究では，Webページの内容を解析するのではなく，リンク解析を行うというアプローチによってWebページの理解容易性を評価する手法を確立することを目指す．リンク解析というアプローチによって，Webページの理解容易性の，文章の読みやすさを推定するだけでは測れない部分も評価することを目的としている．

Webページの理解容易性とWebのリンク構造との関係について，我々は次のことが成り立っているのではないかという仮説を立てた．

- 理解容易なページから理解容易なページへのリンクは多い．
- 理解容易なページから理解困難なページへのリンクは少ない．
- 理解困難なページから理解容易なページへのリンクは少ない．
- 理解困難なページから理解困難なページへのリンクは多い．

この仮説が実際のWeb上で成立しているかどうかは分から

ないが，この仮説のような，Web ページの理解容易性とリンク構造との間の関係について知ることができれば，その関係に基づき，リンク解析によって Web ページの理解容易性を正しく評価する手法を確立することができる可能性があると我々は考えた．

本論文の構成は以下のようになっている．第 2 章では，関連研究について述べる．第 3 章では，Web ページの理解容易性とリンク構造との関係について調査するために行った実験とその結果を述べる．第 4 章では，Web ページの理解容易性評価のための提案手法について述べる．第 5 章では，提案手法の評価実験について述べる．最後に第 6 章において，本論文の結論を述べる．

2. 関 連 研 究

文章を解析し，その分かりやすさを判定しようとする研究がいくつかあり，以下においてそれらを紹介する．

まず，公式を用いるアプローチがある．これは，文章から必要な特徴量を抽出し，抽出した特徴量を公式に当てはめて，文章の可読性を表すスコアを算出するというものである．このような手法はいくつか存在し，抽出する特徴量は手法によって異なるが，文章に含まれる単語や文の長さがよく使用される．例えば，Flesch Reading Ease [1] では，文章中の 1 単語の平均の音節数および，1 文に含まれる単語の平均数という 2 つの特徴量を使用する．

Sato ら [2] は，日本語の文書の難易度を推定するプログラム「帯」を開発した．これは，教科書から抽出したコーパスを用いて難易度の規準とし，13 段階の各難易度に対する入力文書の尤度を，文字の生起確率に基づいて計算し，最大尤度をもつ難易度を出力するというものである．

以下の研究では，文書の内容の専門性を考慮して理解容易性を評価しようとしている．

Yan ら [3] は，米国国立医学図書館の管理する医学用語集である MeSH の情報を解析し，文書中に含まれる単語に着目して，医学に関連する文書の読みやすさを判定する手法について述べている．この手法は医学という特定の専門領域のみにしか適用できないが，以下の 2 つの手法は様々な専門領域に適用できるものとなっている．

中谷ら [4] [5] [6] は，先ほど述べた日本語文書の難易度推定システム「帯」を利用するとともに，Wikipedia を解析して専門用語を抽出し，文書中に専門用語が どの程度含まれるかを考慮して理解容易性の評価を行う手法を提案した．

Zhao ら [7] の手法では，まず，ひとつの専門領域に関連する文書と用語を集める．その後，難しい専門的な用語を多く含む文書は分かりにくく，逆に，分かりにくい文書に現れる用語は難しいという考えに基づいた手法を適用することにより，集められた文書と用語に対して，それぞれ可読性と難しさを評価する．

以上のように，文章の理解容易性を評価する手法はいくつか提案されているが，我々の研究は Web ページ内の文章を解析することによってではなく，Web ページの内容分析は行わずにリンク構造に基づいて理解容易性を評価するというアプローチをとっているという点において異なる．

3. Web ページの理解容易性とリンク構造との関係の調査

本研究の目的は，Web ページの理解容易性を評価する手法を確立することであり，そのためにリンク解析というアプローチをとる．リンク解析によって Web ページの理解容易性を評価する手法を確立するには，前提知識として，Web ページの理解容易性と Web のリンク構造との関係について正しい認識が必要である．

