情報処理学会論文誌　Vol. 52　No. 3　1055–1068 (Mar. 2011)

子供によるWeb検索のための検索結果リランク手法

岩田麻佑[†1]　荒瀬由紀[†2]
原隆浩[†1]　西尾章治郎[†1]

インターネット環境の普及により，子供がWeb検索を行うことが一般的となっている．しかし難解な文章を苦手とし，画像を好むというような子供の特徴を考慮したWeb検索エンジンは存在せず，子供がWeb検索を快適に利用できる環境が整っているとはいえない．そこで本稿では，子供のWeb検索を支援するため，検索エンジンの検索結果を子供向けにリランクする手法を提案する．提案手法では，Webページの文章量などの構成に関する指標，子供向け表現の数などの文章に関する指標を設定し，これらの指標をスコア化する．そして，各指標のスコアを組み合わせて算出した各ページの子供向け度合いに基づき，検索エンジンの検索結果をリランクする．提案手法の有効性を検証するため，33人の小学生に評価してもらったWebページを用いて評価実験を行った．その結果，文章量，子供向け表現の数といった指標により，子供向けページを上位にランクできることを確認した．

A Re-ranking Method of Search Results for Web Search by Children

Mayu Iwata,[†1] Yuki Arase,[†2] Takahiro Hara[†1]
and Shojiro Nishio[†1]

Due to the explosive growth of the Internet technology, children are now familiar with the Internet, i.e., searching information using a search engine and browsing Web pages. However, there is no Web search engine that considers children's characteristics, for example, children are unwilling to read difficult textual contents while prefer images and animations. Therefore, children have to browse a large number of Web pages that are not friendly to them to find the information of interest. In this paper, to support children to use a search engine, we propose a method to re-rank a general seach engine's ranking according to children-friendly score, which is determined based on the structure of a Web page and its textual contents. We conducted an experiment to evaluate the effectiveness of the proposed method. As a ground-truth dataset, we asked 33 elementary school students to judge whether a Web page is children-friendly or not. The result shows that our method can re-rank Web pages for children by considering the amounts of the texts and the number of children-oriented expressions.

1. はじめに

インターネットの爆発的な普及[17]により，子供がインターネットを利用し，Web検索を行うことが一般的となってきている．2009年の統計[6]によると，小学生のインターネット利用率は年々増加しており，小学校6年生では9割近くもの子供が日常的にインターネットを利用している．さらに，gooリサーチによる2009年の調査[6]によると，子供がインターネットでWeb検索する際の目的は“勉強・宿題”に関する検索が53.9%，“趣味・娯楽”に関する検索が53.3%であり，インターネットを利用する際には，Web検索を頻繁に使用していることが分かる．このように，現在の子供は幼児，小学生の頃から日常的にインターネットに触れ，Web検索を行っている．

しかし難解な文章を苦手とし，画像を好む[13)–15)]というような子供の特徴を考慮したWeb検索エンジンは存在しないため，現状では，子供がWeb検索を快適に利用できる環境が整っているとはいえない．子供がWeb検索を行う際は，子供向け検索エンジン，もしくは，一般向け検索エンジンを利用すると考えられる．これらの検索エンジンでは，ランキング形式で表示される検索結果から自分が必要とする情報を探し，選択する必要がある．しかし，一般向け検索エンジンはもちろん，子供向け検索エンジンでも，検索結果のランキングは子供向けであるとはいえない．たとえば，一般向け検索エンジンの検索結果では，Wikipedia（http://ja.wikipedia.org/wiki）の記事が上位にランクされることが多いが，文章が多く，難解な表現が多用されているため，子供にとって分かりやすく，興味を持てるようなページであるとはいえない．また，子供向け検索エンジンでは，あらかじめ登録されたおすすめサイトがある場合のみ，おすすめサイトが上位にランクされるが，ランキングは一般向け検索エンジンのランキングと基本的に同じであり，おすすめサイトであっても必ずしも子供向け

†1 大阪大学大学院情報科学研究科マルチメディア工学専攻
Department of Multimedia Engineering, Graduate School of Information Science and Technology, Osaka University
†2 マイクロソフトリサーチアジア
Microsoft Research Asia

であるとはいえない．そのため，現状の検索エンジンでは，子供が必要とする分かりやすく，興味を持てる情報が上位にランクされるとは限らない．小学生や中学生を対象とした子供の Web 検索行動を調査した既存研究において，子供には，検索結果から自分が必要とするページを選択するのを苦手とし，検索結果の上位 5 件を超えるページをあまり閲覧しないという傾向が明らかになっているため[1),3)]，検索結果のランキングでは，子供向けのページを上位にランクする必要がある．

そこで本研究では，学校の宿題などで頻繁に Web 検索を行うと考えられる小学生を対象に，小学生の Web 検索を支援するため，子供向けページを上位にランクすることを目的とし，一般的な検索エンジンの検索結果を子供向けにリランクする手法を提案する．提案手法では，既存研究で得られた知見に基づき，子供にとって興味を持つことができ，見た目が見やすく，勉強になり，内容が分かりやすいようなページを子供向けページと定義し，一般向けページと子供向けページが混在したページ集合から子供向けページを判定できるよう，ページの子供向け度合いをスコア化する．具体的には，ページ中の画像やアニメーションの量，文章の量，色の数などの構成に関する指標，ページ中の文章の平均文字数，漢字やアルファベットの量，子供向け表現の数などの文章に関する指標を設定する．そして，各指標のスコアを組み合わせて各ページの子供向けスコアを決定し，検索エンジンの検索結果を子供向けスコアの降順に並べ変えることで，子供向けのリランクを行う．これにより，子供向けのページが検索結果上位にランクされるため，子供はランキング下位まで探す手間なしに，子供向けのページを容易に得ることできる．

本稿の構成は以下のとおりである．まず 2 章で関連研究について述べ，3 章で子供向けリランク手法について述べる．4 章で評価実験について述べ，最後に 5 章でまとめと今後の課題について述べる．

2. 関 連 研 究

本章では，まず Web を利用する際の子供の特徴を調査した既存研究について述べ，その後，子供を対象とした Web アプリケーションと既存の子供向け検索エンジンについて述べる．

2.1 Web 閲覧・検索に関する子供の特徴についての研究

子供が Web 閲覧・検索を行う際の特徴を調査した研究は数多くある．Bilal ら[2)]は，大学院生と 7 年生（日本の中学 1 年生）の子供に，ある特定のコンテンツを Web 検索により探すタスクを行ってもらい，大人と子供の間のタスクの成功率や検索行動の違いを調査している．その結果，子供には，検索に失敗したときに次にどうするべきなのか分からない，タスク中でも他に興味のあるコンテンツに気が散ってしまうというような特徴があるため，Web 検索によって必要な情報を探すことが大人よりも苦手であることが示されている．また，Bilal[1)]は，7 年生から 9 年生の子供に，子供向け検索エンジンでタスクを行ってもらい，認識面，身体面，感情面での特徴についても調査している．この調査の中で，ほとんどの子供が検索結果の上位 5 件を超えてページを閲覧しなかったことが報告されている．Druin ら[3)]は，7 歳から 11 歳の子供に自宅で自由に Web 検索を行ってもらう調査により，子供が Web 検索をする際に，スペリング，タイピング，クエリ作成，検索結果の解釈が問題となることを明らかにしている．特に，検索結果の解釈については，多くの子供が検索結果第 1 位のページに依存し，上位 5 件を超えたページを閲覧することがほとんどなかったと示されている．

また，Web ページに対する子供のユーザビリティを調査した研究もある．菊地ら[15)]は，小学校のパソコンの授業時の子供の様子を観察することで，小学生の Web ブラウジングの特徴を分析し，小さな子供ほど難解な漢字を用いた文章量の多い Web ページを嫌い，キャラクタなどのイラスト中心の Web ページを好むという特徴を明らかにしている．Nielsen ら[13),14)]は，6 歳から 12 歳，13 歳から 17 歳の子供に子供向けに作成された Web ページ，大人向けに作成された Web ページの両方を閲覧してもらい，ユーザビリティを調査している．その結果，子供は文字が詰まったようなページよりも見た目に分かりやすいページを好む傾向があることが示されている．特に，6 歳から 12 歳の子供の特徴として，アニメーションなどのマルチメディア要素やカラフルであるなどのデザイン要素を重視する点，スクロールをほとんど行わず，画面上部に見えている部分のみで操作を行うことがほとんどである点が明らかにされている．しかし，単に見た目が魅力的なだけではなく，コンテンツが多すぎず，できる限りシンプルに子供が操作できるようなデザインが必要であるとも述べられている．また，富士通による報告[4)]では，子供向け Web ページを作成する際のユニバーサルデザインについて述べられている．具体的には，子供にとって，分かりやすく，読みやすい文章とするために，文末に語りかけの表現を使用することや，専門用語や難解な漢字，アルファベットを避けること，文章量を少なくすることが必要であり，さらに，イラストなどを用いて見た目に興味を引くように工夫することも重要であると述べられている．

