1P2-4in

検索新聞:可読性に着目した検索情報提示システム

Kensaku Shimbun: A newspaper-style information retrieval system using readability score

関谷 英樹[*1] 祖父江 翔[*2] 浅倉 優介[*2] 田村 哲嗣[*1] 速水 悟[*1]
Hideki Sekiya Sho Sobue Yusuke Asakura Satoshi Tamura Satoru Hayamizu

[*1] 岐阜大学工学部
Faculty of Engineering , Gifu University

[*2] 岐阜大学大学院工学研究科
Graduate School of Engineering , Gifu University

We have been developing a newspaper-style information retrieval system "Kensaku-shimbun", that supports users to retrieve the information and easily understand the results. This paper improves our system by employing "readability" score in the document ranking module in the system; the readability score is obtained by using a textbook corpus. Finally, the effectiveness of the system is investigated by the subjective experiments.

1. はじめに

近年インターネットが普及するにつれ，様々な情報が Web 上にあふれ存在している．その中で，Web を使って情報を得るための手段であるサーチエンジンが多く用いられているが，その情報量は膨大であり，ユーザは複数のページを閲覧しなければ有益な情報を得られないこともある．

また，誰でも気軽に Web 上に文書をアップロードできるようになり，その内容も多岐にわたる．そのため，検索結果には不必要な情報が混濁していることも多々あり，ユーザが望む情報の取得と理解には多くの労力や時間を要してしまう．無論，インターネットに慣れているユーザならば，比較的速やかに望む情報を得ることができる．しかし，そうでないユーザの場合，望む情報をすぐに得ることは困難だと言える．そこで我々は，検索結果を新聞形式にまとめユーザに提示する検索新聞システム(以下本システム)を開発している．

しかし，従来のシステムの提示結果において，文書選択の際の重要度判定の尺度である難易度(可読性)の評価が悪く，システム全体の評価が下がってしまったことが問題点として挙げられた．そこで，本稿では「可読性」に注目し，提示結果の改善を目的としたシステムの改良について示す．

2.「検索新聞」システム概要

検索新聞は，ユーザが検索キーワードを入力することにより Web 上から取得した文書に対し，複数の特徴量を用いて重要度の推定を行い，その結果を新聞形式でユーザに提示するシステムである． 新聞形式でユーザへ提示する利点として，新聞に慣れ親しんでいるユーザに対し読みやすい結果の提示が可能なこと，まとまった形式であるため冗長な情報提示を避けられることなどが挙げられる．

本システムの概要を図 1 に示す．ユーザにとって重要な文書を選択するための特徴量として，3 種類の重要度の評価尺度により文書の重要度の推定を行う．本システムでは，Web 情報の取得を Yahoo!API により行う．

連絡先:関谷英樹
岐阜大学大学院工学研究科応用情報学専攻
速水・田村研究室，岐阜県岐阜市柳戸 1-1，058-293-2763
sekiya@asr.info.gifu-u.ac.jp

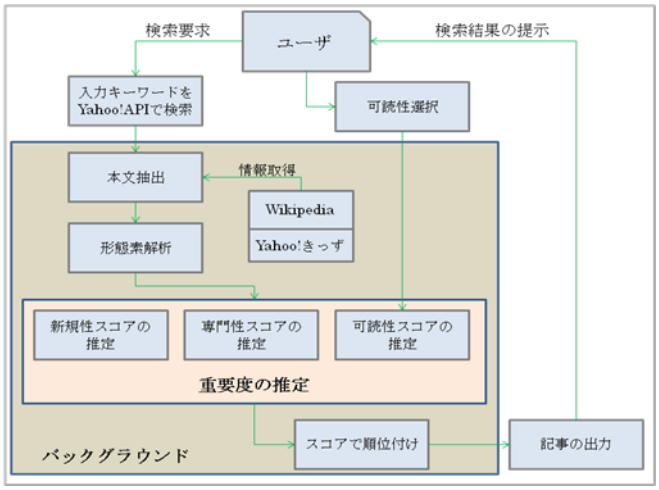

図 1 「検索新聞」システム概要

3. 重要度スコア

検索の自動化や効率化により，ユーザの検索における負担軽減を図ることが可能である．しかし，どの情報がより重要なものであるかは常に変化する不確定な尺度である．そこで，Web 検索で取得した文書に対して重要度の推定を行う．重要度の尺度としては，単語の出現頻度をもとに重要度を算出する手法である，TF*IDF が多く用いられているが，重要度を単一で定めることは困難であり，多角的に判断する必要がある．本システムでは 3 つの評価尺度「専門性，可読性，新規性」を用いることにより，文書の重要度推定を行う．以下に各重要度について示す．

