理解容易度に基づくWebページの検索とランキング

中谷　誠† アダムヤトフト† 大島　裕明† 田中　克己†

† 京都大学情報学研究科
〒 606-8501 京都市左京区吉田本町
E-mail: †{nakatani,adam,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　Web 検索エンジンはユーザの入力したクエリに適合した Web ページの検索を行うが，検索結果に含まれる Web ページの理解容易性はほとんど考慮されていない．そのため，専門用語を含んだクエリが入力された場合，検索結果には一般的なユーザにとって理解が困難なページが含まれてしまうという問題がある．本稿では，文書の読みやすさを測る指標と検索クエリに関連する分野の知識による指標を組み合わせることによって Web ページの理解容易性を評価する手法を提案する．提案手法では Wikipedia のリンク構造とカテゴリ構造を分析することによって効率よく専門用語を抽出し，検索クエリに関連する分野の知識として用いる．

キーワード　Web 検索，リーダビリティ，Wikipedia マイニング，用語抽出

1. はじめに

近年インターネット上から情報を取得するために Web 検索エンジンが頻繁に利用されるようになってきた．Nakamura らが 2007 年に行った 1000 人規模のオンラインアンケート調査 [1] によると，ユーザが検索を行う動機として，

- 検索キーワードについて知らないため (46%)
- 検索キーワードについてより深く知りたいため (36.8%)

の 2 点が大きな割合を占めていることが分かった．詳しくない事柄について調べる場合，ユーザは検索エンジンに入力したキーワードについての理解容易な情報を含む Web ページを求めている．しかし，既存の Web 検索エンジンはクエリとの適合性やリンクによる支持度によって検索結果のランキングを行っており，必ずしも Web ページの理解容易性を考慮しているとは言えない．医療や経済などに関する専門用語を含むようなクエリが検索エンジンに与えられた場合，それらの分野に関して専門的な知識を持たない一般的なユーザにとって理解することが困難なページが検索結果のリスト中に含まれることがある．このような場合，ユーザは検索結果のリストから自分にとって理解容易な Web ページを手動で見つけ出さなければならないが，これはユーザにとって大きな負担となる．例として，医療系の専門知識を持たない一般的なユーザが検索エンジンを利用して「パーキンソン病」について調べるケースを考える．表 1 に示す記述を含む 2 つの Web ページ(注1)(注2)が得られたとする．(i) には「ドーパミン」や「アセチルコリン」,「錐体外路系」などといった多くの専門用語が含まれており，この記述を理解するためにはある程度の専門知識が必要とされる．一方，(ii) はパーキンソン病の具体的な症状について専門用語を使わずに説明しているため，(i) に比べ (ii) の記述を含んだ Web ページの方が一般的なユーザにとって理解が容易であると言えるだろう．しかし，実際には (i) の記述を含む Web ページが上位にランクされており，ユーザは (ii) のページを発見する前に (i) のページを選択してしまう可能性が高いと思われる．このように，既存の検索エンジンから得られる検索結果のリスト中には様々な難易度の Web ページが混在しているため，理解容易な Web ページを手軽に探すことは困難であることが多い．本研究では，検索エンジンを利用して得られる Web ページ集合を理解容易度に基づいて再ランキングする手法を提案する．Web ページの理解容易度はユーザの事前知識に依存するものであるが，本稿では検索キーワードに関する専門的な知識を持たない一般的なユーザにのみ焦点を当て，“Easiest-first” を満たす Web 検索を目標とする．

表 1 「パーキンソン病」についての記述例

(i) パーキンソン病（－びょう，英 Parkinson’s disease）は，脳内の ドーパミン 不足と アセチルコリン の相対的増加とを病態とし，錐体外路系 徴候を示す疾患である．神経変性疾患 の一つである．日本では難病（特定疾患）に指定されている．本疾患と二次性にパーキンソン病と似た症状を来たすものを総称して パーキンソン症候群 と言い，本症はパーキンソン症候群を示す病気の一つである．…
(ii) パーキンソン病は，多くは 40 歳以後に発症し，手足のふるえ，筋の固さ，動作の遅さ，歩行の拙劣さ，転びやすさなどの症状がみられる病気です．最初から全 部の症状がそろっているわけではありませんが，発症して数年経つとこれらの症状の大部分がみられるようになります．初期の症状で一番多いのは手のふるえです．…

以下，第 2 節で本研究のアプローチについて述べ，第 3 節では関連研究について述べる．第 4 節では Wikipedia の構造を用いて検索クエリに関する専門用語を抽出する手法について述べ，第 5 節では検索結果を理解容易度に基づいて再ランキングする手法の提案を行う．第 6 節でプロトタイプシステムについての説明と実行例を示し，第 7 節でまとめと今後の課題について述べる．