さらに，Web とは異なるが，子供向けの文章として，湯浅[21)]は，子供向けの小学生新聞などの記事と一般向けの新聞の記事を比較し，子供向け文章の特徴を調査している．その結果，子供向け文章は一般向け文章よりも，漢字の割合が少ないこと，1 文の文字数が少ない

こと，語りかけ表現や話し言葉が多用されていることが明らかになっている．

これらの既存研究の結果より，子供が好むページには，見た目に興味を引くように，画像やアニメーションが用いられ，カラフルであること，子供が容易に操作できるように，サイズが大きすぎず，複雑な構成でないこと，内容が分かりやすいように，子供に親しみのある表現が用いられ，漢字や難解な表現ができるだけ使用されていないことが重要であると考えられる．

さらに，子供が Web 検索を行う際には，検索クエリの入力や検索結果の選択など多くの問題があることが分かる．特に，現在の Web 検索エンジンでは，検索結果はランキングに基づいて表示されるのに対して，小学生や中学生といった子供は検索結果の上位しか閲覧しないことが明らかになっている．そのため，子供にとって，分かりやすく，興味をひくようなページがランキングの下位に存在すると，子供はそのようなページを探し出すことが困難である．そこで，上記に述べた子供が好むページの特徴を満たす，見た目に興味をひきやすく，文章が分かりやすいようなページを検索結果の上位にランクすることで，子供の Web 検索をサポートできると考えられる．

2.2　子供を対象とした Web アプリケーション

美馬ら[11] は，子供のための Web 情報検索支援アプリケーションを提案している．このアプリケーションでは，小学校の教科書を分析してオントロジを構築し，オントロジをもとに検索クエリを子供の学習向けに拡張する．たとえば，“りんご” を検索クエリに指定した場合，産地である “青森” や，栄養素である “食物繊維” などの関連語により果物としての意味的制約を加える検索クエリを作成する．これにより，検索を通じた学習までを考慮した子供用の Web 検索環境が実現できる．Nakaoka ら[12] は，幼小者の生活様式オントロジを構築し，子供の生活環境に密着した Web 検索を可能とするシステムを提案している．生活様式オントロジには，幼小者の体験するイベントに関連する事柄を記述する．たとえば，“クリスマス” ならば “人気のクリスマスプレゼント” などの具体例を記述する．このオントロジを用いることで，幼小者の検索意図を推測し，Web 検索のキーワードの想起を支援する．これらのシステムは，子供にとって分かりやすく有益な Web 検索環境の実現を目的とした，Web 検索のクエリの拡張をベースとした方法であり，検索結果として返されるページが子供にとって分かりやすく，親しみやすいものであるかどうかは考慮していない．

一方，商用化されている子供向けの Web 検索エンジンとして，以下のようなものがある．Yahoo!きっず[20] は，あらかじめ手作業で安全と判断されたサイトのみ検索可能である．検索結果のランキングは，サイト検索の結果をまず表示し，その後ページ検索の結果を表示する．ページ検索結果のランキングは基本的に Yahoo! JAPAN[19] と同じであるが，Yahoo!きっずに登録されていないサイトのページはランキングに表示されない．キッズ goo[9] は，Web 上のどのようなページでも検索可能であるが，フィルタリング機能を持ち，有害と判断されたページは選択しても閲覧することができない．検索結果のランキングは，サイト検索の結果の後に goo[5] と同じページ検索の結果が表示される．

このような子供向け検索エンジンは，有害情報の削除を主な目的としているため，子供は安全に Web 検索を行うことができる．しかし，ページ検索結果のランキングが一般的な検索エンジンと同じであり，子供向けのページが上位であるとは限らない．子供は，検索結果上位 5 件程度しか閲覧しない傾向があるため[1),3)]，子供向けページを上位にランクすることが重要と考えられる．

3.　子供向けリランク手法

3.1　子供向けページの定義

本研究では子供向けページとして，2.1 節で述べた子供の特性に合致するものを上位にランクする．具体的には，Web ページの構成と内容について，網羅的に以下の特徴量を考慮する．

- Web ページの構成
  ページの大部分が文字で構成され，含まれている情報量が多く，サイズの大きいページは，子供にとって面白みに欠け，どのコンテンツに注目すべきか分からない．さらに，子供にとってスクロール操作は負担である[2)]．そこで，子供にとっては，子供の興味をひき，分かりやすい構成が必要と考えられるため，以下の点を満たすページを子供向けと定義する．
  - 画像やアニメーションが使用されている[14)]．
  - テキストは適度な量である[14)]．
  - リンクはどこをクリックすべきか分かるように把握しやすい量である[14)]．
  - スクロール操作が少量で済むページサイズである[2),14)]．
  - カラフルな色使いである[14)]．
- Web ページの文章
  子供向けのページとしては，子供に親しみや興味を持たせる文章であることが重要と考えられる．また，子供は学習のために Web ページを閲覧することが多いため[6)]，子供にとって勉強になる分かりやすい文章が必要と考えられる．そのため，以下の点を満た

すページを子供向けとする．

– 一般的な文と比べると，長さは短めで，1 文に含まれる情報が少ない[21]．
– 漢字，アルファベットや難解な表現が少ない[14),21)]．
– 語りかけの表現などの子供向け特有の表現が含まれている[14),21)]．
– 難解な表現が使用されず，理解が容易な文である[14),21)]．

3.2　指標の設計

3.1 節で述べた定義に基づき，一般向けページと子供向けページが混在するページ集合から子供向けページを判定できるよう，**表 1** に示す指標を設定した．これらの指標は，既存研究で得られた知見をもとに，試験的に設計しており，4 章の評価実験で，各指標の効果を調査する．設定した指標は，構成に関する指標 6 個，文章に関する指標 5 個の計 11 個であり，これらの指標を用いて Web ページのスコア化を行い，スコアが大きいほど子供向けとする．

以降では，それぞれの指標の詳細を述べる．指標の値域については，値が小さいほど子供向けである指標は −1〜0，値が大きいほど子供向けである指標は 0〜1 の範囲となるよう，正規化を行う．これは，各指標を組み合わせてページのスコアを決定する際に，値が大きいほど良い指標は加算，値が小さいほど良い指標は減算して組み合わせるためである．

3.2.1　構成に関する指標

子供向けのページとして，Web ページのサイズやコンテンツ量というページの構成は，子供にとって情報を把握しやすく，興味を持てる形式である必要がある．そこで，以下の 6 つの指標を定義する．

表 1　スコア化に用いる指標
Table 1　Features of Web pages.

タイプ	指標	説明
構成	*Size*	ページの面積
	ImageRate	ページ中の画像やアニメーションの割合
	TextRate	ページ中の文章の割合
	AnchorRate	リンクの数
	Component	コンポーネントの数
	Color	使用されている色の数
文章	*KanjiRate*	テキスト中の漢字，アルファベットの割合
	ChildrenExpression	テキスト中の子供向け表現の割合
	DifficultExpression	テキスト中の難解表現の割合
	Easy	テキストの難易度
	SentenceLength	テキストの平均文字数

- ページのサイズ（*Size*）
  子供はスクロールを苦手とする特徴があるため[4)]，スクロールが必要な大きいサイズのページは子供向けのページとはいえない．そこで，ページの面積に基づいた指標を *Size* とし，**図 1** (a) に示すような面積が小さければ子供向けページであると判断する．具体的には，ページ i の $Size_i$ は式 (1) に基づき計算する．値が小さいほど子供向けなので，値域は $-1 \leq Size \leq 0$ とする．

$$Size_i = \begin{cases} -1 & (size\ of\ page_i \geq maxsize) \\ -\dfrac{size\ of\ page_i}{maxsize} & (size\ of\ page_i < maxsize) \end{cases} \quad (1)$$

  ここで，$size\ of\ page_i$ はページ i の面積，$maxsize$ はページの面積の最大値であり，ランダムな 500 ページの面積を調査した結果に基づき，1,000 × 5,000 [pix] と設定した．
- 画像量（*ImageRate*）
  画像やアニメーションは，子供の興味をひき，理解のサポートにもなるため，子供向け

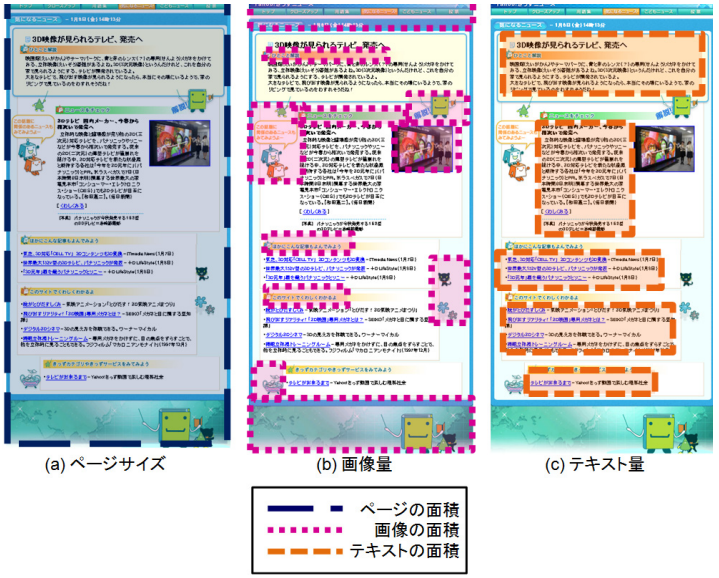

図 1　指標の例（*Size*，*ImageRate*，*TextRate*）
Fig. 1　Example of features (*Size*, *ImageRate*, *TextRate*).