3.1 専門性

利用者が情報を重要と位置付ける要素の 1 つとして，その情報がどの程度専門的な内容を含んでいるか，という指標を考える．その尺度を本論文では「専門性」と呼ぶことにする．ある単語が一般的ではなく，一部の専門分野でのみ使われることが多い場合や，あるジャンルに偏って多く出現する場合に専門的であると考えられる．

専門性のスコア付けとして，CRF(Conditional Random Fields)を使用したスコア付け方法を用いる．CRF は条件付き確率に基づく推定モデルで，入力特徴量に制約がなく定

義しやすい点と推定確率を直接推定できるという特徴がある．本研究では，毎日新聞で定められたジャンルの中から 8 ジャンル（文化，経済，娯楽，国際，科学，社会，スポーツ，家庭）に対し，ジャンルごとに言語モデルを作成する．また，本研究における CRF は，入力特徴量を入力形態素の情報(対象の形態素より前後 2 形態素の表層形および品詞名)，出力ラベルを形態素ごとの各推定ジャンルの確率とする．

スコア付けには，形態素の中から「名詞」，「動詞」を用いる．例えば，入力文書が「ウィンブルドンでテニスの大会がある。」の場合を考える．この文書に CRF を適用し「名詞」，「動詞」に着目すると，図 2 のような結果が得られる．

ウィンブルドン	スポーツ 0.8, 国際 0.15, 社会 0.021,…
テニス	スポーツ 0.9, 家庭 0.08, 社会 0.02
大会	スポーツ 0.5, 国際 0.2, 科学 0.15, …
ある	科学 0.3, スポーツ 0.2, 国際 0.2, …

図 2 CRF を使用したスコア付けの結果例

この結果のスコアをジャンル毎に加算し，単語数との商を計算する．一番大きいジャンルのスコアをその文書のスコアとする．

3.2 可読性

情報がどの程度複雑な内容を含んでいるか，どの程度読みやすいかという指標を考え，その尺度を「可読性」と呼ぶ．ユーザがキーワードを入力すると同時に，3 種類の可読性「やさしい」，「ふつう」，「詳細」から 1 つを選択できる．この選択により文書のスコア付けの際の重みが変化し，新聞紙面に提示される文書が変わる．文書のスコア付けには，先行研究である教科書コーパスを用いたテキストの難易度推定を利用した．

3.2.1. 教科書コーパスを用いた日本語テキストの難易度推定

この手法は，テキストの難易度を表す区分として学年区分を使用するものである．テキストの難易度は，小学 1 年から高校 3 年までの 12 学年に，大学を加えた計 13 学年の段階区分で表わされる．教科書コーパスは，小・中・高の英語を除く全教科の教科書，大学は，教養課程の講義で使用されているものを用いる．小・中・高で用いられる各教科書は，その対象学年が比較的明確であり，文部科学省の検定を受けているため，用いられている言語表現も対象学年に適した難しさに調整されていると考えられる．難易度が既知である教科書を利用することで，幅広い学年において，漢字や語彙だけでなく言語表現も考慮した難易度推定が可能であると考えられる．

それぞれの区分の教科書をコーパスとして作成された bigram の言語モデルを用い，次式を用いて尤度を計算することでテキストの難易度を推定する．

$$L(Mi|T) = \sum_{z \in T} C(z,T)\, log\, P\,(z|Mi) \qquad (1)$$

ここで，$P(z|Mi)$ は各難易度クラスの言語モデル Mi における，連続する 2 単語 z が生起する確率，$C(z,T)$ は，入力された日本語テキスト T における z の出現回数，$L(Mi|T)$ は，言語モデル Mi における入力テキスト T の尤度である．本研究では，この手法で得られた 13 個の尤度のうち，最大の尤度を各文書のスコアとする．

