3D1-2

検索新聞：新聞形式による検索情報要約システムの提案

Kensaku shimbun : A summarization system for information retrieval

祖父江 翔[*1] Sho Sobue　瀬合 将士[*2] Masashi Sego　山本 孝二[*2] Kouji Yamamoto　田村 哲嗣[*2] Satoshi Tamura　速水 悟[*2] Satoru Hayamizu

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

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

This work proposes a framework to generate content by using the search results of web pages for user's input in a newspaper style. The content consists of summarized documents of retrieved articles that are adopted according to three importance scores, i.e. specialty, novelty and readability; the specialty is based on domain-specific unigram scores of every word in the article. The novelty is estimated by the upload time and the trend information of keywords in the article. The readability is based on the similarity with commentary articles. Not only web pages but also news articles and pictures are obtained using the same scheme. The output of the system is automatically generated as a portable document file (PDF). Evaluation experiments were conducted by several users, and then effective points as well as future works of the system are turned out.

1. はじめに

近年，インターネットの発展に伴い，誰でも気軽にブログなどの WEB サイトを作成できるようになった．そのため大量の WEB サイトが存在し，その内容も多岐にわたり，ユーザが望む情報を得ることが困難になってきた．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，どのような情報があるのかを理解していない状況において，その分野について造形を深めようとしている段階であれば，望む情報を得ることは困難だと言える．

本研究では，検索エンジンから得られた結果をもとに重要な情報の選択を行い，ユーザに見やすくすることを考慮した新聞紙面型の情報提示システム（検索新聞）の開発について述べる．

2. 重要度

2.1 3つの重要度の尺度

重要度の尺度としては TF*IDF が多く用いられている．これは単語の出現頻度を基に重要度を算出する手法である．しかし，重要度はユーザによって異なる指標であり，常に変化する不確定な尺度である．これ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．本研究では重要度の尺度を 3 種類に分割し，それぞれを「専門性（specialty）」，「難易度（readability）」，「新規性（novelty）」と定め，重要度推定を行う．

2.2 専門性（specialty）

「専門性」では，そのテキストがどの程度専門的な内容を含んでいるかを考慮する．その第一段階として，そのテキストがどのジャンルに属しているかを推定する必要がある．今回は 2008 年の毎日新聞の記事を利用し，「文化」や「経済」など，全 10 ジャンルの記事に対し，それぞれ TermExtract[前田]を使用しジャンルごとに unigram を作成した．

ここで作成した unigram を用い，(1)式より入力テキストと各ジャンルの unigram に対する尤度を計算する．尤度を判定するにあたり，入力テキストを MeCab[工藤]で形態素解析をして，ジャンルiに対するテキストTの尤度を算出する．

$$Li(T,i) = c(w) * \sum_{w \in T} \log \frac{S_i(w)}{M+N} \qquad (1)$$

$c(w)$はテキストT中に出現する単語wの出現数を表し，$S_i(w)$はジャンルiのモデルにおけるwのスコア，Mはモデルテキスト中の単語ののべ数，Nはモデルテキスト中の単語の異なり数をそれぞれ表す．この尤度をテキストのスコアと考え，最も高くなるジャンルを，入力テキストのジャンルと推定する．

2.3 難易度（readability）

「難易度」の推定は，そのテキストがどの程度難解な日本語を使用しているかで判断する．この研究については佐藤ら[佐藤 2007][佐藤 2008(a)][佐藤 2008(b)]が既に行っている．現段階では，その予備実験として 2008 年の毎日新聞の記事から「社説」を抽出し，それを他の記事と比較をする．「社説」との類似性について，(1)式を使用して尤度を算出し，この尤度を「難易度」のスコアとして代用する．

2.4 新規性（novelty）

「新規性」の判断においては，WEB ページが更新された時刻，話題になったキーワードがどの程度使用されているかを評価の対象とする．以下の 2 つの基準で求めたスコアを加算することで，新規性のスコアとした．

1. **更新時刻によるスコア**
   古くからある記事に比べ，新しく更新された WEB ページの方が重要であると考えられるためである．Yahoo! API [Yahoo] を利用し，検索結果とその WEB ページの更新時刻も同時に取得する．もし 1 カ月以内の記事であれば，この更新時刻と現在時刻との差をスコアとする．
2. **話題となっているキーワード**
   ブログの出現単語をランキングとしている kizasi.jp [kizasi]を利用する．ここに出現する単語，共起する単語を使用している WEB ページを重要と判断してスコアを付ける．

連絡先：祖父江翔
岐阜大学 大学院 工学研究科 応用情報学専攻
岐阜県岐阜市柳戸 1-1,E-mail s-sobue@asr.info.gifu-u.ac.jp

3. 本システムの概要

3.1 システム全体像

本システムの概要を図 1 に示す.