本章では，Web ページの理解容易性とリンク構造との関係について調査するために行った実験について述べる．我々は，3.1 節に述べる方法によって，与えたキーワードに関連する Web ページからなる Web ページ集合を生成し，実験のためのデータセットとした．このデータセットに対して，Web ページの理解容易性とリンク構造との関係について調査するための実験を，2 種類行った．2 種類の実験は，それぞれ英語の Web ページ，日本語の Web ページを対象としたものであり，3.3 節，3.4 節において，各実験について述べる．

3.1 Web ページ集合の作成

我々は，Web ページの集合を作成し，実験のためのデータセットとした．以下に述べる手順に従ってデータセットを構築すれば，与えたキーワードに関連する Web ページからなる集合を得ることができる．後に述べるいくつかのキーワードのそれぞれに対して，Web ページ集合を以下のように作成した．

まず，Web 検索 API を提供する，Yahoo! Search BOSS(注1) を利用して，キーワードに対する検索結果上位 N 件の Web ページをダウンロードする．さらに，こうして得られた検索結果ページのそれぞれについて，inlink ページおよび outlink ページを取得する．なお，inlink ページの URL は，こちらも Yahoo! Search BOSS の提供する API（BOSS Site Explorer API）を利用して取得した．また，outlink ページの URL は，検索結果ページの内容を解析することで取得した．以上の手順でダウンロードした検索結果ページ，inlink ページ，outlink ページを，キーワードに対する Web ページの集合とした．

つまり，実験のためのデータセットとしては，図 1 に表されるような Web ページが対象となる．

こうして作成した Web ページ集合に対して，Web ページの理解容易性とリンク構造の関係を調べるための実験を行った．

3.2 データセットの Web ページに対する理解容易性の評価

前節によって得られた Web ページ集合に対して，集合内の Web ページの内容を解析すれば，それぞれの Web ページがどの Web ページにリンクを張っているかがわかり，Web ページ集合におけるリンク構造を知ることができる．

データセットの Web ページ集合に対して，Web ページの理解容易性とリンク構造との関係を調べるためには，リンク構造の情報と合わせて，さらに集合内の Web ページの理解容易性

(注1)：http://developer.yahoo.com/search/boss/

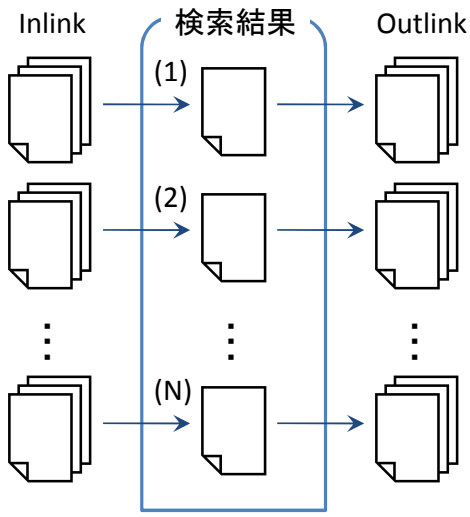

図 1　検索結果ページとその inlink ページおよび outlink ページ

を知る必要がある．Web ページの理解容易性を調べる方法として，次の 2 つが考えられる．

（1） 人手による評価．人が実際に Web ページを見て，その理解容易性を判定する．

（2） コンピュータによる自動化が可能な既存の手法を Web ページに適用する．例えば，Web ページから文章を抽出し，抽出した文章に対して，第 2 章に述べたような文章の分かりやすさを推定する手法のいずれかを適用する．

人手による評価では，大量の Web ページを評価するのに多くの時間や労力が必要だという問題がある．コンピュータを使えば大量の Web ページを素早く処理することができるが，人が感じる分かりやすさからは乖離した結果が含まれてしまうことがある．

3.3　英語の Web ページを対象とした実験

表 1 に示す 10 個のキーワードのそれぞれについて，以下に述べる手順で実験を行った．

表 1　実験に用いたキーワード

Alzheimer's disease, Parkinson's disease,
bipolar transistor, quantum computer,
comparative advantage, derivative, complex number,
mitochondrion, black hole, Halley's comet

まず，3.1 節で述べた手順によって，キーワードに対する Web ページ集合を作成した．このとき，検索結果ページについては，英語で書かれた Web ページを対象にしたので，検索結果ページとその inlink ページおよび outlink ページからなる Web ページ集合には，英語で書かれた Web ページが多いことが期待できる．また，今回の実験では，検索結果ページは検索結果の上位 30 件のページを用い，inlink ページおよび outlink ページは，各検索結果ページにつき最大 50 件まで取得した．