ページに不可欠な要素であると考えられる．そこで，画像やアニメーションの量を判断する指標を *ImageRate* とし，図 1 (b) に示すようなページ全体の面積のうちの画像やアニメーションの占める面積の割合が大きいほど，子供向けのページであると判断する．具体的には，ページ i の $ImageRate_i$ は式 (2) に基づき計算する．値が大きいほど子供向けであるので，値域は $0 \leq ImageRate \leq 1$ とする．

$$ImageRate_i = \frac{\sum_{j=1}^{N} size\ of\ image_j^i}{size\ of\ page_i} \tag{2}$$

ここで，$size\ of\ image_j^i$ はページ i 内の j 番目の画像とアニメーションの面積，N はページ i に含まれる画像とアニメーションの数，$Size_i$ は前述したページ i の面積である．画像とアニメーションの面積は，HTML テキストより，$\langle img \rangle$，$\langle script \rangle$ などの HTML タグ内の情報を抽出して決定する．

- テキスト量（*TextRate*）
  文章量が多いと，子供にとっては面白みに欠け，ページを閲覧する意欲を持てないと考えられるため，文章量が多いページは子供向けであるとはいえない．そこで，文章量を判断する指標を *TextRate* とし，図 1 (c) に示すようなページ全体の面積のうちの文章の面積の占める割合が小さいほど子供向けであると判断する．具体的には，ページ i の $TextRate_i$ は，式 (3) に基づき計算する．値が小さいほど子供向けであるので，値域は $-1 \leq TextRate \leq 0$ とする．

$$TextRate_i = -\frac{\sum_{j=1}^{N} length\ of\ text_j^i \cdot fontsize}{size\ of\ page_i} \tag{3}$$

ここで，$length\ of\ text_j^i$ はページ i の j 番目の文の文字数，N はページ i の文の数，*fontsize* は文字のフォントサイズであり，今回は標準の 16 [point] をすべての文字のサイズと設定した．$Size_i$ は前述したページ i の面積である．

- リンク量（*AnchorRate*）
  リンクの量が多いと，子供はどのリンクをたどれば閲覧したいページにたどりつけるのか判断できなくなってしまう．そこで，リンクの量を判断する指標を *AnchorRate* とし，値が小さいほど子供向けのページであると判断する．ページ i の $AnchorRate_i$ は，式 (4) に基づき計算する．リンクテキストが少ないほど子供向けであるので，値域は $-1 \leq AnchorRate \leq 0$ として定義する．

$$AnchorRate_i = -\frac{number\ of\ anchor_i}{maxnumber\ of\ anchor} \tag{4}$$

ここで，$number\ of\ anchor_i$ はページ i に出現するリンクの数，$maxnumber\ of\ anchor$ はリンクの数の最大値であり，ランダムな 500 ページを調査した結果に基づき，300 個と設定した．

- コンポーネント（*Component*）
  コンポーネントとは，Web ページ内の関連する情報が集まったブロックである．コンポーネントの数が多いほど，ページの構成は複雑になる傾向があり，子供にとっては，注目すべきコンテンツを判断するのが難しくなる．そこで，コンポーネントの数に関する指標を *Component* とし，コンポーネントの数が少ないほど，子供向けのページであると判断する．具体的には，筆者らの所属する研究グループが提案した方式[10] でコンポーネントを抽出し，ページ i の $Component_i$ は式 (5) に基づき計算する．数が少ないほど子供向けであるので，値域は $-1 \leq Component \leq 0$ とする．

$$Component_i = \begin{cases} -1 & (number\ of\ comp_i \geq max\ of\ comp) \\ -\dfrac{number\ of\ comp_i}{max\ of\ comp} & (number\ of\ comp_i < max\ of\ comp) \end{cases} \tag{5}$$

ここで，$number\ of\ comp_i$ はページ i のコンポーネントの数，$max\ of\ comp$ はコンポーネント数の最大値であり，ランダムな 500 ページを調査した結果に基づき，20 個と設定した．

- 色数（*Color*）
  使用されている色の数が多いほど，見た目に興味をひく子供向けの表示と考えられる．そこで，ページに表示されている色の数を判断する指標を *Color* とし，色の数が多いほど子供向けページと判断する．ページ i の $Color_i$ は式 (6) に基づき計算する．値が大きいほど子供向けであるので，値域は $0 \leq Color \leq 1$ とする．

$$Color_i = \frac{number\ of\ color_i}{maxnumber\ of\ color} \tag{6}$$

ここで，$number\ of\ color_i$ はページ i で使用されている色の数で，ページで使用される色については，JPEG 形式の Web ページのスクリーンショット画像から異なる色の数をカウントし，$count\ of\ color_i$ を決定する．$maxnumber\ of\ color$ は色の数の最大値であり，JPEG 画像の最大色数である 1,670 万色と設定した．

3.2.2　文章に関する指標

子供向けのページとして，Web ページの文章は，子供にとって分かりやすく，興味を持

てる形式である必要がある．そこで，以下の 5 つの指標を定義する．

- 漢字・アルファベット量（*KanjiRate*）
  漢字，アルファベットの量が多いほど，その文章は子供にとって理解するのが困難だと考えられる．そこで，文章中の漢字，アルファベットの量に関する指標を *KanjiRate* とし，ページに含まれる文章中の漢字とアルファベットの文字数の占める割合が少ないほど，子供向けのページであると判断する．ページ i の $KanjiRate_i$ は式 (7) に基づき計算する．値が小さいほど子供向けであるので，値域は $-1 \leq KanjiRate \leq 0$ とする．

$$KanjiRate_i = -\frac{number\ of\ kanji_i + number\ of\ alphabet_i}{length\ of\ text_i} \tag{7}$$

  ここで，$number\ of\ kanji_i$ はページ i の文章に含まれる漢字の文字数の合計，$number\ of\ alphabet_i$ はページ i の文章に含まれるアルファベットの文字数の合計，$length\ of\ text_i$ はページ i に含まれる文章の文字数の合計である．

- 子供向け表現の量（*ChildrenExpression*）
  “〜しているよ” などの子供向け表現が多く含まれているページは，子供にとって親しみやすく，内容も容易な文章で書かれている可能性が高い．そこで，文章に含まれる子供向け表現に関する指標を *ChildrenExpression* とし，子供向け表現が多く含まれているほど，子供向けページと判断する．子供向け表現は，Yahoo!きっず[20]，キッズ goo[9]，NHK 週刊こどもニュース（http://www.nhk.or.jp/kdns/）をはじめとする子供向けページ 1,000 ページから形態素，文末表現などの頻出表現を抽出して利用する．ページ i の $ChildrenExpression_i$ は式 (8) に基づき計算する．値が大きいほど子供向けであるので，値域は $0 \leq ChildrenExpression \leq 1$ とする．

$$ChildrenExpression_i = \frac{\sum_{j=1}^{N} number\ of\ children\ expression_j}{number\ of\ term_i} \tag{8}$$

  ここで，$number\ of\ children\ expression_j$ はページ i の文章の j 番目に出現する子供向け表現，N はページ i の文章中の子供向け表現の出現数であり，あらかじめ作成した辞書の子供向け表現と合致する数とする．$number\ of\ term_i$ はページ i の文章に含まれる形態素，文末表現の総数である．

- 難解表現の量（*DifficultExpression*）
  難解な表現が多く含まれるページでは，子供は読む意欲を失い，内容を理解することが困難だと考えられる．そこで，文章に含まれる難解な表現に関する指標を *DifficultExpression* とし，難解な表現の量が少ないほど，子供向けページであると判断する．難解表現は，Yahoo!ニュース（http://headlines.yahoo.co.jp/hl）1,000 ページ，Wikipedia の記事 1,000 ページから頻出表現を抽出して利用する．子供にとって一般向けに作成されたニュースページ，また専門用語などを解説するようなページは難解であると考えられるため，Yahoo!ニュースと Wikipedia を用いた．ページ i の $DifficultExpression_i$ は式 (9) に基づき計算する．値が小さいほど子供向けであるので，値域は $-1 \leq DifficultExpression \leq 0$ とする．

$$DifficultExpression_i = -\frac{\sum_{j=1}^{N} number\ of\ difficult\ expression_j}{number\ of\ term_i} \tag{9}$$