（注1）：http://ja.wikipedia.org/wiki/パーキンソン病
（注2）：http://www.tmin.ac.jp/medical/01/parkinson1.html

2. アプローチ

Web ページの理解容易性を評価するアプローチは大きく二つある．一つはクエリ非依存なアプローチであり，もう一方はクエリ依存なアプローチである．クエリ非依存なアプローチとして文書の読みやすさを測るリーダビリティテストがある．リーダビリティテストは文書中の単語の音節数や平均文長といった文書の表層的な特徴のみを分析することによって文書の理解容易性の評価を行う．リーダビリティの指標は簡単に計算することができるという利点があるが，専門用語を含む検索クエリが与えられたときに得られる Web ページの理解容易性を評価する上でいくつかの問題が生じる．まず，そのような場合に取得された Web ページには検索クエリに関連する専門用語が多く含まれているが，古典的なリーダビリティテストではそれらの専門用語を考慮して理解容易性を推定することができない．また，一般的な語が特定の専門分野においては特異な意味を持つ場合もある．これらの理由により，我々はクエリ非依存なアプローチだけでは，理解容易度に基づく Web 検索を実現する上で不十分であると考えている．

クエリ依存なアプローチとして，本研究ではクエリに関連する専門用語を用いることによって Web ページの理解容易性を評価する手法を提案する．専門分野の知識を利用して Web ページの理解容易性を評価する研究はほとんどない．Xin ら [2] は，医療系のシソーラスである MeSH(注3) を用いて概念ベースで文書の理解容易性を評価する手法を提案している．Xin らの手法は特定の分野における情報検索に焦点を当てたものであったが，一般の Web 検索エンジンにおいては多様な分野についての検索クエリが入力される．シソーラス間で構造は標準化されておらず，また必ずしも適切なシソーラスが見つかるという保証がないため，クエリごとに異なるシソーラスを利用することは非現実的である．また，ほとんどのシソーラスは専門家によって手動で構築されており，多くの場合最新の話題を含んでいないという問題がある．

本研究では，検索クエリの分野に非依存なアプローチを目指す．分野に非依存なアプローチをとるためには，クエリに関連する専門知識を効率よく獲得する必要がある．この要求を満たすために，本研究ではオンライン百科事典である Wikipedia の構造的特徴を用いて専門用語を取得する手法を提案する．Wikipedia は幅広い分野を網羅しており，かつ最新の話題も含んでいるため，検索クエリに関連する専門用語を抽出する上で有用であると考えられる．専門用語を用いたクエリ依存なアプローチに加え，クエリ非依存なアプローチであるリーダビリティテストを併用することによって Web ページの理解容易性の評価を行う．

提案手法の概要を図 1 に示す．まず，既存の検索エンジンを用いてクエリキーワードを含む Web ページ集合を取得する．本稿での提案手法は検索クエリに関していくつかの仮定を置いている．まず，検索クエリは単一語であると仮定する．これは知らない語について調べるために検索を行うときには，最も自然な

(注3)：http://www.nlm.nih.gov/pubs/factsheets/mesh.html

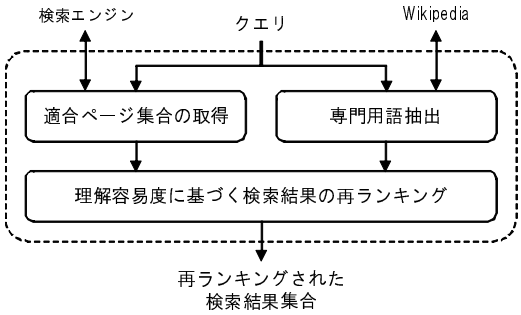

図 1　提案手法の概要

クエリの入力方法であると思われるためである．次に，検索クエリは Wikipedia 記事の見出し語であるものとする．Wikipedia は通常の百科事典などと比べて非常に多くの記事を含んでいるが，必ずしも検索クエリについての記事が存在するとは限らない．検索結果集合に含まれるページの多くは検索クエリに適合したものであるが，一部のページはクエリに関する情報を含んでいない可能性がある．そのため，理解容易性を評価する上でノイズとなるようなページを元の検索結果集合から除去する．提案手法においては，LexRank アルゴリズム [3] を利用することによって，重要性の低いページを除去する．ノイズとなるページを除去したのち，リーダビリティテストと Wikipedia から得られた専門用語を用いて Web ページの理解容易度を求め，それを元に検索結果を再ランキングする．