次に，Web ページ集合内の Web ページの HTML 文書からタグ等を除去して得られたテキストに対して，文章の可読性の評価手法である，Flesch Reading Ease テストを適用した．Flesch Reading Ease テストでは，文章を解析して，文章中の 1 単語の平均の音節数および，1 文に含まれる単語の平均数という 2 つの特徴量を抽出し，それらを次の公式に当てはめて計算を行うことで，その文章の可読性を表すスコアを得る．

$$206.876 - 1.015ASL - 84.6ASW \qquad (1)$$

ただし，ここで ASL，ASW はそれぞれ，1 文に含まれる単語の平均数，1 単語の平均の音節数を表す．

また，Wikipedia によると(注2)，Flesch Reading Ease テストによって算出されるスコアは表 2 のように解釈される．

表 2　Flesch Reading Ease テストにおけるスコアの解釈 (注2)

Score	Notes
90.0—100.0	easily understandable by an average 11-year-old student
60.0—70.0	easily understandable by 13- to 15-year-old students
0.0—30.0	best understood by university graduates

Web ページ集合内の各 Web ページについて，その HTML 文書からタグ等を除去して得られたテキストに対して，Flesch Reading Ease テストによる可読性のスコアを算出した．

こうして算出されたスコアの，検索結果ページのスコアとそのページに対する inlink ページのスコアの平均との相関係数，検索結果ページのスコアとそのページに対する outlink ページのスコアの平均との相関係数を，表 3 に示す．相関係数は-1 から 1 までの実数値をとり，1 に近ければ正，-1 に近ければ負の相関があり，0 に近ければ相関が弱い．表 3 のとおり，すべての要素について正の値が得られた．キーワードが「comparative advantage」のときは，検索結果ページのスコアとその inlink ページのスコアの平均との相関係数も，検索結果ページのスコアとその outlink ページのスコアの平均との相関係数も，0 に近い値になったが，その他のキーワードに関しては概ね高い相関が得られた．

表 3　相 関 係 数

	検索結果 と inlink	検索結果 と outlink
derivative	0.7003	0.6493
black hole	0.6217	0.6218
Parkinson's disease	0.5873	0.6473
mitochondrion	0.5170	0.4594
Alzheimer's disease	0.4904	0.6150
quantum computer	0.4891	0.7344
complex number	0.4647	0.6654
Halley's comet	0.3883	0.3746
bipolar transistor	0.3081	0.7669
comparative advantage	0.1416	0.0349
平均	**0.4708**	**0.5569**

また，次のような実験も行った．

算出されたスコアにしたがい，表 2 の解釈をもとに，Web ページ集合内の Web ページを次のように分類した．

(注2)：http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test

- Easy: スコアが 60 から 100 であった Web ページ.
- Medium: スコアが 30 から 60 であった Web ページ.
- Difficult: スコアが 0 から 30 であった Web ページ.

そして，この Web ページの分類に基づいて，inlink ページから検索結果ページへのリンク，検索結果ページから outlink ページへのリンクについて，次のように分類した．あるリンクが，Web ページ A から Web ページ B へのリンクであるとする．このとき，Web ページ A および Web ページ B の分類に従って，そのリンクの分類を (Web ページ A の分類)→(Web ページ B の分類) と決定した．たとえば，Web ページ A の分類が Easy，Web ページ B の分類が Medium であれば，Web ページ A から Web ページ B へのリンクは Easy→Medium と分類される．Web ページの分類が Easy，Medium，Difficult の 3 種類あるので，リンクの分類は全部で 9 通りとなる．

実験結果を表 4 に示す．表の行がリンク元ページの分類を，列がリンク先ページの分類を表す．表の各要素は，分類されたリンクの，各分類に対するリンクの数を示す．ただし，括弧内は，リンク元の Web ページの分類がその行の表す分類であった場合に，リンク先の Web ページの分類がその列の表す分類であった割合を示す．例えば，実験結果において，Easy→Difficult の分類にあるリンクの数は 316 であったことが表から読み取れる．また，表の Easy→Difficult の要素について，括弧内が 4.68%となっているので，Easy に分類された Web ページからのリンクは，4.68%の割合で Difficult に分類された Web ページへのリンクであったことがわかる．なお，表 4 は 10 個の各キーワードに対する実験結果を合算したものである．