  ここで，$number\ of\ difficult\ expression_j$ はページ i の文章の j 番目に出現する難解表現，N はページ i の文章中の難解表現の出現数であり，あらかじめ作成した辞書の難解表現と合致する数とする．$number\ of\ term_i$ はページ i の文章に含まれる形態素，文末表現の総数である．

- 難易度（*Easy*）
  単純に漢字・アルファベットの量，難解な表現の量という部分的な要素だけでなく，文章全体の難易度が低い方が，子供にとって理解が容易なため，文章全体の難易度に関する指標を *Easy* とし，難易度が低いほど子供向けのページと判断する．難易度は，教科書から作成したコーパスを用いて日本語の文章の難易度を推定するツールである帯[16]を利用する．帯によって推定される難易度を用い，ページ i の $Easy_i$ は式 (10) に基づき計算する．値が小さいほど子供向けであるので，値域は $-1 \leq Easy \leq 0$ とする．

$$Easy_i = -\frac{level_i}{13} \tag{10}$$

  ここで，$level_i$ は帯で推定したページ i のテキストの難易度であり，13 は帯で推定する難易度の最大値である．

- 文の平均文字数（*SentenceLength*）
  1 文が長いほど，1 文に含まれる情報が多くなり，子供がその文の内容を理解することが難しくなると考えられる．ここで，1 文の長さとは，漢字を平仮名に変換してカウントしたときの文字数と定義する．文字数の少ない文は，含まれている情報が簡潔で，子供にとって容易に理解できる文と考えられる．そこで，1 文の長さに関する指標を *SentenceLength* とし，1 文の文字数が少ないほど子供向けのページと判断する．具体的には，ページに含まれるすべての文の文字数を平均し，*SentenceLength* とし，ペー

ジ i の $SentenceLength_i$ は式 (11) に基づき計算する．値が小さいほど子供向けであるので，値域は $-1 \leq SentenceLength \leq 0$ とする．

$$SentenceLength_i = \begin{cases} -1 & (average\ len_i \geq maxlen) \\ -\dfrac{average\ len_i}{maxlen} & (average\ len_i < maxlen) \end{cases} \quad (11)$$

ここで，$average\ len_i$ はページ i の文章中のすべての文の平均文字数である．$maxlen$ は 1 文の文字数の最大値であり，ランダムな 500 ページの 1 文の平均文字数を調査した結果に基づき，100 文字と設定した．

3.3　リランク手順

3.2 節で述べた指標に基づき，Web ページのスコアを算出し，リランクを行う．まず，以下のような手順で Web ページの子供向けスコアを計算する．

(1) クロールを行い，Web ページを収集する．
(2) 子供が安全に Web 検索を行えるよう，フィルタリングで有害なページを削除する．
(3) 指標に基づき，各ページのスコアを算出する．具体的には，−1〜1 の各指標のスコアを加算して組み合わせ，各ページのスコアとする．

そして，ユーザがクエリを発行すると，以下の手順でリランクを行う．

(1) ユーザがクエリを入力する．
(2) クエリに対する検索結果を検索エンジンより取得する．
(3) 取得した検索結果の Web ページのスコアを取得する．
(4) スコアの降順にリランクして表示する．

4. 評 価 実 験

本章では，各指標の有効性，指標間の関係を調査し，その結果に基づいて指標を組み合わせた際の有効性を検証するために行った評価実験について述べる．

4.1　データセット

提案手法を評価するためには，実際に子供に Web ページを評価してもらった正解データが必要である．そこで，以下のようにデータセットを作成した．まず，Yahoo!きっず，キッズ goo における週間検索キーワードランキングの 12 月から 3 月のランキングの上位（10 位まで）より選んだ “地球温暖化”，“介助犬”，“ゲーム”，“うらない”，“冬至”，“百人一首” の 6 個のキーワードを実験に使用するクエリとし，各クエリごとに Yahoo! JAPAN, Yahoo!きっずの検索結果のランキングのそれぞれの上位から 25 の Web ページを取得した．そして，合計 300 の Web ページを**表 2** に示す小学校低学年（1，2 年生）から高学年（5，6 年生）の男女 33 人に評価してもらい，Yahoo! JAPAN 用，Yahoo!きっず用のデータセットをそれぞれ作成した．具体的には，それぞれの被験者に，Web ページを 30 秒程度閲覧してもらい，“読みたいと思うか？”，“見た目が見やすいか？”，“勉強になるか？”，“内容が分かりやすいか？” の 4 つの質問に Yes，No の 2 択で答えてもらい，1 ページあたり 4 人から 7 人による評価を得た．そして，Yes を 1 点，No を 0 点とし，各ページごとに点数を平均したものをそのページの子供向けスコアとし，子供向けスコアの降順に並べ変えたものを理想ランキングとし，データセットを作成した．この際，スコアが同じページは，Yahoo! JAPAN のランキングに基づいた順位とした．4 つの質問で評価を行ったのは，Web ページを子供向けであると判断するためには，ページの構成や見た目といった外観に基づく指標，文章など内容に基づく指標など多面的な評価が必要であり，それらを個々に検証するためである．具体的には，“読みたいと思うか？” の質問で子供にとって興味を持ちやすいページがどのようなものなのか，“見た目が見やすいか？” の質問で子供にとって見やすいと感じるページがどのようなものなのか，“勉強になるか？” の質問で子供にとって学習になるページがどのようなものなのか，“内容が分かりやすいか？” の質問で子供にとって理解が容易なページがどのようなものなのか調査するためである．

表 2　被験者の内訳
Table 2　Participants.

	小学校低学年	小学校中学年	小学校高学年
男	2 人	6 人	15 人
女	2 人	4 人	4 人

ここで，Yahoo! JAPAN，Yahoo!きっずのそれぞれでデータセットを作成したのは，Yahoo! JAPAN と Yahoo!きっずでは検索結果に含まれるページが大きく異なるためである．Yahoo!きっずのランキングでは，クエリに対応する登録サイトがあれば，そのサイトを検索結果上位に表示した後，Yahoo! JAPAN と同様のランキングを表示するが，Yahoo!きっずに登録されていないページは表示されない．そのため，Yahoo!きっずのランキングによるデータセットでは，Wikipedia などの明らかに子供向けでないページは含まれておらず，Yahoo! JAPAN のランキングと比べて，子供向けページが多く含まれる傾向がある．

4.2　評 価 指 標

評価指標には，正解データがクエリへの多段適合度を持つ場合の順位付き検索結果の性能

を測る *NDCG*（Normalized Discounted Comulative Gain）を用いた．クエリ q に対する *NDCG* は，適合度の高い順に並べた理想的な結果とのずれを表す指標で，式 (12) より求める．

$$NDCG_q = \frac{1}{IDCG_q}\left(rel_1 + \sum_{i=2}^{l}\frac{rel_i}{\log_2 i}\right) \tag{12}$$

rel_i は検索結果 i 番目のページのデータセットで定義されたスコア，l は検索数であり，子供が一般的に閲覧するといわれている上位 5 件で評価を行った．$IDCG_q$ はクエリ q に対する $NDCG_q$ の理想値，つまりデータセットで定義された理想ランキングの *NDCG* 値である．

4.3 実験結果

提案手法によるリランク結果と Yahoo! JAPAN，Yahoo!きっずによるランキング結果を比較した結果について述べる．ここで，Yahoo!きっずのデータセットには子供向けページを多く含むため，Yahoo! JAPAN よりも *NDCG* 値は高くなる傾向がある．具体的には，Yahoo!きっずの平均 *NDCG* 値が 0.68 であったのに対して，Yahoo! JAPAN の平均 *NDCG* 値は 0.60 となった．データセットでは，“読みたいか？”，“見た目が見やすいか？”，“勉強になるか？”，“内容が分かりやすいか？” の 4 つの質問で理想ランキングを定義しているため，それぞれの質問ごとに結果を述べる．

4.3.1 単独指標によるリランク

Yahoo! JAPAN のデータセットにおいて指標を単独に用いたリランクの *NDCG* の平均値を**図 2** に，Yahoo!きっずのデータセットにおいて指標を単独に用いたリランクの *NDCG* の平均値を**図 3** に示す．グラフでは，緑が Yahoo! JAPAN，Yahoo!きっずのランキング，青が構成に関する指標，赤が文章に関する指標によるリランクの結果を示す．