3. 関 連 研 究

3.1 リーダビリティ

リーダビリティとは文書の読みやすさのことであり，Web ページの理解容易性に大きく影響を与える要因の一つである．言語学や教育学の分野ではリーダビリティに関する多くの指標が提案され利用されている．Gunning-Fog Index や ARI [4] といった古典的なリーダビリティの指標の多くは，単語の音節数や文の長さのような文書の表層的な情報を用いて計算される．Dale-Chall Readability Index [5] などの一部のリーダビリティの指標は，語彙の難易度のようなやや意味的な情報も考慮する．Gray ら [6] の指摘によると，文書のリーダビリティは，

（1） **内容** ： 主張やその一貫性
（2） **文体** ： 文の長さや語彙の難易度
（3） **デザイン** ： レイアウトや文字の書体・色
（4） **構造** ： 章構成や見出し

の 4 つの要因に影響される．古典的なリーダビリティテストの多くは文書の文体を評価するものであるが，先に述べた通り専門的な内容の Web ページの理解容易性を評価するためには十分ではない．本稿では，リーダビリティの指標と検索クエリに関連する専門的な知識を併用することによって理解容易性に基づく Web 検索を目指す．Xin ら [2] は古典的なリーダビリティテストに加えて文書に含まれる概念の深さと概念間の結合性を考慮することによって専門的な内容を含んだ文書の理解容易性を評価する手法を提案した．我々のアプローチは Xin らのものに近いが，分野に非依存であるという点で異なる．また，直接

的に Web ページの理解容易性やリーダビリティを評価したものではないが，デザインや構造の観点から Web ページの品質の評価を行う試みもある [7] [8]．これらの研究では，機械学習を用いたアプローチがとられている．

3.2 Wikipedia マイニング

Wikipedia は誰でも編集することのできるオンライン百科事典であり，相互にリンクを張りあった非常に多くの記事を有している．Wikipedia の統計情報(注4)によると，2008 年 6 月時点で英語版 Wikipedia には約 240 万の記事が投稿されている．近年，Wikipedia はデータマイニングのためのコーパスとして注目されてきており，「Wikipedia マイニング」という新しい研究分野が生まれている．Nature 誌によると自然科学のトピックにおいて Wikipedia は Britannica 百科事典と同程度の精度を示しているという [9]．また，Milne らの実験 [10] は，農学分野のシソーラスである Agrovoc と比較して，Wikipedia には専門用語や上位・下位といった用語間の意味的関係が十分に含まれていることを示した．

Wikipedia の構造を利用して概念間の関連性を求めようといういくつかの試みがある．Strube ら [11] は Wikipedia 記事のカテゴリ構造に，Ito ら [12] は Wikipedia 上でのリンクの共起性に着目した手法を提案している．また，Wikipedia から抽出された知識を利用したいくつかのアプリケーションが提案されている．Koru [13] は Wikipedia から得られた関連語によって自動的にクエリ拡張を行うものである．Mihalcea らの Wikify! [14] は，Wikipedia から得られたキーワードを用いて通常のウェブページに含まれるキーワードに Wikipedia 記事へのリンクを付与するシステムである．これらのシステムの目的がユーザの検索や学習をサポートすることであるのに対し，本研究は Wikipedia から抽出された専門用語を元に Web ページの理解容易性を評価することを目的としている．

4. Wikipedia からの専門用語抽出

4.1 抽出手法

Wikipedia のリンク構造とカテゴリ構造を用いて検索クエリに関連する専門用語を効率良く取得する手法について述べる．ここで検索クエリに関連する専門用語とは，検索クエリの属する分野においてのみ頻出するような語である．例えば，クエリとして「サポートベクターマシン」が与えられた場合，機械学習や分類アルゴリズムに関する用語を説明するためだけに頻繁に用いられる語が専門用語であると言える．Wikipedia 記事は通常一つ以上のカテゴリに属している．また，各記事はその記事を理解する上で役に立つ他の記事へリンクを張っている．提案手法においては，カテゴリ構造は検索クエリの属する分野を検出するために用いられ，リンク構造はカテゴリ内外での用語の出現の偏り度合を求めるために用いられる．以下，専門用語抽出のアルゴリズムの詳細について述べる．

4.1.1 検索クエリの分野の検出

クエリキーワードを q とする．ただし，Wikipedia に q に関する記事があるものとする．クエリ q の Wikipedia 記事の属しているカテゴリ集合 $C_q = \{c_1, ..., c_m\}$ を取得する．各カテゴリ c_i は直観的にはクエリ語の上位語となっており，取得されたカテゴリ集合を検索クエリの属する分野と見なすことができる．例えば，「パーキンソン病」の記事は「特定疾患」や「神経変性疾患」のカテゴリに属している．一部の記事は「編集保護中の記事」のように記事の状態を表すカテゴリに属しているが，このようなカテゴリは検索クエリの属する分野を検出する上では必要ないので取り除く．