表 4　実験におけるリンクの分類結果

		リンク先ページの理解容易性		
		Easy	Medium	Difficult
リンク元ページの理解容易性	Easy	3594 (53.2%)	2840 (42.1%)	316 (4.68%)
	Medium	3233 (28.8%)	6381 (56.8%)	1630 (14.5%)
	Difficult	355 (18.8%)	903 (47.8%)	631 (33.4%)

この実験結果によると，ある Web ページ A が別の Web ページ B にリンクを張っていた場合，Web ページ A の分類と Web ページ B の分類とには，ある程度の相関が見られることが分かる．表 4 を見ると分かる通り，Web ページ A の分類が Easy であった場合，リンク先の Web ページ B の分類は Easy である確率が最も高く，Web ページ A の分類が Medium であった場合も同様に，Web ページ B の分類が Medium である確率が最も高いという結果になった．リンク元の Web ページ A の分類が Difficult であった場合のみ結果が異なり，Web ページ B の分類は Diffucult ではなく Medium が最も多かった．

Web ページの理解容易性と Web のリンク構造との関係について，我々が以下の仮説を立てたことは既に述べた．

- 理解容易なページから理解容易なページへのリンクは多い.
- 理解容易なページから理解困難なページへのリンクは少ない.
- 理解困難なページから理解容易なページへのリンクは少ない.
- 理解困難なページから理解困難なページへのリンクは多い.

Easy に分類された Web ページを理解容易なページ，Difficult に分類された Web ページを理解困難なページとみなすと，この仮説が成立しているように見える．理解容易なページからは，理解困難なページへのリンク（Easy→Difficult （4.68%））よりも，理解容易なページへのリンク（Easy→Easy（53.2%））のほうが多い．また，逆に，理解困難なページからは，理解容易なページへのリンク（Difficult→Easy（18.8%））よりも，理解容易なページへのリンク（Easy→Easy（33.4%））のほうが多い．

第 1 章で述べたように，Web ページには，文章だけでなく，画像，動画，音声なども含まれ，Web ページの理解容易性はそれらに加えてレイアウトやデザインなども含めた複合的な要因によって決定されると考えられる．しかし，Web ページの理解容易性と，Web ページから抽出された文章の可読性の間にも，一定の相関が存在すると考えられる．今回の実験では，Web ページから抽出した文章の可読性のスコアとリンク構造との間に相関がみられ，Web ページの理解容易性とリンク構造との間にも，相関が存在するという示唆を与えるものといえる．

3.4　日本語の Web ページを対象とした実験

表 5 に示す 10 個のキーワードのそれぞれについて，以下に述べる手順で実験を行った．

表 5　実験に用いたキーワード

複素平面 脂肪酸 アルツハイマー病 筋ジストロフィー
ページランク ブラックホール ドップラー効果
ストックオプション シナプス ガイア理論

まず，3.1 節で述べた手順によって，キーワードに対する Web ページ集合を作成した．今回の実験では，検索結果ページについては，日本語で書かれた Web ページを対象にしたので，検索結果ページとその inlink ページおよび outlink ページからなる Web ページ集合には，日本語で書かれた Web ページが多いことが期待できる．

次に，Web ページ集合内の Web ページの HTML 文書からタグ等を除去して得られたテキストに対して，第 2 章で述べた，日本語文書の難易度を推定するプログラム「帯」(注3) のスコアを算出した.「帯」のスコアは 1 から 13 までの整数値のいずれかであり，数値は次のように学年を表す.

