# 漢字難易度を利用した Web ページのランキング提示手法

友廣 翔太 [1]　　藏 培慶 [2]　　椎名 広光 [3]

[1,2] 岡山理科大学大学院 総合情報研究科 情報科学専攻

[3] 岡山理科大学 総合情報学部 情報科学科

{i12im07ts[1], im12im05zb[2]}@st.ous.ac.jp, shiina@mis.ous.ac.jp[3]

## 1　はじめに

現在，多くの検索エンジンではユーザの検索履歴やページの閲覧履歴などを利用することでユーザの興味や関心に合った Web ページの順位を上げる検索結果のパーソナライズが行われている．このような手法は検索エンジンだけでなく，オンライン通販サイトでの商品の推奨等にも広く利用されている．一方で，学生や留学生等のような語学力が成長段階にあるユーザや，学習目的に Web 検索を利用するユーザに対しては，興味や関心といった要因以外に，ユーザの基礎的な日本語の理解力に関する項目を考慮する必要があると考えられる．

そこで本研究では，多くの検索エンジンと同様に検索履歴やページの閲覧履歴を蓄積していき，ユーザに理解されやすい Web ページの漢字難易度の級別分布の特徴を日本語の難易度として抽出する．そして，新たに得られた検索候補の漢字難易度の分布の特徴がユーザに理解されるかどうかを確率的に評価することで，ユーザの日本語理解力に沿ったランキングに再編成することを目標としている．

## 2　Web ページの文章の難易度

本研究では Web ページの日本語の難度として利用する指標として，Web ページの文章中に出現する漢字の日本漢字能力検定（以下，漢検 [1]）の級別を利用する． 漢検の級別は 1 級から 10 級の 12 段階あり 1 級が最も難しく 10 級に近づくにつれて易しくなっていく．例えば， 図 1 では単語「単純ベイズ分類機」で検索した本システムの出力結果である．ただし， 検索候補の表示順については， Yahoo!API[2] で得られた結果のままである．


図 1: Web ページの学年別漢字出現比率分布

## 3　作成したシステム

本研究で作成したシステムでは， ユーザが閲覧した Web ページの内容に対する理解についてのアンケートを行い，級別漢字の出現データを蓄積する．蓄積した漢字の出現データとユーザの Web ページの理解率を利用することで， Web ページのランキングを再構成してから提示している．作成したシステム (図 2) は，つぎの 4 つの手順で処理を行う．

(1)Yahoo!API の Web 検索 API から検索ワードに関する Web ページを 10 件抽出する．

(2)Web ページに含まれる文から漢検の難易度 [3] で分けた漢字の出現データを求める．

(3) 漢字の出現データをユーザモデルに当てはめ， 事前分布， 尤度から Web ページの理解確率を計算し，確率が高い順にランキングを出力する．

(4) 提供しているランキングの閲覧と理解に関するアンケートから事前確率を更新し，次の Web ページの

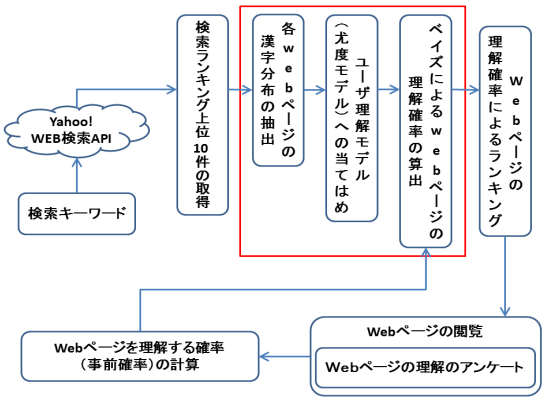


図 2: システム構成

理解確率の計算に利用する．

## 4 Web ページの理解確率

本研究ではユーザの Web ページの閲覧履歴の漢字難易度の級別分布を利用することで各 Web ページの理解確率を求める過程でベイズ推定 [4, 5] を利用した．以下に本研究におけるベイズ推定のそれぞれの式の意味を説明する．

(1) 事前確率 $P(A)$:

ユーザが任意のページを理解する確率を示す．ユーザが Web ページを閲覧した際に，その日本語を理解できたか否かのアンケートを実施することでデータを更新する．なお，$A$ はユーザの Web ページの日本語に対する理解状況を表す確率変数で，$A$=0 ならば理解できない，$A$=1 ならば理解できることを示すカテゴリカル変数である．

(2) 証拠 $P(F)$:

ユーザの閲覧した Web ページから級別漢字の出現パターン $F$ が得られる確率．これはユーザが過去に閲覧した Web ページの級別漢字の出現データを基に計算される．

級別漢字の出現パターン $F$ とは 1 級から 10 級の漢字が出現するか否かを表す確率変数 $f_1, ..., f_{12}$ をまとめてベクトル化したものである．各 $f_n$ はそれぞれ $n$ に対応した級の漢字が $f_n = 0$ ならば出現しない，$f_n = 1$ ならば出現することを表している．

(3) 尤度 $P(F|A)$:

ユーザの理解状況を条件付けた上で級別漢字の出現パターン $F$ が得られる確率を表す．各級の漢字の出現の独立性を仮定して，$P(F|A)$ は各 $P(f_n|A)$ の同時確率として表すこととしている．

$$P(F|A) = \prod_i P(f_i|A)$$

(4) 事後確率 $P(A|F)$:

Web ページの級別漢字の出現パターン $F$ が与えられたとき，その Web ページが理解されるかどうかの確率を表す．

事後確率 $P(A|F)$ は，ベイズの定理より事前確率，尤度，証拠を用いて次式のように表す．

$$P(A|F) = \frac{P(A) \cdot P(F|A)}{P(F)}$$

ここで尤度については上述したような理由から，式の書き換えが可能である．また，証拠についてもユーザが Web ページを理解するしないに関わらず同じ値をとる (確率変数 $A$ に依存しない) ため定数的な扱いをすることができる．これらを踏まえて式を書き換えると以下の式になる．

$$P(A|F) \propto P(A) \cdot \prod_i P(f_i|A)$$

本研究では，この事後確率が高い Web ページを検索ランキングの上位に表示することで，ユーザが理解する確率の高い $F$ を持つ Web ページほど，ユーザに閲覧されやすい状態にするのが目的である．

## 5 ユーザ理解モデル

ユーザ理解モデルを漢字のレベルを用いて測るが，本研究の離散型モデルを採用し，漢検の漢字の各レベルの理解の有無を 0,1 で測るモデルを利用している．

例えば，図 1 の漢字級別頻度分布は，図 3 の有無 (0,1) に置き換えられる (頻度分布 $(0,0,0,0,0,1,1,7,7,1,6,2)$ を $F = (0,0,0,0,0,1,1,1,1,1,1,1)$ に置き換えている)．