3.2.2. 予備実験

本システムでは，ユーザが選択した可読性（「やさしい」，「ふつう」，「詳細」）と 3.2.1 を用い，文書ごとに可読性のスコア付けを行う．そこで，3.2.1 により得られる推定結果をどの可読性と対応付けるのが適切であるかについて評価を行った．

被験者は 18 人，評価文書はランダムに選んだ 16 文書（中学 2 年:3 文書，中学 3 年:3 文書，高校 1 年:3 文書，高校 2 年:3 文書，高校 3 年:1 文書，大学:3 文書）であり，それぞれどの可読性に当てはまるかをアンケートにより評価してもらった．各文書の難易度はあらかじめ知らせていない．結果を図 3 に示す．

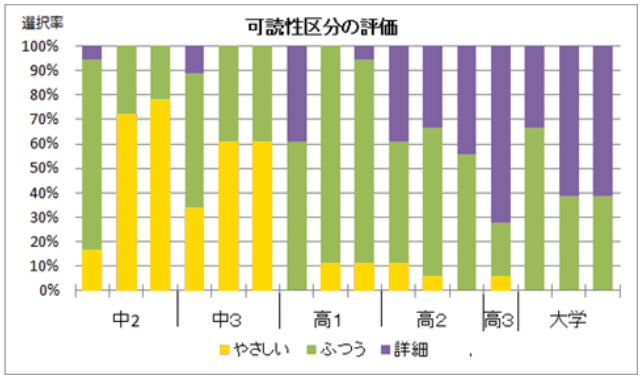

図 3 可読性区分の評価結果

縦軸は 18 人を 100%とした被験者の人数のパーセンテージを示し，各棒グラフは，横軸の学年を正解とした各文書を示している．文書により差はあるものの，「やさしい」と評価される文書は中学 2 年から中学 3 年，「ふつう」は高校 1 年から高校 2 年，「詳細」は高校 3 年から大学と推定される文書が選ばれていることが分かる．この結果により，ユーザの可読性選択により重要度を上げる文書を図 4 の通りに設定した．

「やさしい」	:	～中学 3 年
「ふつう」	:	高校 1，2 年
「詳細」	:	高校 3 年，大学

図 4 可読性と推定結果の対応

3.2.3. Yahoo!きっず・Wikipedia 文書の利用

3.2.1 によりテキストの難易度を求めることができる．しかし，本システムの検索は一般のニュースや Web 記事を対象としているため，難易度推定の際に，低学年（小学生から中学生前半）向けであると推定される文書は少ないという問題がある．そこで，小・中学生のためのポータルサイトである Yahoo!きっずの「こどもニュース」，「用語集」(話題になった用語の解説)の文書を用いる．これは，ユーザが可読性「やさしい」を選択した場合に提示し，ユーザが理解し易い文書の提示を行うことを目的とする．

また，入力キーワードに対するユーザの理解がより深まることを期待し，ニュース検索などでは取得できない，キーワードに対する説明を Wikipedia から取得する．キーワードの説明として，「第一文」，「概要」，「関連項目」の 3 つに注目した．「第一文」は，Wikipedia の冒頭に表示される「”キーワード”とは～である。」の部分を指し，キーワードについての簡単な説明が書かれてお

り内容がつかみ易い.「概要」では，記事の要点が示されており，第一文よりも詳しい説明がされている.「関連項目」は，キーワードに対する関連語が示されている部分である. これらを提示することにより，ユーザの興味を広げ，関連語を次の検索へ役立てることができると考える.

3.3 新規性

情報がどの程度流行を反映しているか，という指標を考え，その尺度を「新規性」と呼ぶ. これは，「ユーザが新しい情報を望む傾向にある」という考えに基づいている. 取得した Web ページがどの程度流行を反映しているか推定するため，ブログ内の出現単語をランキングにした kizasi.jp のツールである kizAPI を利用し，上位 30 キーワードと，共起の関連語情報を使用する. kizasi.jp は，ブログに出現する単語の出現情報を見ることにより，注目キーワードの動向を確認することのできるサイトである.