- 1 - 6：小学（1 年 - 6 年）
- 7 - 9：中学（1 年 - 3 年）
- 10 - 12：高校（1 年 - 3 年）
- 13：大学

Web ページ集合内の 2 つのページ A，B に対して，2 つのページに付与された「帯」のスコアの差 $ScoreDistance(A, B)$ を以下の式により定義した．

(注3)：http://kotoba.nuee.nagoya-u.ac.jp/sc/readability/obi.html

$$ScoreDistance(A,B) = |Score(A) - Score(B)| \qquad (2)$$

ここで，$Score(A)$，$Score(B)$ はそれぞれ，ページ A，ページ B に付与された「帯」によるスコアを表す．例えば，$Score(A) = 4$，$Score(B) = 7$ であった場合，$ScoreDistance(A,B) = |4-7| = |-3| = 3$ と計算される．「帯」のスコアは 1 から 13 までの整数値のいずれかなので，$ScoreDistance$ は 0 から 12 までの整数値のいずれかとなる．

Web ページ集合内のリンクについて，リンク元ページとリンク先ページとの $ScoreDistance$ の値を計算し，その値に従ってリンクを分類し，各分類のリンクが何個あるかを数えた．ただし，今回の実験ではサイト内リンクについては考慮せず，サイト外リンクのみを数に含めた．これは同一サイト内の Web ページの理解容易性のレベルは似ているだろうという考えに基づく．なぜなら，同一サイト内の Web ページは，同一の作者によって作成された可能性が高いと考えられるからである．

実験結果を表 6 に示す．なお，結果は 10 個のキーワードのそれぞれに対する Web ページ集合について実験を行った結果を合算したものである．

表 6　実験におけるリンクの分類結果

$ScoreDistance$	リンクの数	割合（%）
0	112433	39.74
1	72183	25.51
2	36832	13.02
3	32967	11.66
4	17477	6.18
5	4822	1.70
6	3172	1.12
7	2820	0.99
8	134	0.0474
9	9	0.00318
10	10	0.00353
11	18	0.00636
12	17	0.00601

表 6 によると，$ScoreDistance$ の値が小さいほどリンクの数が多い．これは，付与した「帯」のスコアが近い Web ページ同士がリンクを張っていることが多いことを意味する．例えば，同じ「帯」のスコアをもつ 2 つのページ間によるリンクの数（$ScoreDistance$ の値が 0 であるリンクの数）は，すべてのリンクの数のほぼ 40%と，非常に高いことがわかる．「帯」のスコアの差が 1 以内であるページ間のリンクの数とすると，全体のほぼ 3 分の 2 に達している．

Web ページに付与された「帯」のスコアが，その Web ページの理解容易性をよく表しているとすれば，この実験結果から，同じような理解容易性をもつ Web ページどうしがリンクしあっていることが多く，Web ページの理解容易性とリンクとの間に強い相関があるといえる．

4. Web ページの理解容易性の評価手法

本章では，前章において得られた，Web 上のリンクにおけるリンク元ページの理解容易性と，リンク先ページの理解容易性の関係についての知見を基にした，理解容易性の評価手法を提案する．

我々の提案手法は TrustRank アルゴリズム [8] を元にしている．

本章の以下では，まず，TrustRank アルゴリズムについて説明し，その上で提案手法について述べる．

4.1 TrustRank アルゴリズムの概要

TrustRank は，スパムページからスパムでない良質なページをより分けるためのリンク解析アルゴリズムである．TrustRank は，「スパムでないページからスパムページへのリンクはめったにない」という経験的観測に基づいている．この観測によれば，スパムでないページにリンクされているページもまた，スパムでない可能性が高い．TrustRank では，まず，ページ集合の中からいくつかページをシードとして選択し，それらがスパムであるかどうかを人手で判定しておく．シードページのうち，スパムでないと判定されたものから，biased PageRank アルゴリズムによって，リンク構造に従ってスコアを伝播させることで，ページ集合全体の各ページについてスパムでなさそうな度合いをスコアとして算出する．TrustRank は，シードページについては人手でチェックする必要があるが，その他の操作をコンピュータで自動化できる，半自動的な手法であるといえる．

各ページの biased PageRank スコアを要素にもつベクトル r は次のように定義される．

$$\mathbf{r} = \alpha \cdot \mathbf{T} \cdot \mathbf{r} + (1-\alpha) \cdot \mathbf{d} \qquad (3)$$