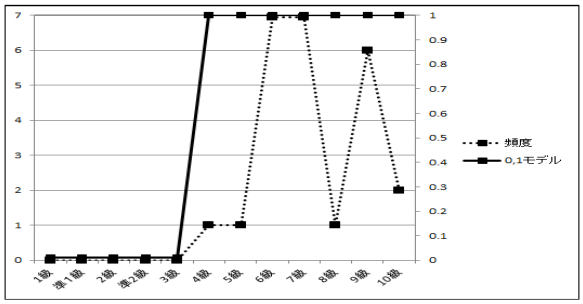


図 3: ユーザ理解モデル

表 1: ユーザによる Web ページの順位の相違

| 検索語 | ベイズ分類器 | | | | 日本語能力試験 | | | | 岡山理科大学 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Yahoo | 日本人学生 | | 留学生 | | 日本人学生 | | 留学生 | | 日本人学生 | | 留学生 | |
| 順位 | 順位 | 理解確率 | 順位 | 理解確率 | 順位 | 理解確率 | 順位 | 理解確率 | 順位 | 理解確率 | 順位 | 理解確率 |
| 1 | 7 | 0.054156086 | 3 | 0.44091764 | 4 | 0.132262884 | 10 | 0.286755559 | 5 | 0.126286688 | 7 | 0.218988189 |
| 2 | 5 | 0.064248443 | 1 | 0.508336593 | 3 | 0.139970473 | 8 | 0.394921525 | 10 | 0.021717842 | 2 | 0.600771970 |
| 3 | 8 | 0.054019617 | 4 | 0.440237172 | 1 | 0.279661241 | 5 | 0.446250584 | 1 | 0.519268962 | 10 | 0.039913539 |
| 4 | 9 | 0.053928668 | 5 | 0.440010085 | 2 | 0.244078789 | 9 | 0.380253896 | 2 | 0.222535040 | 8 | 0.179976514 |
| 5 | 3 | 0.277373317 | 8 | 0.059087490 | 5 | 0.063074320 | 1 | 0.503938411 | 8 | 0.050163042 | 3 | 0.562737937 |
| 6 | 4 | 0.101602606 | 6 | 0.424849333 | 9 | 0.053065861 | 6 | 0.435899759 | 4 | 0.132150222 | 6 | 0.447532207 |
| 7 | 1 | 0.397934949 | 10 | 0.006190131 | 6 | 0.062914443 | 2 | 0.503238518 | 9 | 0.025792596 | 1 | 0.620288766 |
| 8 | 2 | 0.368635071 | 9 | 0.014673617 | 7 | 0.062861163 | 3 | 0.502771091 | 6 | 0.070922265 | 4 | 0.506960592 |
| 9 | 10 | 0.022124077 | 7 | 0.362484794 | 8 | 0.062754621 | 4 | 0.502537128 | 3 | 0.150949274 | 9 | 0.072581915 |
| 10 | 6 | 0.063500929 | 2 | 0.505565696 | 10 | 0.052788127 | 7 | 0.434519997 | 7 | 0.070741558 | 5 | 0.506267135 |

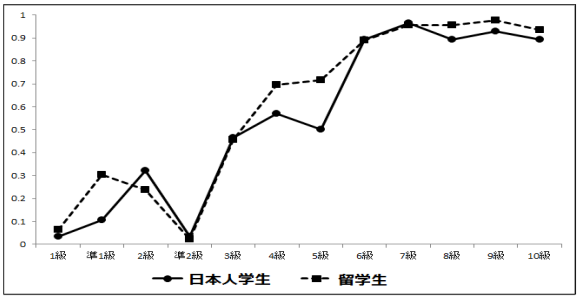


図 4: ユーザの理解分布 (尤度グラフ)

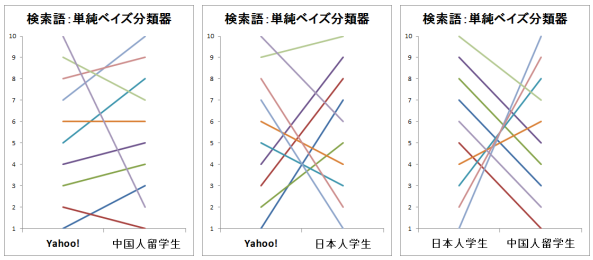


図 5: 検索語 “ 単純ベイズ分類器 ”

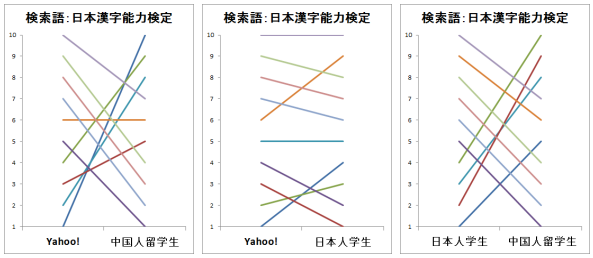


図 6: 検索語 “ 日本の漢字能力検定 ”

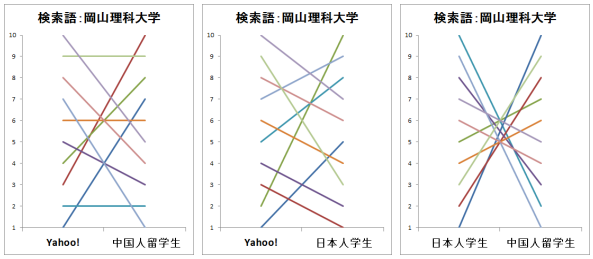


図 7: 検索語 “ 岡山理科大学 ”

# 6 評価実験

本研究では， 日本人学生と留学生 (中国) による評価を行った.