キーワードのランキング順位でスコアを調節し，共起の関連語にはそのキーワードの 10 分の 1 のスコアを与える. 入力文書の中に，きざしランキングによりスコア付けされた単語がある場合，入力文書のスコアにその単語のスコアを加算していく. こうして，すべての単語について加算したスコアとその文書の単語数の商を，その文書の最終的なスコアとする.

3.4 重要度スコアの統合

総合的な重要度のスコアを算出するため，正規化した 3 種類の尺度を 2 式で統合し，スコアが高い文書から順に提示する.

$$S(T) = \alpha * Exp(T) + \beta * Rea(T) + \gamma * Nov(T) \qquad (2)$$

ここで，$S(T)$ は入力されたテキスト T の重要度の総合スコア，$Exp(T)$ ，$Rea(T)$ ，$Nov(T)$ はそれぞれ「専門性」，「可読性」，「新規性」の正規化済みスコア，α ，β ，γ は各重要度の重み付けに用いる値である. 本研究では，可読性の選択による提示文書の変化を調査し，適度な変化を持たせるため，重みの値を $\alpha = 1$ ，$\beta = 1.5$ ，$\gamma = 1$ のように設定した. また，重要な文は文書の初めの方に存在し易いという考えに基づき，文書が長い場合は「。」までを 1 区切りとし，400 文字以内に納まるよう残りを切り捨てる.

また，選択された可読性により表示結果を変化させる. 3.2 の結果をもとに，「やさしい」が選択された場合は図 5,「ふつう，詳細」が選択された場合は図 6 の優先度で文書を提示する.

1. Yahoo!きっず　用語
2. Yahoo!きっず　こどもニュース
3. Wikipedia　第一文
4. Wikipedia　関連項目
5. 重要度スコア順に重みを付ける「やさしい」>「ふつう」>「詳細」

図 5 「やさしい」の優先度

1. Wikipedia　第一文
2. Wikipedia　概要
3. Wikipedia　関連項目
4. 重要度スコア順に重みを付ける「選択可読性」>「他の可読性」

図 6「ふつう・詳細」の優先度

本システムにより提示される新聞画面の例を図 7 に示す. これは，キーワード「インフルエンザ」により提示された紙面である. 右下部には，キーワードと kizasi.jp から取得した現在の流行ワードが書き出される. また，Wikipedia，Yahoo!きっずの記事のタイトルを赤色，Web 記事のタイトルを青色，ニュース記事のタイトルを緑色，画像検索の記事のタイトルを灰色で提示した.