ここで，α はダンピングファクター，$\mathbf{T}$ はページ間の遷移行列，$\mathbf{d}$ は要素がすべて非負かつ要素の総和が 1 であるベクトルである．通常の PageRank では，$\mathbf{d}$ の要素はすべて等しく，ページ数を N とすると $1/N$ となる．

TrustRank では，biased PageRank アルゴリズムにおいて，スパムでないと判定された良質なページに対応する $\mathbf{d}$ の要素にのみ正の値を与え，他の要素を 0 とすることによって，良質なページからスコアを伝播させる．

4.2 TrustRank の Web ページの理解容易性の評価手法への応用

TrustRank はスパムでないページからスパムページへのリンクはめったにないという経験的観測に基づいているが，逆に，スパムページからスパムでないページへのリンクが少ないということは仮定していない．Web ページがスパムであるかどうかということと，リンクとの関係を，表 7 に示す．表において，G はスパムでないページ，B はスパムページを表す．また，→ はリンクを表す．例えば，「G→B　少ない」は，スパムでないページからスパムページへのリンクが少ないことを表す．

表 7　Web ページがスパムかどうかとリンクとの関係

G→G　多い
G→B　少ない
B→G　多い
B→B　多い

TrustRank アルゴリズムでは，スパムでないページからリンク構造に従ってスコアを伝播させる．表 7 に示した，Web ページがスパムかどうかとリンクとの関係によると，スパムでないページから伝播させたスコアは，スパムでないページに流れることが多いといえる．ところが，逆にスパムページからスコアを伝播させても，スパムページからスパムでないページをより分けるという目的には役に立たない．表 7 によると，スパムページから伝播されたスコアは，スパムでないページとスパムページのどちらにも流れやすいからである．

表 7 の関係の代わりに，Web ページの理解容易性と Web のリンク構造との関係が得られたとき，理解容易なページから理解容易なページへのリンクが多く，理解容易なページから理解困難なページへのリンクが少なかったと仮定する．このとき，スパムでないページの代わりに，理解容易なページから biased PageRank アルゴリズムによってスコアを伝播させれば，理解容易なページにスコアが流れやすいといえる．逆に，理解困難なページから理解困難なページへのリンクは多いが，理解困難なページから理解容易なページへのリンクが少なかったと仮定すると，理解困難なページから biased PageRank アルゴリズムによってスコアを伝播させれば，理解困難なページにスコアが流れやすいといえる．

第 3 章に述べた実験結果は，Web ページの理解容易性とリンク構造との間に強い相関があり，Web ページは自身と同等の理解容易性をもつ Web ページにリンクを張ることが多いということが窺えるものであった．

提案手法は，TrustRank アルゴリズムを応用し，スパムでないページの代わりに，理解容易なページまたは理解困難なページから，biased PageRank によってスコアを伝播させることで，理解の容易さを表すスコアまたは理解の困難さを表すスコアを計算するというものである．

提案手法を適用した場合，「理解容易なページから理解容易なページへのリンクは多く，理解容易なページから理解困難なページへのリンクは少ない」または「理解困難なページから理解困難なページへのリンクは多く，理解困難なページから理解容易なページへのリンクは少ない」という前提条件が成立していれば，Web ページの理解容易性を正しく評価できると考えられる．

5. 評　　価

提案手法の評価実験のためのデータセットとなる Web ページ集合は，3.4 節に述べたものと同じもの，表 5 に示す 10 個のキーワードのそれぞれについて，3.1 節で述べた手順によって作成したものである．

TrustRank では，Web ページの集合に対して Inverse PageRank アルゴリズムによって算出されたスコアの高いものをシードページとして選択し，その中からスパムでないページを人手で選び，スコアの伝播元となるページを決定する．それに対して，我々は，今回の実験において，「帯」によって計算されたスコアに基づいて，スコアの伝播元となるページを選択した．