図 2 (a)，図 3 (a) に示すページの読みたさについては，Yahoo! JAPAN，Yahoo!きっずともに，*Size* のみ 2% *NDCG* 値が下がり，それ以外の指標では 1%から 14% *NDCG* 値が上がった．*Size* の *NDCG* 値が下がったのは，必ずしもサイズが小さいほど読みたいと感じるわけではなく，最適値が存在し，それより小さいページは興味を失う傾向があるためと考える．ただし，Wikipedia の記事のようにスクリーンの数倍以上のサイズのページは読みたくないと判断した子供が多かったので，そのようなページを排除する補助的な指標として使用することが有効と考える．精度の上がった指標の中で，特に *TextRate*，*Color*，*ChildrenExpression* は 7%から 18% *NDCG* 値が向上しており，子供にとっては，テキストが少なく，カラフルであり，さらに子供向け表現が含まれるページを読みたいと思うものと考えられる．

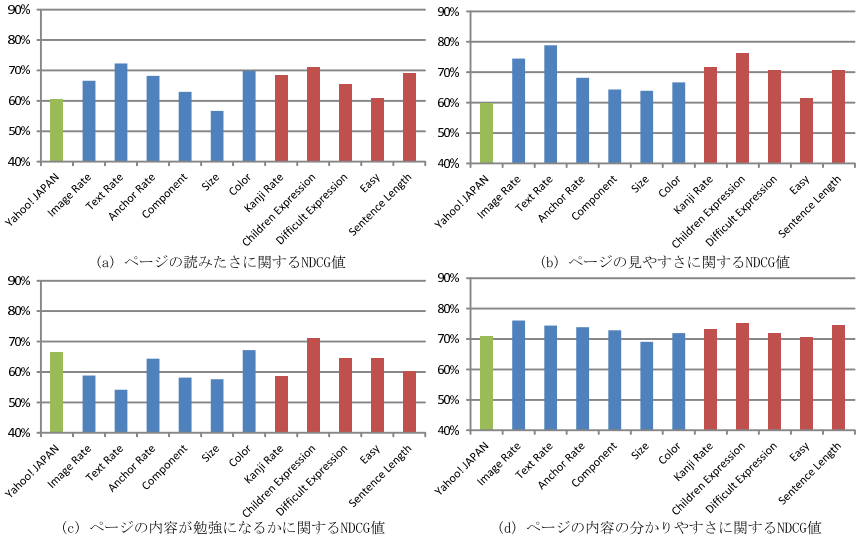

図 2　各指標の *NDCG* の平均値（Yahoo! JAPAN）
Fig. 2　*NDCG* of our methods using each feature and Yahoo! JAPAN.

図 2 (b)，図 3 (b) に示すページの見やすさについては，Yahoo! JAPAN ではすべての指標で 2%から 20%，Yahoo!きっずでもすべての指標で 4%から 10% *NDCG* 値が向上していた．特に，*ImageRate*，*TextRate* は，Yahoo! JAPAN，Yahoo!きっずどちらにおいても単独に指標を用いるだけで，5%から 18%も *NDCG* 値が向上しており，子供にとっては，画像量が多く，テキスト量が少ないページならば見た目に見やすいと思う傾向が強いものと考えられる．

図 2 (c)，図 3 (c) に示すページの内容が勉強になるかについては，*ChildrenExpression*，*Color* は 1%から 11% *NDCG* 値が向上し，それ以外の指標では *NDCG* 値は下がっていた．これは，30 秒程度ページを閲覧しただけでは勉強になるかどうかを判定するのが難しく，文章が多いページを単純に勉強になると判断した子供が多かったため，多くの指標が有効に働かなかったものと考えられる．*ChildrenExpression* は，あらかじめ定義した辞書に “なぜ”，“学ぼう” などの子供用の勉強ページに含まれるような表現が多数含まれていたため，精度良く判定できたものと考えられる．*Color* についても，子供用にカラフルに作成された

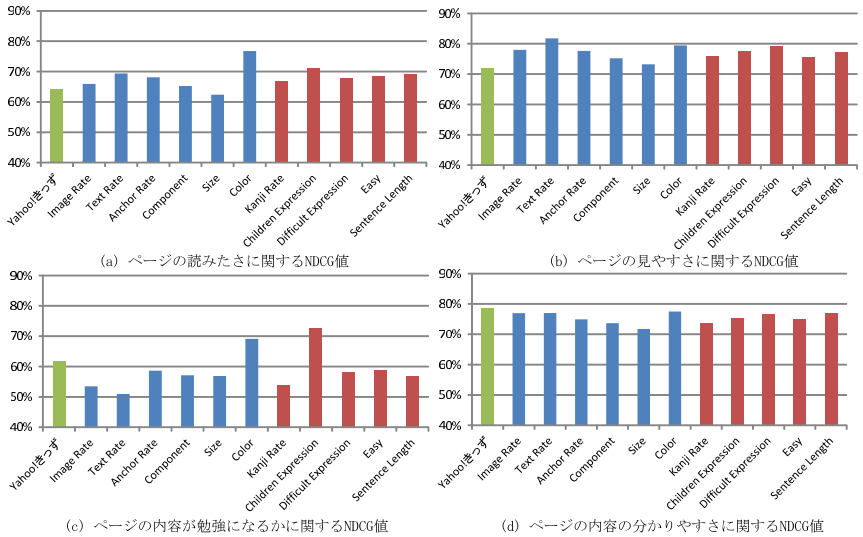

図 3　各指標の *NDCG* の平均値（Yahoo!きっず）
Fig. 3　*NDCG* of our methods using each feature and Yahoo! KIDS.

勉強用のページを判定できたものと考えられる．

図 2 (d)，図 3 (d) に示すページの内容の分かりやすさについては，Yahoo! JAPAN ではどの指標でも *NDCG* 値は 5%以下の向上にとどまっており，Yahoo!きっずではすべての指標で精度が下がった．これは，画像，テキストの両方をある程度用いて説明しているページを分かりやすいと見なす子供が多く，単純にテキストが少ないほど分かりやすいと判断することが不十分であったと考えられる．特に，Yahoo!きっずでは，多くのページが小学校高学年程度の子供にとって十分に分かりやすいため，むしろ，文章が多めで詳しく解説されているページを分かりやすいと見なす傾向があった．そのため，学年が上がるほど文章は多めが良いというように，各指標のスコアを学年ごとで適切な値とする必要がある．

これらの結果より，指標を単独に用いただけでも，多くの場合で Yahoo! JAPAN，また元々精度の高い Yahoo!きっずのランキングの精度を上回ることができることを確認した．また，どのような子供向けページを上位にランクするかで効果の大きい指標が異なることが分かる．“読みたさ”，“見やすさ”，“分かりやすさ” については画像量やテキスト量などのページの構成が子供向けであることが共通して必要であり，“勉強になるか” についてはそのような構成とはあまり関係がない．そして，“読みたさ”，“勉強になるか”，“分かりやすさ” については文章が子供向けであることが共通して重要である．しかし，“勉強になるか”，“分かりやすさ” については文章が容易なだけではなく，年齢に応じた適度なテキスト量，難易度である必要がある．つまり，“読みたさ”，“見やすさ” は組み合わせて使うことができるが，“分かりやすさ”，“勉強になるか” は単独で使うべきであると考える．

4.3.2　指標間の関係

4.3.1 項において指標単独でのリランクの結果を示したが，さらに精度を向上させるためには，指標を組み合わせることが有効と考えられる．そこで本項では，指標間の相関性を考慮し，組み合わせる指標を検討する．3.2.1 項で述べた 11 の指標のうち，以下に示す指標は独立性が低く，互いに相関性が強いと考えられる．

- *Size*（ページのサイズ），*Component*（コンポーネント数）：
  ページのサイズが大きいほど，含まれるコンポーネント数は増え，ページの構成が複雑になる傾向がある．そのため，ページのサイズとコンポーネント数は，ページの構成の複雑さを示すために用いる類似した指標であり，相関性が強いと考えられる．
- *ImageRate*（画像量），*Color*（色の数）：
  ページ内に多くの画像が用いられるほど，色の数が増え，見た目に興味をひきやすい構成となる傾向がある．そのため，画像量と色の数は，どちらも見た目に興味をひく構成に関係する要素であり，相関性が強いと考えられる．
- *KanjiRate*（漢字・アルファベット量），*ChildrenExpression*（子供向け表現の数），*DifficultExpression*（難解表現の数），*Easy*（難易度），*SentenceLength*（平均文字数）：
  文章に関する指標はどれも子供にとって分かりやすい文章であるかどうかを推定する指標であり，独立性が低いと考えられる．具体的には，子供にとって分かりやすく難易度の低い文章は，漢字やアルファベット，難解な表現が少ない一方，子供向け表現が多く，平均文字数が短い傾向があると考えられる．

そこで，これらの指標間の相関性を調査するために，指標ペア間の偏相関係数を算出した．偏相関係数は，複数の変数がある場合に，他の変数の影響を取り除いて，2 つの変数の相関関係を調査するために使用される．**表 3** に独立でないと考えられる指標間の偏相関係数を示す．偏相関係数は 3.2 節で述べた式により算出した各指標のスコアを用いて算出した．一般的に，相関係数は ±0.2〜0.4 で低い相関があるとされる[8]．

Size と *Component* については，偏相関係数は 0.654 と他と比較しても大きく，相関性が強いことが分かる．*ImageRate* と *Color* については，偏相関係数は 0.283 であり，*Size* と

表 3　指標間の偏相関係数
Table 3　Correlation coefficient between features.