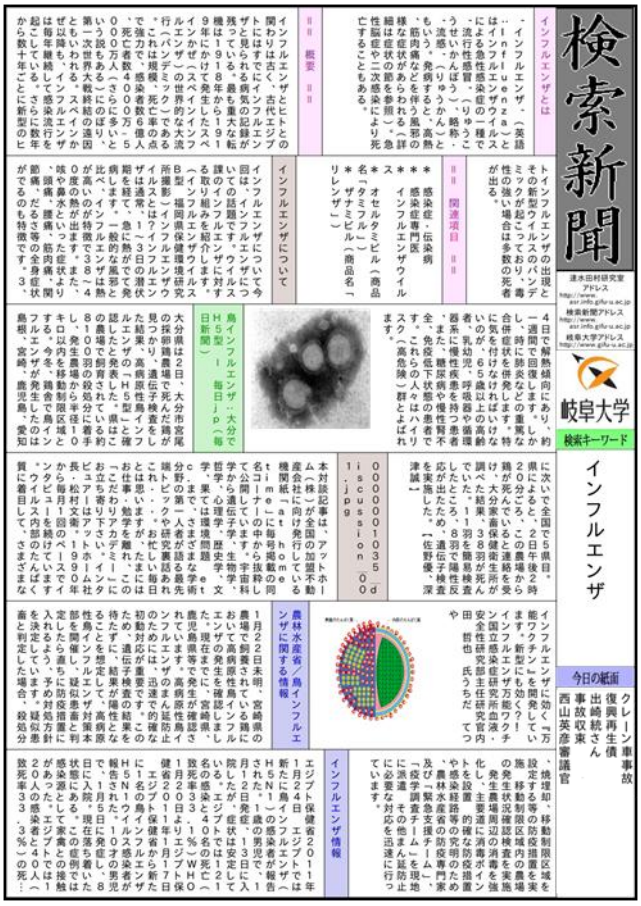

図 7 提示結果の例

4. システム評価実験

Yahoo!API を用いて取得した各文書に対し，3 種類の重要度スコアをそれぞれ推定する. スコアを統合し，重要度スコアが高い上位 6 件の記事を用い，ユーザに提示する記事の選択を行うモジュールの評価を行った. 被験者 18 人に対して以下の 3 つの内容で評価を行った.

(1) 以前の検索新聞と提示結果を比較し，どちらが有用か. さらに，Wikipedia の情報を提示することにより，キーワードに対する理解が深まるか.

(2) 可読性「やさしい，ふつう，詳細」の結果が適切に提示できているか.

(3) 関連項目を提示することが有用であるか，また，関連語に興味を持つことができるか.

4.2 従来のシステムとの比較

(1)では，従来の検索新聞との提示結果の比較，Wikipedia の情報を提示することの有用性の評価を行う. 評価条件として，キーワード「オリンピック」により提示される，従来の検索新聞と提案手法（可読性「ふつう」）の結果において，「キーワードそのもの」の情報に対し，より理解を深めることができるのはどちらか. 有用な情報を含んでおり，取得したい結果はどちらか，について評価を行った. それぞれの結果を図 8，図 9 に示す.

これらの図は，18 人を 100%とする被験者の選択をパーセンテージで示している. A が従来の検索新聞，B が本システムにより提示された結果である. 図 8 では，一般的なキーワード「オリン

ピック」に対し，約 60%の被験者が本システムの方がキーワードに対する理解を深めることができると評価している．この結果は，よく知られている一般的なキーワードに対しても説明文を提示することの有用性を示している．さらに，図 9 では 100%の被験者が，本システムが有用な情報を含んでおり結果を取得したいと評価している．これは，可読性のシステムを取り入れたことで理解のし易い文書を提示できているためであると考えられる．

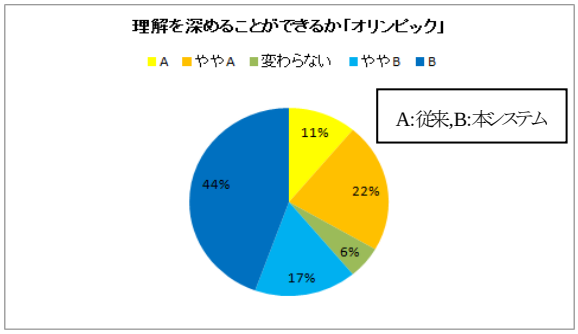

図 8 理解評価

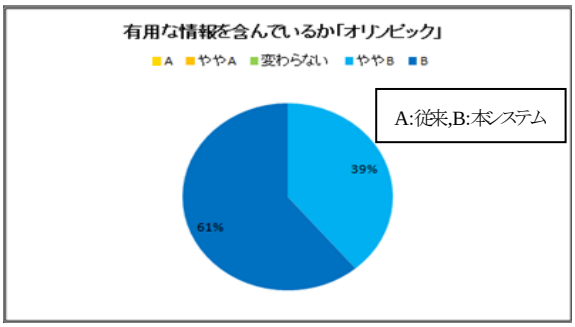

図 9 有用性評価

4.3 可読性による提示結果の評価

(2)では，可読性選択により結果を適切に提示できているかの評価を行う．キーワード「レアアース」により提示される，可読性「やさしい」，「ふつう」，「詳細」の結果において，それぞれどの可読性に当てはまるかを選択してもらった．被験者には，提示結果がどの可読性のものであるかは知らせていない．結果を図 10 に示す．これらの図の縦軸は，18 人を 100%とする被験者の人数をパーセンテージで示している．横軸は提示結果の正解の可読性であり，その下にはそれぞれの可読性を選択した人数を示している．半数以上の被験者が正解の可読性を選択している．これより，可読性の変更により提示文書を適切に変化させることができていると考えられる．