つまり，今回の実験において，我々は以下の仮定をおいている．

- 「帯」のスコアの低いページは理解容易性が高い．
- 「帯」のスコアの高いページは理解容易性が低い．

4.2 節で述べた考えに従って，理解容易なページからスコアを伝播させる場合と，理解困難なページからスコアを伝播させる場合の 2 通りについて，提案手法がうまくいくかどうか評価したい．そこで，「帯」のスコアが低い Web ページをスコアの伝播元とした場合と，「帯」のスコアが高い Web ページをスコアの伝播元とした場合の 2 通りについて，biased PageRank によるスコアの伝播を行った．今回は，「帯」のスコアが低い Web ページを選ぶ場合には，スコアが 1〜6 の Web ページのすべてをスコアの伝播元として選び，「帯」のスコアが高い Web ページを選ぶ場合には，スコアが 13 の Web ページのすべてをスコアの伝播元として実験を行った．この手法がうまくいけば，「帯」のスコアが低い Web ページをスコアの伝播元として選んだ場合には，アルゴリズムの適用後，「帯」のスコアが低い Web ページに対して提案手法によるスコアが高くなる．逆に，「帯」のスコアが高い Web ページをスコアの伝播元として選んだ場合には，アルゴリズムの適用後，「帯」のスコアが高い Web ページに対して提案手法によるスコアが高くなる．

この手法の評価を，次のようにして行った．

「帯」のスコアが低い Web ページをスコアの伝播元として選んだ場合の評価は，上位 k 件を選んだ時の適合率を以下のように定義し，Web ページを提案手法により算出されたスコアが降順になるように並べた場合についての適合率を調べた．この結果を，PageRank アルゴリズムによって算出されたスコアが降順になるように並べた場合および，昇順になるように並べた場合についての適合率と比較を行った．

$$\frac{\text{「帯」のスコアが閾値以下の Web ページの数}}{k} \tag{4}$$

閾値を 8 として，表 5 に示した 10 個の各キーワードに対する結果の平均をとったものを，図 2 に示す．

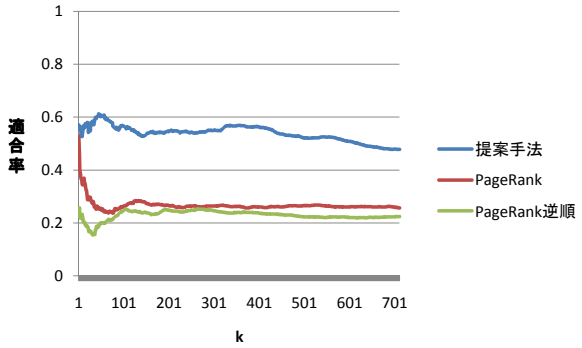

図 2　「帯」のスコアが 1〜6 のページを伝播元，閾値 8 とした場合の結果

「帯」のスコアが高い Web ページをスコアの伝播元として選んだ場合の評価も同様に，上位 k 件を選んだ時の適合率を以下のように定義し，Web ページを提案手法により算出されたスコアが降順になるように並べた場合についての適合率を調べ，この結果と，PageRank アルゴリズムによって算出されたスコ

アが降順になるように並べた場合および，昇順になるように並べた場合についての適合率との比較を行った．

$$\frac{\text{「帯」のスコアが閾値以上の Web ページの数}}{k} \tag{5}$$

閾値を 10 として，表 5 に示した 10 個の各キーワードに対する結果の平均をとったものを，図 3 に示す．

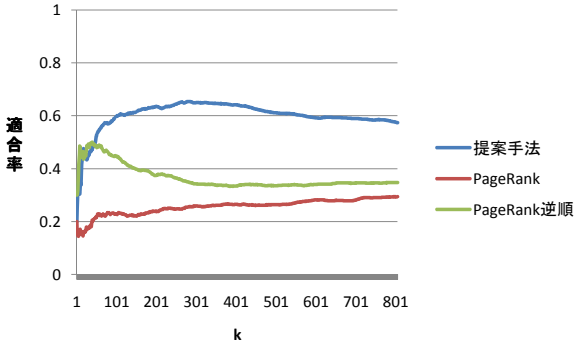

図 3　「帯」のスコアが 13 のページを伝播元とし，閾値 10 とした場合の結果

図 3，2 を見ると，図 3 の最初の数 10 件以外については，提案手法による適合率が，PageRank，PageRank 逆順による適合率を上回っていることがわかる．「帯」によるスコアが Web ページの理解容易性とよく合致しているという仮定のもとでは，提案手法によって，Web ページの理解容易性をある程度正しく評価できているといえる．