指標のタイプ	指標のペア	相関係数
構成	*Component* と *Size*	0.654
	ImageRate と *Color*	0.283
文章	*ChildrenExpression* と *DifficultExpression*	0.312
	KanjiRate と *Easy*	0.244
	KajiRate と *ChildrenExpression*	0.200
	Easy と *SetenceLength*	0.170
	DifficultExpression と *Easy*	0.080
	ChildrenExpression と *SentenceLength*	0.062
	KanjiRate と *DifficultExpression*	−0.001
	ChildrenExpression と *Easy*	−0.020
	KanjiRate と *SentenceLength*	−0.086
	DifficultExpression と *SentenceLength*	−0.094

Component 間ほどではないが，低い相関が認められる．*KanjiRate*，*ChildrenExpression*，*DifficultExpression*，*Easy*，*SentenceLength* については，10 の指標ペアのうち，*KanjiRate* と *ChildrenExpression*，*KanjiRate* と *Easy*，*ChildrenExpression* と *DifficultExpression* の 3 ペアで 0.2 以上の偏相関係数となり，*Easy* と *SentenceLength* でも 0.17 と 0.2 に近い偏相関係数となった．つまり，漢字・アルファベットの量と子供向け表現の数，難易度や子供向け表現の数と難解表現の数，難易度と平均文字数は，文章の指標の中でも，互いに低い相関があり，独立性がそれほど高くないことが分かる．

上記以外のペアについては，相関係数は 0.1 以下となった．これは，*SentenceLength* が，短いキーワードで表されたリンク文字列であるメニューや広告などの影響を受けやすく，そのようなページでは平均文字数が短くなる可能性があることが 1 つの原因であると考えられる．つまり，*SentenceLength* については，他の文章に関する指標とは，独立性が比較的高いと考えられる．

また，*ChildrenExpression* や *DifficultExpression* が，他の指標と異なり，あらかじめ用意した辞書内の表現に依存していることも 1 つの原因と考えられる．*KanjiRate* と *ChildrenExpression* については，辞書内に定義した子供向け表現には，直接的に平仮名が多い傾向があるため，相関が見られたが，それ以外の指標との間には大きな相関は見られなかった．

このように，11 の指標の中には，独立性が十分でなく，実際に他の指標との相関の高い指標が含まれていることが分かる．そのため，指標を組み合わせる際には，互いに相関性の強いグループから代表的な指標を選択して用いることが有効と考えられる．

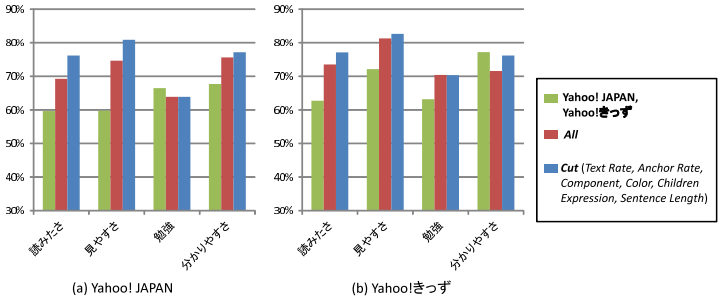

図 4　指標を組み合わせた *NDCG* の平均値
Fig. 4　*NDCG* of our methods combining features.

4.3.3　指標の組合せによるリランク

前項の議論に基づいて，指標を組み合わせた際の有効性を検証する．具体的には，教師付き機械学習を用いたランキング関数の学習の代表的な手法である Ranking SVM[7] を用いて，指標を組み合わせて利用した場合の性能を評価した．実装には svm_rank（http://www.cs.cornell.edu/People/tj/svm_light/svm_rank.html）を用い，カーネルには線形カーネルを用いた．Ranking SVM による学習を行うことで，各指標を適切に組み合わせることが可能となる．ここで，学習は，5 種類のクエリに対応する 250 ページで行い，残り 1 種類のクエリに対応する 50 ページで評価を行うことを 1 セットとし，6 種類のクエリすべてで評価できるように，6 セットの評価を行った．

指標を組み合わせた際の，Yahoo! JAPAN のデータセットにおけるリランクの *NDCG* 値の平均を**図 4** (a) に，Yahoo!きっずのデータセットにおけるリランクの *NDCG* 値の平均を図 4 (b) に示す．緑色が Yahoo! JAPAN，Yahoo!きっずのベースラインとなる *NDCG* 値，赤色の *All* が単純にすべての指標を用いた場合の *NDCG* 値，青色の *Cut* が，4.3.1 項，4.3.2 項の結果に基づいて独立性の観点から不要と考えられる指標を取り除いた場合の *NDCG* 値を示す．具体的に *Cut* は，4.3.2 項で述べた互いに相関性が強いと考えられるグループの中で，4.3.1 項で述べた指標単独での *NDCG* 値の低い指標を除いたものであり，以下のように決定した．

- *Component*，*Size*：
  NDCG 値の高い *Component* を使用．

- *ImageRate*，*Color*：
  NDCG 値の高い *Color* を使用．
- *KanjiRate*，*ChildrenExpression*，*DifficultExpression*，*Easy*，*SentenceLength*：
  NDCG 値の高い *ChildrenExpression*，および，ある程度 *NDCG* 値が高く他の指標と相関の小さい *SentenceLength* を使用．

この結果，*Cut* は，*Size*，*ImageRate*，*KanjiRate*，*Easy*，*DifficultExpression* の 5 つの指標を除いた，*TextRate*，*AnchorRate*，*Component*，*Color*，*ChildrenExpression*，*SentenceLength* の 6 つの指標の組合せとなる．

まず，すべての指標を組み合わせた *All* では，図 4 (a) に示す Yahoo! JAPAN では，勉強になるかについては Yahoo! JAPAN と同程度，それ以外の質問で 5%から 15%の *NDCG* 値の向上が見られた．この結果より，単純にすべての指標を用いても，指標を組み合わせる効果があることが分かる．また，図 4 (b) に示す，*NDCG* 値がもともと高い Yahoo!きっずでは，分かりやすさでは Yahoo!きっずの *NDCG* 値を下回ったが，それ以外の質問で 5%から 10%の *NDCG* 値の向上が見られた．分かりやすさで *All* の *NDCG* 値が下がってしまったのは，指標を単独で用いた場合でも Yahoo!きっずの *NDCG* 値を上回る指標がないため，それらを組み合わせることで性能が悪化したものと考えられる．

不要と考えられる指標をカットした *Cut* では，図 4 (a) に示す Yahoo! JAPAN では，*All* の *NDCG* 値と同程度から 7%の向上が見られた．また，図 4 (b) に示す Yahoo!きっずでも，同程度から 5%の *NDCG* 値の向上が見られた．この結果より，他の指標と相関しており，独立性の低い指標を除く効果が大きいことが分かる．ただし，勉強になるかの質問については，指標を除いたことによる *NDCG* 値の向上は見られなかった．これは，勉強になるかでは，他の質問と異なり，単独で *NDCG* 値の向上する指標が *Color* や *ChildrenExpression* のみであり，他の指標では *NDCG* 値が大きく下落するため，それらの指標の影響が大きく，指標を組み合わせても，*NDCG* 値が向上しなかったものと考えられる．そのため，勉強になるかの質問については，他の質問とは異なる指標の組合せが必要だと考えられる．

以上の結果より，適切な指標を組み合わせて Web ページを評価することで，単純にすべての指標を用いる場合と比べておおむね *NDCG* 値を向上させることができることを確認した．

4.4 考　　察

指標を Ranking SVM により学習して組み合わせた結果，多くの場合で，Yahoo! JAPAN や Yahoo!きっずよりも提案手法が子供向けページを上位にランキングできることを確認した．しかし，Yahoo! JAPAN の勉強になるかどうかでは，指標を組み合わせた *NDCG* 値が 0.65 であるのに対して，*ChildrenExpression* を単独に用いた *NDCG* 値が 0.72 であったように，単独指標の結果と比較して，指標を組み合わせることで性能が低下する場合があった．そのため，本節では，さらに精度を向上させるために必要と考えられるクエリごとの特徴，また子供の学年ごとの特徴について考察し，本手法のさらなる改善方法について検討する．

4.4.1 クエリごとの特徴

Yahoo!きっずの分かりやすさの *NDCG* 値が向上しなかった 1 つの原因はクエリによる差が大きいことがあげられる．そこで，本項ではクエリごとにどのような特徴があるのかを考察する．