## 6.1 評価方法

(1) 評価実験の方法としては， 初めに自由に検索し 50 件の Web ページを閲覧する．それぞれの Web ページに対して理解できたできないの情報を収集し，それぞれの級別漢字の出現率より尤度 $P(F|A) = \prod_i P(f_i|A)$ を作成し，事後確率を計算している.

図 4 は日本人学生と留学生から収集した結果より得られた尤度 $P(f_i = 1|A = 1), i = 1 \ldots 12$ のグラフである.

(2) それぞれのユーザで共通の検索語を検索し， Web ページのランキングの相違を比較している．ここでは， 本稿では検索語に 3 種類 “ 単純ベイズ分類器 ”, “ 日本の漢字能力検定 ”, “ 岡山理科大学 ” を用いて Yahoo!API, 日本人学生， 留学生の Web ページのランキングを比較している． ランキングと理解確率の表を表 1 に示し， それぞれのランクの変動を図示したものを図 5,6,7 に示す.

### 6.2 評価実験に対する考察

図 4 のユーザの理解分布 (尤度グラフ) より日本人学生と留学生が理解した Web ページの級別漢字の出現率には，それほど大きな差は見られない．しかし実際は，図 5,6,7 に見られるように，それぞれのランキングを比較すると大きな差が見られる．これについてはユーザの Web ページ閲覧履歴の級別漢字の出現データを利用して，検索ランキングをパーソナライズするという目的が概ね達成できてるのではないかと考える．なお， ユーザの理解分布 (尤度グラフ) や検索語による Web ページのランキングは，これを見る限り日本人と留学生の母語の違いより， アンケートに答えた学生の傾向によって相違するとも考えられる．

検索語自体が持つ漢字級別に着目して見ると,「岡山理科大学」という単語は 10 級漢字と 9 級漢字のみで,「日本漢字能力検定」は 10，8，6 級構成されている．これらの級別の漢字は日本人学生と留学生間では出現率はほとんど変わらない級別の漢字である．一方,「単純ベイズ分類器」は 9，7，5 級と日本人学生と留学生とで出現率が大きく違う 5 級漢字を持っている．しかし，順位の変動を見てみると「単純ベイズ分類器」と「日本漢字能力検定」の方がむしろ近い順位変動をしている． このことから検索語に含まれる漢字の級別の影響についてはあまり考える必要がないのではないかと考えられる．

## 7 おわりに

本研究を通して，ユーザが閲覧した Web ページ中の級別漢字の出現データだけも検索ランキングをある程度パーソナライズが可能なことが分かった．

一方で，パーソナライズされたランキングが本当にユーザの日本語能力に沿ったものなのかという妥当性を問うような研究はほとんど行っていない．今後はより能力に応じたランキングのパーソナライズを行うことを目標に漢字級以外の情報も日本語理解力のパラメータとして取り入れていきたい．例えば，単語の難易度の推定 [6] などを利用した単語の知識，助詞の運用に関する能力などの語彙能力を踏まえた日本語能力の情報を反映させることを考えている．

また， 今回は日本人学生と中国人留学生という明らかに日本語能力に差がある二者を対象にランキングの違いを示したが，ほぼ同程度の日本語能力を持つと考えられる日本人同士のランキングを比較も必要であると考えている．

## 参考文献


[1] 財団法人　日本漢字能力試験 http://www.kanken.or.jp

[2] Yahoo! Japan デベロッパーネットワーク, http://developer.yahoo.co.jp/

[3] 徳弘:日本語学習のためのよく使う順漢字 2100, 三省堂, 2008.

[4] C. M. Bishop, “ Pattern Recognition and Machine Learning ”, Springer, 2006.

[5] R. O. Duda, P. E. Hart and D. G. Stork, “ Pattern Classification ”, Wiley Inter-Science, 2000.

[6] K.Nakanishi, N.Kobayashi, H.Shiina, F.Kitagawa Estimating word difficulty using semantic descriptions in dictionaries and Web data, 2012 IIAI International Conference on Advanced Applied Informatics, pp324-329,2012.