6. 結　　論

本論文では，Web ページの理解容易性とリンク構造との関係について調査する実験を行うとともに，TrustRank アルゴリズムの応用により，リンク解析によって Web ページの理解容易性を評価する手法を提案した．

Web ページの理解容易性の評価手法は，Web ページの内容の分かりやすさを考慮した Web 検索への適用を目指している．また，リンク解析というアプローチをとることは，Web ページの内容解析を行う困難を避ける目的がある．

Web ページの理解容易性とリンク構造との関係を調べる 2 つの実験では，Web ページの理解容易性とリンクとの間に一定の関係が存在していることが窺えた．

また，評価実験の結果から，提案手法を適用することによって，Web ページの理解容易性がある程度正しく評価できそうだといえる．

提案手法では，スコアの伝播元を，「帯」のスコアによって決定した．TrustRank においては，Web ページの集合からいくつかのシードページを選択し，そのシードページについて人手でチェックしてスコアの伝播元を決めるが，これと同様に，シードページを選択し，人手でその理解容易性を判定することによって，スコアの伝播元を決定してから，biased PageRank にしたがってスコアを伝播した場合にはどうなるかについて調べるべきであると考えており，今後さらに研究を進める必要がある．

また，提案手法では，「理解容易なページから理解容易なページへのリンクは多いが，理解容易なページから理解困難なページへのリンクは少ない」または「理解困難なページから理解困難なページへのリンクは多いが，理解困難なページから理解容易なページへのリンクは少ない」という前提が成立している必要がある．任意の 2 つの Web ページについて，それらの間にリンクが存在するというだけでなく，さらに条件を加えれば，2 つの Web ページの理解容易性が同程度である確率は向上するかもしれない．例えば，リンクが存在する 2 つの Web ページの内容の類似度を考慮したり，アンカーテキストやその周辺テキストを分析し，2 つの Web ページの理解容易性が同程度であるかどうかを示唆するキーワードが含まれていないか調べたりすることが考えられる．このようなことを考慮して今回の手法を改良すれば，算出される Web ページの理解容易性の精度が向上する可能性があり，今後の課題として検討していきたいと考えている．

謝辞　本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」（研究代表者：田中克己），文部科学省科学研究費補助金若手研究 (B)「Towards time-focused Web Search and Mining」(研究代表者：Adam Jatowt，番号 22700096）によるものです．ここに記して謝意を表します．

文　　献

[1] R. Flesch, “ A new readability yardstick“ , Journal of Applied Psychology, 32(3), pp. 221-233, 1948.

[2] S. Sato, S. Matsuyoshi and Y. Kondoh, “ Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus ”, Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008), ELRA, 2008.

[3] X. Yan, D. Song and X. Li, “ Concept-based Document Readability in Domain Specific Information Retrieval ”, Proceedings of the 15th Conference on Information and Knowledge Management (CIKM 2006), ACM, pp. 540-549, 2006.

[4] 中谷 誠, アダム ヤトフト, 田中克己, “ 理解容易性を考慮した用語説明のランキング手法 ”, Web とデータベースに関するフォーラム (WebDB Forum) 2009.

[5] M. Nakatani, A. Jatowt, H. Ohshima and K. Tanaka,“ Quality Evaluation of Search Results by Typicality and Speciality of Terms Extracted from Wikipedia”, Proceedings of the 14th International Conference on Database Systems for Advanced Applications (DASFAA 2009), pp. 570-584, 2009.

[6] M. Nakatani, A. Jatowt and K. Tanaka, ”Easiest-First Search: Towards Comprehension-based Web Search”, In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009), ACM, pp. 2057-2060, 2009.

[7] J. Zhao and M.-Y. Kan, “ Domain-Specific Iterative Readability Computation ”, Proceedings of the 10th annual joint conference on Digital libraries (JCDL 2010), ACM, pp. 205-214, 2010.

[8] Z. Gyöngyi, H. Garcia-Molina and J. Pedersen, “ Combating Web Spam with TrustRank ”, Proceedings of the 30th international conference on Very large data bases, VLDB Endowment, pp. 576-587, 2004.