図 5 に指標を単独で用いたリランク結果の中で，クエリによる差が顕著であった，Yahoo!きっずのデータセットにおける 4 つの指標（*ImageRate*，*TextRate*，*ChildrenExpression*，*KanjiRate*）のクエリごとの *NDCG* 値を示す．この結果，クエリごとに効果のある指標，効果のない指標の差が大きいことが分かる．たとえば，図 5 (c) の Yahoo!きっずの分かりや

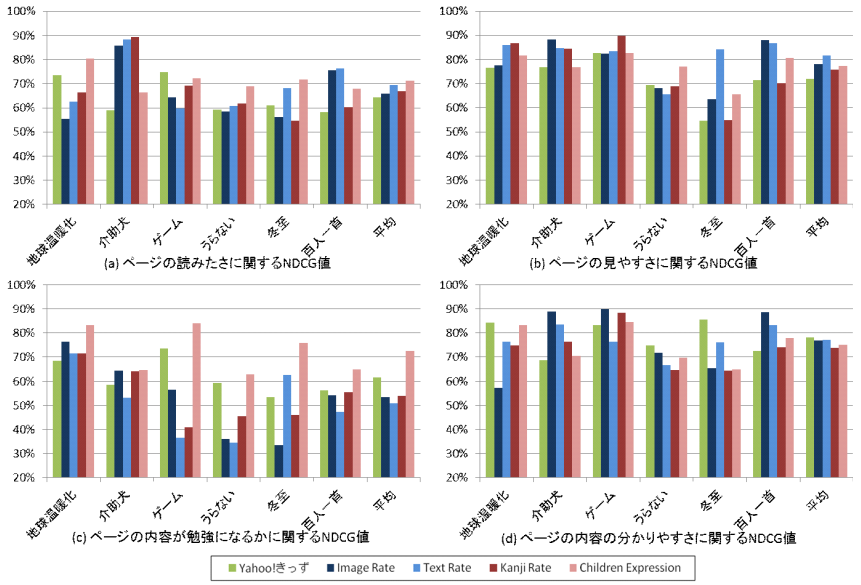

図 5　各指標のクエリ別の *NDCG* の平均値（Yahoo!きっず）
Fig. 5　*NDCG* of our methods on each query and Yahoo! KIDS.

すさで大きな差が見られ，“地球温暖化” や “冬至” はどの指標でも Yahoo!きっずの *NDCG* 値を下回っているが，“介助犬” や “百人一首” はどの指標も向上している．そのため，クエリによる傾向の違いを考慮して組み合わせる指標を変更する必要があると考えられる．具体的には，以下のような点を考慮する必要がある．

“ゲーム”，“うらない” という遊びや趣味に関するクエリは，図 5 (a) に示す読みたさや (b) に示す見やすさについて，他のクエリと比べて，*NDCG* 値が下がったり，向上率が小さかったりするという特徴が見られる．これは，個人の嗜好の影響が強いため，他のクエリと同じような組合せ方では *NDCG* 値が向上しないことが多いものと考えられる．たとえば，今回の実験でページの評価をしてもらった被験者には男子が多かったため，サンリオやディズニーのページは全体的にスコアが低くなっていた．そのため，遊びに関するクエリについては，個人の嗜好に合わせた指標の組合せ方が必要になると考えられる．

“介助犬” のように子供にとって親しみのないクエリでは，図 5 (a) に示す読みたさ，(b) に示す見やすさ，(d) に示す分かりやすさすべてにおいて，*ImageRate* や *TextRate* を考慮した *NDCG* 値が良く，他のクエリよりも *NDCG* 値の向上率が大きい傾向が見られる．このことより，子供にとって親しみのないクエリに対しては，ページに画像が多く含まれ，テキストが少ないことが子供の興味をひくために重要であるといえる．

また，“介助犬” のクエリの特徴の 1 つとして，他のクエリと異なり，*ChildrenExpression* よりも *KanjiRate* を考慮した方が *NDCG* 値が高い．これは，子供にとって親しみのないクエリでは，子供向けに作成されたページが少なく，そのような中では子供向け表現を含むページは限られ，単純に漢字があまり用いられていないページを判定する方が精度が良くなったためと考えられる．つまり，子供向けに作成されたサイトが少ないクエリに対しては，子供向け表現の数を考慮するだけでは不十分であると考えられる．

図 5 (d) に示す Yahoo!きっずの “地球温暖化”，“冬至” の分かりやすさについてはすべての指標で *NDCG* 値が下がり，(a) に示す読みたさでも *NDCG* 値が下がる傾向が見られる．この結果は，子供の勉強用に作成された解説ページ，子供にとって少し難易度の高い解説ページの両方が混在し，その両方を読みたい，分かりやすいと見なした子供が多かったことを示している．これは，難易度の高めの文章の多いページでも，コンテンツごとに色分けをしたりすることで，どこに何についての情報があるのかが分かりやすい構成になっているため，子供は内容を分かりやすく感じたものと考える．そのため，ある程度難易度の高いページでも整理された分かりやすい構成であれば，子供は内容が分かりやすいと感じることを考慮する必要がある．

また “地球温暖化”，“冬至” では，図 5 (d) に示す分かりやすさで，*ImageRate* を考慮した *NDCG* 値が 20%程度も下がっている．これは，クエリと関連のないキャラクタなどの画像のみを使用しているようなページについて，特に高学年の子供の分かりやすさの評価が低かったことが原因と考えられる．この結果より，単純に画像を多く用いているだけでは，分かりやすさは不十分であり，クエリと関連している画像を用いることが子供にとっての分かりやすさに重要であると考えられる．そのため，子供向け解説ページでも，画像が直接的にクエリに関連しているものであるのかを考慮する必要がある．

以上の議論から，個人の嗜好の影響の大きい遊び系のクエリ，検索エンジンのランキングに子供向け解説ページなどを多く含む授業で習う内容のクエリ，子供向けページの少ない子供に親しみのない内容のクエリなど，クエリ特性に応じて指標の選択や組合せを決定することが，精度の向上に重要と考えられる．

4.4.2　学年ごとの特徴

データセットを作成するために，本評価実験に参加した小学生は 60%が高学年を占めているが，学年により嗜好に差があるものと考えられる．そこで，この差を調べるために，**図 6**

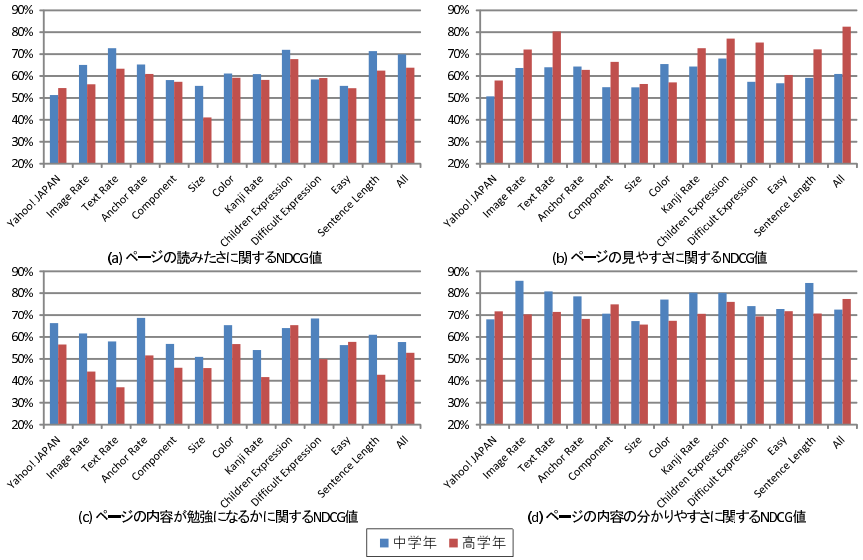

図 6　学年別の *NDCG* 値（Yahoo! JAPAN）
Fig. 6　*NDCG* of our methods on each grade for Yahoo! JAPAN.