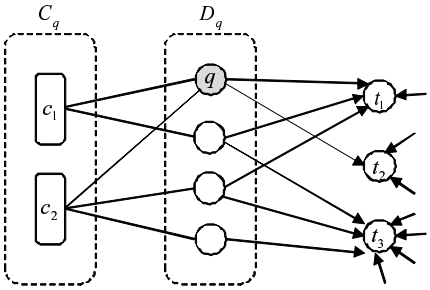

図 2　Wikipedia の構造を用いた専門用語抽出

4.1.2 リンク構造分析による専門用語判定

得られたカテゴリ集合 C_q の各要素 c_i について，それぞれ含まれる記事を全て取得して記事集合 D_q を生成する．語 t がクエリ q に関する専門用語であるとは，語 t の Wikipedia 記事へのリンク元の記事が記事集合 D_q 内に偏っていることということである．図 2 において，t_1 のリンク元は D_q 内に偏っているため t_1 は専門用語であるといえるが，t_2 や t_3 は D_q に含まれていない記事からも多くリンクされているため専門用語であるとはいえない．ここではリンクの偏り度合を求めるためにカルバック・ライブラー情報量 (KLD) を用いることで，「記事集合 D_q に含まれることが，語 t の記事へのリンクの出現に，どれだけ影響を与えるか」を求める．

$P(t)$：　ランダムに選ばれた Wikipedia 記事が語 t の記事へのリンクを持っている確率，

$P(\neg t)$：　ランダムに選ばれた Wikipedia 記事が語 t の記事へのリンクを持っていない確率，

$P(t|D_q)$：　記事集合 D_q からランダムに選ばれた Wikipedia 記事が語 t の記事へのリンクを持っている確率，

$P(\neg t|D_q)$：　記事集合 D_q からランダムに選ばれた Wikipedia 記事が語 t の記事へのリンクを持っていない確率

とすると，KLD は次の式によって表わされる．

$$KLD(t; D_q) = P(t) \log \frac{P(t|D_q)}{P(t)} + P(\neg t) \log \frac{P(\neg t|D_q)}{P(\neg t)} \qquad (1)$$

$KLD(t; D_q) \geqq \theta_{KLD}$ を満たす語 t をクエリ q に関する専門用語として取得する．

4.2 予備実験

Wikipedia の構造を用いた専門用語抽出手法の精度を評価するための予備実験を行った．実験のために「天文学」「医療・健康」「情報学」「経済・金融」の 4 つの分野についてそれぞれ 5 つの検索クエリを用いた．複数分野の検索クエリを用いるのは，Wikipedia を用いた我々の提案手法がどのような分野の検索ク

(注4)：http://en.wikipedia.org/wiki/Special:Statistics

表 2　予備実験結果

天文学		
クエリ	抽出数	適合率
ブラックホール	149	0.7584
地動説	19	0.7895
暗黒物質	74	0.8649
ニュートリノ	83	0.9157
ビッグバン	77	0.7922
平均	80.4	**0.8184**

医療・健康		
クエリ	抽出数	適合率
パーキンソン病	86	0.7326
動脈硬化症	42	0.7143
大腸癌	69	0.7826
レーシック	11	0.7273
メタボリックシンドローム	12	0.4167
平均	45	**0.7273**

情報学		
クエリ	抽出数	適合率
サポートベクターマシン	28	0.7500
SQL インジェクション	25	0.8400
形態素解析	31	0.7419
相互情報量	138	0.7174
ポリモーフィズム	60	0.8833
平均	56.4	**0.7695**

経済・金融		
クエリ	抽出数	適合率
サブプライムローン	22	0.5000
スタグフレーション	46	0.5870
モラル・ハザード	11	0.6364
ケインズ経済学	44	0.6364
市場経済	20	0.6000
平均	28.6	**0.5944**

エリにも対応できることを示すためである．専門用語であるかどうかの判定を行う対象語は，検索クエリについての Wikipedia 記事と同一カテゴリの記事集合中で 5 つ以上の記事にアンカーテキストとして出現する語を用いた．判定のためのパラメータは $\theta_{KLD} = 0.01$ とした．評価尺度は抽出された語のうち専門用語であると判断できる語の割合を表す適合率を用いる．