4.4 関連項目の評価

(3)では，関連項目を提示することの有用性についての評価を行う．評価条件として，被験者 12 人に対し，2 つのキーワード「クラスター爆弾」に関して，関連語に興味を持つことができるか，関連項目の提示は有用であるかについて 5 段階で評価を行った．結果を図 11 に示す．縦軸は選択した被験者の人数を，横軸は選択項目を示している．図の通り，関連項目の提示が有用であるという意見が多数であった．特に一般的ではないキーワードに対して関連項目の提示を行うことが有用であると考えられる．今回は，キーワードを指定して評価を行ったが，ユーザ自身が本システムを利用しキーワード入力を行うことで，各ユーザにとってさらに興味を持つことのできる有用な情報となると考える．

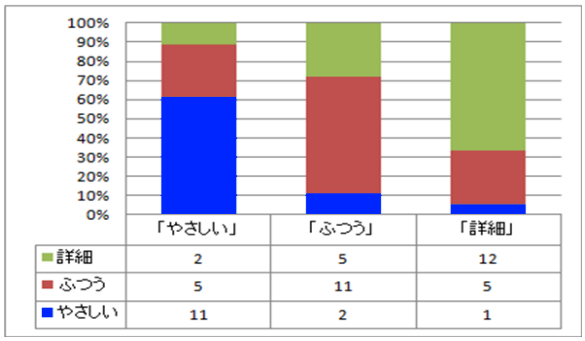

図 10 可読性評価

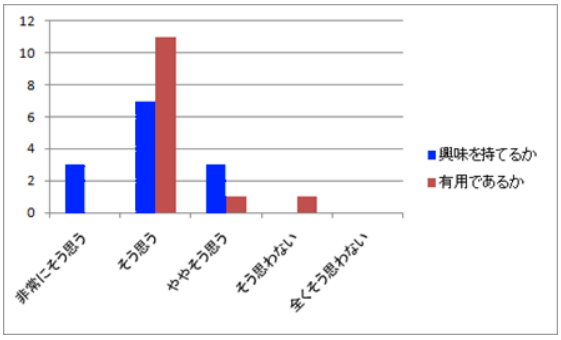

図 11 関連項目の評価

5. まとめ

本研究では，昨年構築された検索新聞の提示結果を改善させるため，テキストの読みやすさに着目し，システムを再構築した．テキストの読みやすさのスコアを算出する重要度の尺度である「可読性」を提案し，Wikipedia，Yahoo!きっずの情報と併せてユーザの可読性選択に応じて提示する文書を変化させた．また評価実験により，従来のシステムとの比較を行い提示結果の改善を示せた．また，可読性選択に応じて適切な難易度の結果を提示できていることが評価できた．

今後の課題として，より良い「検索新聞」を目指し，これ以外にユーザが重要と考える情報とは何かを模索し，ユーザの嗜好に合わせた結果を提示することが必要であると考える．また，難易度の異なる同意語，文の構成などを変更することにより，代替テキストを用意することなくテキストの可読性を変化させることが可能であると考えられる．

参考文献

[祖父江 2010] 祖父江 翔，瀬合 将士，山本 孝二，田村 哲嗣，速水 悟：検索新聞 ：新聞形式による検索情報要約システムの提案， 第 24 回人工知能学会全国大会，3D1-2, 2010.

[Lafferty 2001] John Lafferty, McCallum Andrew, Pereira Fernando. : Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceeding of the 18th International Conference on Machine Learning, 2001.

[近藤 2008] 近藤 陽介，松吉 俊， 佐藤 理史 ：教科書コーパスを用いた日本語テキストの難易度推定，言語処理学会第 14 回年次大会, pp.1113-1116, 2008.