に Yahoo! JAPAN をデータセットとした場合の単独指標の *NDCG* 値を，小学校中学年，高学年それぞれで示す．Yahoo! JAPAN，Yahoo!きっずをデータセットとした両方で，同様の特徴が見られたため，Yahoo!きっずをデータセットとした結果については省略する．また，低学年については，十分な数のページの評価を得られなかったため，結果を省略する．

図 6 (a) に示す読みたさについては，中学年，高学年の結果は類似している．しかし，中学年の方が *Size*，*SentenceLength* の *NDCG* 値が高学年よりも高くなっている．これは，学年が上がるほど，Web ページを閲覧することに慣れているため，ページのサイズがある程度大きかったり，文がある程度長かったりしても，読みたさにそこまで影響を与えなかったものと考えられる．それに対して，中学年ではページのサイズが小さく，1 文も短い方がそのページを読みたいと思う傾向がある．

図 6 (b) に示す見やすさについては，高学年では，構成に関する指標，文章に関する指標の両方の指標で *NDCG* 値の向上が見られるが，中学年では，構造に関する指標のみで *NDCG* 値の向上が見られる．この結果より，学年が上がるにつれて，ページの見た目だけでなく，書かれている文章にも影響を受けることが分かる．

図 6 (c) に示す勉強になるかどうかについては，高学年，中学年とも多くの指標で Yahoo! JAPAN の精度を下回っているが，中学年の方が構成に関する指標の *NDCG* 値が高い．これは，高学年の方が，難解な文章が多いようなページを勉強になると見なす傾向があったためと考えられる．また，中学年は高学年と異なり，*DifficultExpression* の *NDCG* 値が高い結果となった．この結果より，学年が小さいほど，難解な表現を好まないものと考えられる．

図 6 (d) に示す内容の分かりやすさについては，中学年と高学年で結果が大きく異なる．高学年ではほとんどの指標で Yahoo! JAPAN の *NDCG* 値を下回っているのに対して，中学年では上回っており，提案手法が中学年に効果的であることが分かる．これは，学年が上がるほど，簡潔で親しみやすいページでなく，多くの説明を含むようなページを分かりやすいと見なす傾向があるためと考えられる．特に，中学年では，構成に関する指標の効果が高く，見た目が分かりやすさに影響を受けやすいことが分かる．

以上のように，中学年ほどページの見た目に影響を受けやすく，高学年になるほど見た目よりもページの中身に影響を受けやすくなることが分かる．このように，学年ごとに差が見られるため，学年に応じて指標のスコア化の方法，組合せ方を変更する必要があると考えられる．特に，分かりやすさについては，その影響が顕著であり，学年が上がるごとにページの文章への理解度が増すことを考慮することが必要である．

5. まとめと今後の課題

本稿では，子供の Web 検索を支援するために，検索結果を子供向けにリランクする手法を提案した．提案手法では，Web ページの文章量や色の数などの構成に関する指標，文の長さや子供向け表現の数などの文章に関する指標をスコア化する．そして，各指標のスコアを組み合わせて各ページのスコアを決定し，スコアの降順に検索結果を並べ替える．

提案手法の有効性を検証するため，小学生の子供 33 人に，6 個のクエリに関する 300 ページの評価を行ってもらったデータセットを作成し，そのデータセットを用いて評価実験を行った．その結果，子供向け表現を中心とした指標を組み合わせることで，子供にとっての“読みたさ”，“見た目の見やすさ”，“勉強になるか”，“内容の分かりやすさ” の 4 つの側面において，提案手法によるリランクは，Yahoo! JAPAN，Yahoo!きっずのランキングより *NDCG* 値が最大 20%向上した．この結果より，提案手法はランキング上位に子供向け度合いの高いページを多くランクでき，子供が検索結果から分かりやすいページを容易に選択できるランキングを実現できたことが分かる．

今後は，テキスト量，画像量，文章の難易度などの年齢に応じた適切な値を調査し，指標のスコア化の方法をさらに改善し，クエリや学年の特徴に従った各指標の最適な組合せ方法を検討する予定である．

謝辞　本研究の一部は，文部科学省グローバル COE プログラム（研究拠点形成費）の研究助成によるものである．ここに記して謝意を表す．

参 考 文 献

1) Bilal, D.: Children’s Use of the Yahooligans! Web Search Engine: I. Cognitive, Physical, and Affective Behaviors on Fact-based Search Tasks, *Journal of the American Society for Information Science*, Vol.51, No.7, pp.646–665 (2000).
2) Bilal, D. and Kirby, J.: Differences and Similarities in Information Seeking: Children and Adults as Web Users, *Information Processing and Management*, Vol.38, No.5, pp.649–670 (Sep. 2002).
3) Druin, A., Foss, E., Hatley, L., Golub, E., Guha, M.L., Fails, J. and Hutchinson, H.: How Children Search the Internet with Keyword Interfaces, *Proc. IDC 2009*, pp.89–96 (June 2009).
4) 富士通：富士通キッズコンテンツ作成ハンドブック (2007). http://jp.fujitsu.com/about/kids/handbook/
5) goo. http://www.goo.ne.jp/

6) goo リサーチ．http://research.goo.ne.jp/
7) Herbrich, R., Graepel, T. and Obermayer, K.: Large Margin Rank Boundaries for Ordinal Regression, *Advances in Large Margin Classifiers*, pp.115–132, MIT Press (2000).
8) 稲垣宣生，山根芳知，吉田光雄：統計学入門，裳華房 (Dec. 1992).
9) キッズ goo．http://kids.goo.ne.jp/
10) 前川卓也，原　隆浩，西尾章治郎：モバイル端末のための Web ページ自動スクロール方式，日本データベース学会 Letters，Vol.4, No.2, pp.29–32 (Sep. 2005).
11) 美馬秀樹，尹　泰聖：子供のためのウェブ情報検索支援システムの開発，情報処理学会夏のプログラミング・シンポジウム報告集，pp.17–23 (Aug. 2003).
12) Nakaoka, M., Shirota, Y. and Tanaka, K.: Web Information Retrieval Using Ontology for Children based on Their Lifestyles, *Proc. ICDEW 2005*, p.1260 (Apr. 2005).
13) Nielsen, J.: Usability of Websites for Children: 70 Design Guidlines based on Usability Studies with Kids, *Nielsen Norman Group Report* (2002).
14) Nielsen, J.: Teenagers on the Web: 61 Usability Guidelines for Creating Compelling Websites for Teens, *Nielsen Norman Group Report* (2005).
15) 菊地秀文，赤堀侃司：小学校情報教育における児童の Web ブラウジングの特徴分析，日本教育工学会論文誌，Vol.27, No.2, pp.143–153 (2003).
16) Sato, S., Matsuyoshi, S. and Kondoh, Y.: Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus, *Proc. LREC 2008*, pp.28–30 (May 2008).
17) 総務省．http://www.soumu.go.jp/johotsusintokei/field/tsuushin01.html
18) Yahoo!デベロッパーネットワーク．http://developer.yahoo.co.jp/
19) Yahoo! JAPAN. http://www.yahoo.co.jp/
20) Yahoo!きっず．http://kids.yahoo.co.jp/
21) 湯浅千映子：子ども向け文章の情報の配列—「小学生新聞」を対象に，文体論研究，Vol.52, No.52, pp.41–56 (Mar. 2006).

(平成 22 年 5 月 17 日受付)
(平成 22 年 11 月 5 日採録)

岩田　麻佑（学生会員）

2009 年大阪大学工学部電子情報エネルギー工学科卒業．現在，同大学大学院情報科学研究科博士後期課程在学中．Web 情報システムおよび検索技術に興味を持つ．日本データベース学会の学生会員．

荒瀬　由紀（正会員）

2006 年大阪大学工学部電子情報エネルギー工学科卒業．2007 年同大学大学院情報科学研究科博士前期課程修了．2010 年同博士後期課程修了．博士（情報科学）．同年 Microsoft Research Asia に入社し，Natural Language Computing Group に所属，現在に至る．日本語自然言語処理，特に統計的機械翻訳，またモバイル Web，モバイル端末のインタフェース，Web データマイニングの研究に従事．ACM 会員．

原　　隆浩（正会員）

1995 年大阪大学工学部情報システム工学科卒業．1997 年同大学大学院工学研究科博士前期課程修了．同年同大学院工学研究科博士後期課程中退後，同大学院工学研究科情報システム工学専攻助手，2004 年より同大学院情報科学研究科マルチメディア工学専攻准教授となり，現在に至る．工学博士．2000 年電気通信普及財団テレコムシステム技術賞受賞．2003 年本学会研究開発奨励賞受賞．2008 年，2009 年本学会論文賞受賞．データベースシステム，分散処理の研究に従事．IEEE，ACM，電子情報通信学会，日本データベース学会の各会員．

西尾章治郎（フェロー）

1975 年京都大学工学部数理工学科卒業．1980 年同大学大学院工学研究科博士後期課程修了．工学博士．京都大学工学部助手，大阪大学基礎工学部および情報処理教育センター助教授，大阪大学大学院工学研究科情報システム工学専攻教授を経て，2002 年より大阪大学大学院情報科学研究科マルチメディア工学専攻教授となり，現在に至る．2000 年より大阪大学サイバーメディアセンター長，2003 年より大阪大学大学院情報科学研究科長，その後 2007 年より大阪大学理事・副学長に就任．この間，カナダ・ウォータールー大学，ビクトリア大学客員．データベース，マルチメディアシステムの研究に従事．現在，Data & Knowledge Engineering 等の論文誌編集委員．本会理事を歴任．本会論文賞を受賞．電子情報通信学会フェローを含め，ACM，IEEE 等 8 学会の各会員．