実験結果は表 2 に示す通りである．「天文学」「医療・健康」「情報学」の分野の検索クエリについては，70%から 80%程度の抽出精度を示すことが分かった．しかし，「経済学」の分野の検索クエリからの抽出の精度が他の分野の場合よりやや低く，分野間で抽出精度に差が生じる結果となった．これは分野によって Wikipedia 記事の充実度に差が存在していること，また他記事へのリンクの張り方に違いがある可能性を示唆している．表 3 に「パーキンソン病」「サポートベクターマシン」「ブラックホール」「サブプライムローン」を入力としたときに抽出された専門用語の例を示す．表に示されるように，多様な分野の用語に対して専門用語を取得できることが分かる．「パーキンソン病」に対する「病気」や「サブプライムローン」に対する「売買」や「住宅」のように，一般的な語彙も一部含まれてしまっている．

表 3　専門用語の抽出例

ブラックホール
シュヴァルツシルトの解，アインシュタイン方程式，ロイ・カー，シュヴァルツシルト半径，事象の地平面，一般相対性理論
パーキンソン病
レボドパ, パーキンソン症候群, 不随意運動, 固縮，振戦，アセチルコリン，病気，遺伝，血液脳関門，ウイルス，肺線維症，線条体
サポートベクターマシン
人工知能，ニューラルネットワーク，教師あり学習，最適化問題，パターン認識，線形分類器，知能，プログラム，線型性
サブプライムローン
不動産，証券化，賃貸借，抵当権，住宅ローン，売買，貯蓄貸付組合，有価証券，住宅，債権，建築基準法

5.　理解容易度に基づく検索結果の再ランキング

5.1　文書の理解容易度

5.1.1　リーダビリティ

文書の読みやすさは Web ページの理解容易性を大きな影響を与える要因である．3.1 では，英語を対象言語としたリーダビリティテストについて述べた．本研究では，日本語で記述された Web ページの評価を行うため，日本語を対象としたリーダビリティテストとして Sato ら [15] の開発した「帯」(注5) を用いる．帯では 13 段階の難易度を持つ教科書をコーパスとして利用し，与えられた文書中の文字の生起確率に基づいて各難易度に対する尤度を計算して，最大の尤度を持つ難易度を 1（小学 1 年生レベル）から 13（大学生レベル）までの整数値として出力する．文字の生起確率によって文書のリーダビリティを推定しているため，句読点がない等の体裁の整っていない文書を含むような Web ページのリーダビリティも計算することができる．本稿では帯の出力を理解容易性の尺度として用いる．

$$C_{readability}(p) = Obi(p)^{-1} \tag{2}$$

ただし，$Obi(p)$ は帯の出力値を表す．$C_{readability}$ 値の高いページほど読むために必要とされるリーディング能力が低いため，理解容易性が高いと考えられる．

5.1.2　専門用語の出現密度

Web ページ中に含まれる専門用語は，その理解容易度に影響を与えるクエリ依存な要因である．専門用語を多く含んでいる Web ページを理解することは，多くの前提知識を必要とするため一般的なユーザにとって困難であると考えられる．一方，単純にページ中に含まれる専門用語の異なり数を用いると，文書長の長いページは専門用語を多く含む可能性が高いため，全体として見ると理解容易であるとしてもその通りに判定されないという問題がある．そのため，ここではページの文書長を考慮した専門用語の出現密度を用いて理解容易性を評価する．評価式は次に示すとおりである．

$$C_{technical}(p,q) = \exp\left\{-\frac{n_t(p,q)}{\log|p|}\right\} \tag{3}$$

ここで，$n_t(p,q)$ とはページ p 中に含まれるクエリ q に関する

(注5)：http://kotoba.nuee.nagoya-u.ac.jp/sc/readability/obi.html

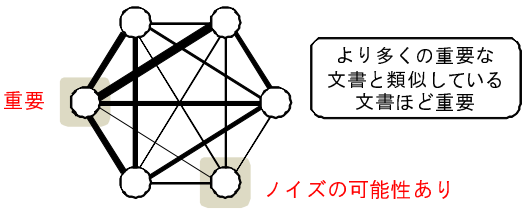

図 3　LexRank の考え方

専門用語の異なり数であり，$|p|$ はページ p の文書長を表す．$C_{technical}$ 値の高いページほど読むために必要とされる専門分野に関する事前知識が少ないため，理解容易性が高いと考えられる．

5.2　検索結果の再ランキング

検索エンジンを利用することで得られる検索語を含む Web ページ集合を理解容易性の尺度に基づいて再ランキングする．しかしながら，検索結果中には理解容易性の評価を行う上でノイズとなるページが存在する場合がある．ここでのノイズページとはサイトのトップページや検索クエリに不適合なページのことであり，これらのページには検索語に関する情報が少ないために専門用語がほとんど含まれておらず，誤って上位に再ランキングされてしまう恐れがある．本稿では，検索結果の上位には多くの適合ページが含まれており，かつノイズページと適合ページのテキストの類似度は低いという仮定を置き，文書要約のアルゴリズムとして提案された LexRank [3] を用いることでそれらのノイズページを除去する．ノイズページが除去した上で，各検索結果の理解容易性を評価することで検索結果の再ランキングを行う．

5.2.1　LexRank を用いたノイズページの除去

LexRank は文書集合を無向グラフで表し，グラフ上での中心性を求めることによって各文書の重要性を算出する，PageRank に類似したアルゴリズムである．LexRank の基本的な考えは図 3 に示す通りであり，多くの重要な文書と類似している文書に高いスコアが与えられる．本稿では LexRank の概要を簡潔に述べるに留め，アルゴリズムの詳細は [3] に譲る．

まず，検索エンジンを用いて取得された Web ページをダウンロードし，HTML のタグ構造を利用したヒューリスティックな方法により本文領域を抽出する．取得された文書を形態素解析したのち，tfidf 法による単語の重み付け手法 [16] を用いて特徴ベクトル化する．次に，任意の二つの文書間のコサイン類似度を計算し，各要素 $s_{i,j}$ が i 番目の文書と j 番目の文書の類似度を表す対称な行列 S を求める．各文書はグラフ上で節点として表され，二節点間にはそれらが表す二つの文書間のコサイン類似度によって重み付けされた枝が張られる．

PageRank と同様に，LexRank(LR) は次のように再帰的に定義される．

$$LR = dS^* \times LR + (1-d)p, \quad where \quad p = [\frac{1}{n}]_{n\times 1} \qquad (4)$$

ただし，n は文書数を，S^* は類似度行列 S の各列を正規化した行列を表す．また，d は dumping factor を表し，本稿では $d = 0.85$ とする．LexRank の値が低いページは他の適合ページと類似していないということを意味しており，不適合なページである可能性が高いと考え元の検索結果集合から除去する．

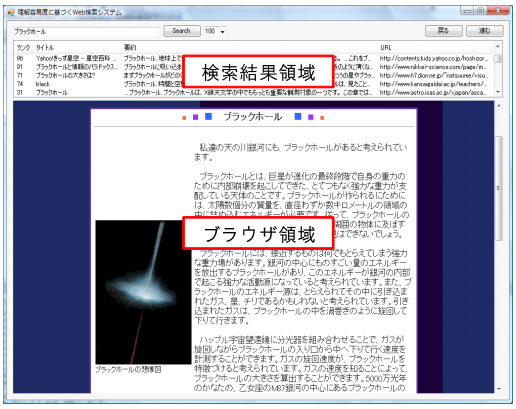

図 4　プロトタイプシステム

5.2.2　理解容易性の評価とランキング

リーダビリティテストに基づくクエリ非依存な尺度と専門用語の出現密度に基づくクエリ依存な尺度を併用することで，Web ページの理解容易性の評価を行う．具体的な評価式は次に示す通りである．

$$C(p,q) = (1-\alpha)\cdot C_{readability}(p) + \alpha \cdot C_{technical}(p,q) \qquad (5)$$

ただし、α は $0 \leqq \alpha \leqq 1$ を満たす実数である．C 値の高いページほど理解容易性が高いと考えて，検索結果集合を C 値の降順に再ランキングする．

6.　アプリケーション

6.1　実　　装

本稿での提案手法に基づいて Web 検索結果を提案手法により得られた理解容易度の順に再ランキングを行うシステムのプロトタイプを実装した．システムの動作例を図 4 に示す．システムは Web 検索の API を利用してユーザの入力したクエリを含む Web ページ集合を取得し，Wikipedia から抽出された検索クエリに関する専門用語とリーダビリティテストを用いて各 Web ページのスコアリングを行い，結果を理解容易度順でリストに表示する．ユーザが検索結果のリストから項目を選択することで，ブラウザコンポーネントに Web ページが表示される．Web 検索サービスは Yahoo!JAPAN の Web 検索 API(注6) を用いている．また，Wikipedia データベースは 2008 年 7 月 24 日の日本語版 Wikipedia のダンプデータをダウンロードしたものを利用している(注7)．

6.2　検索結果の再ランキング例

検索クエリとして「ブラックホール」「パーキンソン病」を検索エンジンに与えて取得される上位 100 件の検索結果を提案手法によって再ランキングした結果例を表 4 に示す．スコアリングの式におけるパラメータは $\alpha = 0.5$ とした．

「ブラックホール」の例では，1 位と 4 位に子供向けのブラッ

(注6)：http://developer.yahoo.co.jp/webapi/search/websearch/v1/websearch.html
(注7)：http://download.wikimedia.org/jawiki/

表 4　提案手法による Web 検索結果の再ランキング例（括弧内の数字は元の順位を表す）

クエリ：“ブラックホール”		
1 (23)	Cat Chat:Dr. ユニバース:ブラックホールって何ですか?	http://www.tbs.co.jp/catchat/friendpark/universe/que_blackhole.html
2 (46)	ブラックホールについて - 教えて!goo	http://oshiete1.goo.ne.jp/qa390129.html
3 (42)	ブラックホール	http://park1.wakwak.com/ỹumemaru/blackhole.html
4 (98)	Yahoo!きっず星空 - 星空百科 - 星の事典	http://contents.kids.yahoo.co.jp/hoshizora/encyclopedia/dic_blackhole.html
5 (72)	ブラックホールの大きさは?	http://www.h7.dion.ne.jp/ñatsuume/visual/BH1.html
41 (2)	SPACE INFORMATION CENTER : ブラックホール	http://spaceinfo.jaxa.jp/ja/black_holes.html
44 (1)	ブラックホール - Wikipedia	http://ja.wikipedia.org/wiki/ブラックホール
× (3)	BLACK HOLE	http://www.h3.dion.ne.jp/b̃lack.h/
クエリ：“パーキンソン病”		
1 (33)	Neuroinfo Japan:パーキンソン病	http://square.umin.ac.jp/neuroinf/patient/502.html
2 (84)	パーキンソン病 DBS に関する情報サイト — NouProblem.jp	http://www.nouproblem.jp/DBS/index.html
3 (93)	asahi.com : 健康 : 健康相談	http://www.asahi.com/health/soudan/jh030430.html
4 (3)	パーキンソン病	http://www.niigata-nh.go.jp/nanbyo/pd/pdindex.htm
5 (73)	パーキンソン病 をリンクで学ぶ [ペイスケ.com]	http://www.peisuke.com/parkinson/top.htm
50 (2)	難病情報センター—パーキンソン病関連疾患 (3) ...	http://www.nanbyou.or.jp/sikkan/089.htm
× (1)	Parkinson’s Disease	http://www.parkinson.gr.jp/

クホールについての解説ページが現れた．これらのページは子供向けの Web ページは平易な文章で記述されており，また専門用語もほとんど使われていないためであると考えられる．2 位のページは QA サイト内のページであり，回答者は分かりやすくブラックホールとは何かという質問について答えているが，内容の信憑性にやや問題があった．5 位にランクされたページは，ブラックホールの大きさについて例を交えつつ分かりやすく解説しているページであった．一方，元の順位が 1 位であった Wikipedia のブラックホールに関する記事と 2 位であった宇宙情報センターのページは，それぞれ多くの専門用語を用いてブラックホールについて解説しているため大きく順位を落とす結果となった．しかし，これらのページにおいては含まれている専門用語の解説ページへのリンクが張られており，それらを参照することにより内容の理解が助けられるが，ページ内の情報のみを利用している我々の提案手法ではページ外の情報による要因を考慮することができないためこのような結果となった．元順位が 3 位であったページは，天体のブラックホールについてのページではなかったため，不適合ページの除去の段階で取り除かれた．

「パーキンソン病」の例では，1 位と 2 位にパーキンソン病の治療に関する内容を含んだページが現れた．これらのページでは治療法に関するいくつかの専門用語が使用されていたが，本稿での提案手法でそれらの専門用語が抽出できていなかったために高く順位づけされてしまった．これは，理解容易性を評価するための専門用語抽出において，その再現性が重視される必要があることを示している．5 位に現れたページはパーキンソン病の具体的な症状や病気の進行について記述されたものであった．また，元の順位が 1 位であったページはパーキンソン病に関する Web サイトのトップページであったため，ほとんど文書情報を含んでおらず，不適合ページとして除去された．

「ブラックホール」のように子供の興味の対象となりやすい語がクエリとして入力された場合には，検索結果に子供向けの Web ページが含まれることが多い．このような場合，提案手法で考慮している 2 つの観点のうち，リーダビリティの指標が有効に働くものと考えられる．一方,「パーキンソン病」のようなクエリ例では，検索結果として得られる各 Web ページのリーダビリティに大きな差が見られないため，含まれている専門用語の多寡によって順位づけされることになる．このように，我々の着目した 2 つの理解容易性の観点はそれぞれに有効であるケースが異なっていると考察される．

7. まとめと今後の課題

本稿では，理解容易度に基づいた Web ページの検索・ランキング手法についての提案を行った．理解容易性の尺度として，クエリ非依存な尺度であるリーダビリティテストとクエリ依存な尺度である専門用語の出現頻度を用いた．検索クエリに関する専門用語を効率よく取得するため，本稿では Wikipedia のリンク構造とカテゴリ構造を用いて専門用語を抽出する手法を提案した．提案手法によって Web 検索結果を理解容易度順に再ランキングするプロトタイプシステムの実装を行い，システムの実行例を示した．

今後の課題として，Wikipedia からの専門用語抽出手法についての追加実験ならびに理解容易度に基づく Web ページの再ランキング手法の評価実験を行う予定である．専門用語の抽出手法に関しては，特に再現性についての評価が重要であると考えている．また，検索クエリとして複数の語や Wikipedia の見出し語でない語が入力された場合に対応できるように，クエリの関連分野の検出を中心に手法の拡張を行う必要がある．ランキング手法の評価において，理解容易性はクエリへの適合性のように単純に正誤を判断できるようなものではないため，適合率や再現率による機械的な評価は困難である．そのため，ユーザ実験を行って，提案手法によるランキング手法がユーザにとって妥当なものであるかを評価することを考えている．

謝辞　本研究の一部は，京都大学グローバル COE 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，文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」，計画研究「情報爆発に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田

中克己、課題番号 1809041），および，文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己），ならびに，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」，および，文部科学省科学研究費補助金若手研究 (B)「情報検索とウェブアーカイブにおけるマイニング」（研究代表者：Adam Jatowt，課題番号：18700111）によるものです．ここに記して謝意を表します．

文　　献

[1] S. Nakamura, S. Konishi, A. Jatowt, H. Ohshima, H. Kondo, T. Tezuka, S. Oyama and K. Tanaka: “Trustworthiness analysis of web search results”, Proceedings of the 11th ECDL (2007).

[2] X. Yan, D. Song and X. Li: “Concept-based document readability in domain specific information retrieval”, CIKM ’06: Proceedings of the 15th ACM international conference on Information and knowledge management, New York, NY, USA, ACM, pp. 540–549 (2006).

[3] G. Erkan and D. R. Radev: “Lexrank: Graph-based lexical centrality as salience in text summarization”, Journal of Artificial Intelligence Research, **22**, pp. 457–479 (2004).

[4] E. A. Smith and R. J. Senter: “Automated readability index”, AMRL-TR-66-22 (1967).

[5] E. Dale and J. Chall: “Readability Revisited: The New Dale-Chall Readability Formula”, Brookline Books/Lumen Editions (1995).

[6] W. S. Gray and B. Leary.: “What makes a book readable”, Chicago University Press (1935).

[7] M. Y. Ivory and M. A. Hearst: “Statistical profiles of highly-rated web sites”, CHI ’02: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM, pp. 367–374 (2002).

[8] T. Mandl: “Implementation and evaluation of a quality-based search engine”, HYPERTEXT ’06: Proceedings of the seventeenth conference on Hypertext and hypermedia, New York, NY, USA, ACM, pp. 73–84 (2006).

[9] J. Giles: “Internet encyclopedia go head to head”, Nature, **438**, (2005).

[10] D. Milne, O. Medelyan and I. H. Witten: “Mining domain-specific thesauri from wikipedia: A case study”, International Conference on Web Intelligence (2006).

[11] M. Strube and S. P. Ponzetto: “Wikirelate! computing semantic relatedness using wikipedia”, Proceedings of National Conference for Artificial Intelligence (2006).

[12] M. Ito, K. Nakayama, T. Hara and S. Nishio: “Association thesaurus construction methods based on link co-occurrence analysis for wikipedia”, CIKM ’08: Proceeding of the 17th ACM conference on Information and knowledge management, New York, NY, USA, ACM, pp. 817–826 (2008).

[13] D. N. Milne, I. H. Witten and D. M. Nichols: “A knowledge-based search engine powered by wikipedia”, Proceedings of the sixteenth ACM conference on CIKM, New York, NY, USA, ACM (2007).

[14] R. Mihalcea and A. Csomai: “Wikify!: linking documents to encyclopedic knowledge”, Proceedings of the sixteenth ACM conference on CIKM, ACM (2007).

[15] S. M. Satoshi Sato and Y. Kondoh: “Automatic assessment of japanese text readability based on a textbook corpus”, Proceedings of the Sixth International Language Resources and Evaluation (LREC’08) (Ed. by E. L. R. A. (ELRA)), Marrakech, Morocco (2008).

[16] G. Salton and C. Buckley: “Term-weighting approaches in automatic text retrieval”, Inf. Process. Manage., **24**, 5, pp. 513–523 (1988).