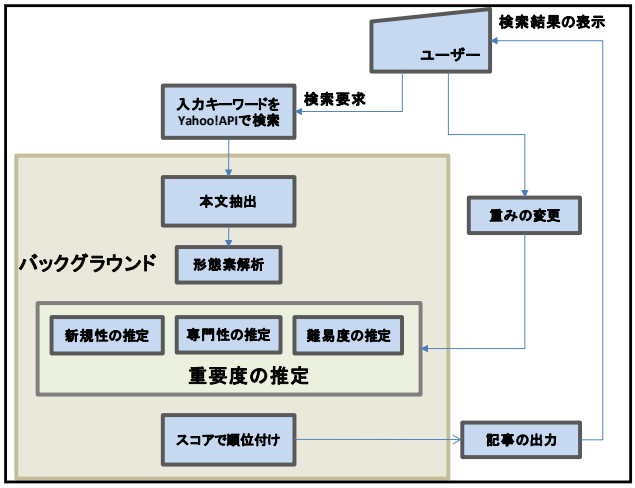

図 1　システムの概要

まず, ユーザが入力した検索キーワードに対し検索を行い, その検索結果で得られる各 URL から Web ページを取得する. その後に形態素解析して重要度を算出し, 検索キーワードの重要度の高い順に Web ページの順位をつける. その順位をもとに, 重要な文書から新聞紙面に書き出しを行い, 結果として出力された新聞紙面をユーザに提示するというものである. 新聞紙面は PDF で出力される.

3.2 ユーザへの出力画面

ユーザが検索キーワードを入力する画面を図 2 に示す.

図 2　キーワード入力画面

この画面ではキーワードの入力と共に, 重要度を判定するために使用する 3 つの尺度の重みと, 検索するキーワードのジャンルを選択することができる. 重みは全ての値の合計が 10 になる値の範囲で選択が可能である. ジャンルは, スポーツ, 社会, 科学, 読書, 国際, 家庭, 芸能, 経済, 文化, 解説の 11 種類から選ぶことができる. また, ジャンルを指定したくないというユーザのために, 指定なしという選択項目も用意している. ここで選択したジャンルは尺度の専門性に影響を与え, ジャンルによって重要であるとされる記事の内容が変化する. そのため, 検索キーワードや, ユーザが望む検索結果に沿う形のジャンルを選択することが望ましい. 新聞作成が終了すると, 画面の一番下に新聞記事の PDF へのリンクが現れ, 記事の PDF を閲覧することができる.

3.3 Web 文書の取得

検索結果の取得には Yahoo! API を用いた. Yahoo! API からニュース記事, Web 記事, 画像の 3 つ検索を行い結果を取得する. 取得する件数はそれぞれ最大 100 件ずつで, 最大で合計 300 件の取得結果を得る. 検索結果から取得するのは URL から取り出した各 Web ページであるが, 画像検索のみ, 画像 URL と HTML の 2 つを取得する. なお, 画像検索については, 画像の掲載されているサイトの URL から HTML を取得する.

取得した HTML からタグ除去を行い, Web ページの本文を取得する. ここで, 本文内の各文(段落)は「見出し」と「内容文」がセットになっていると仮定する.

3.4 尤度による重要度推定

Yahoo! API により取得した検索結果から本文抽出を行った後, まず 1 つの Web ページ内において各文について 2 章で述べたように重要度推定を行う. 重要であるとされた文は, その記事が書き出されるときに優先的に新聞紙面に書き出される. 重要文だと判定された文は Web ページの記事によって内容は様々であるが, 文の長さがあまりにも長く, 1 つの記事のみで紙面が埋まってしまう可能性もあるため, 一定の長さで文を切り取った. 文の文字数が一定数以上を超えたときにそれ以降は切り捨てている.

次に, Web ページ単体の重要度を推定し, 重要度の高い順に記事に順位を付ける. この順位によって新聞紙面に書き出される記事の順番が決まる.

3.5 新聞紙面作成

前節で作成した記事の順番をもとに新聞紙面に本文を書き出す. 画像検索で得られる結果について, 本来の Yahoo!検索における画像検索では画像のみが羅列された画面が検索結果として表示される. 本研究で提案する検索新聞は検索支援のシステムのため, 表示形式は元の検索結果に似たものにする. そのため, ランキング内に画像検索で取得された Web 文書があり, その文書を書き出す場合, HTML 内の本文の書き出しは行わずに画像の表示のみを行う. 画像に対しても重要度を求める処理を行うべきであるが, 今回は文の重要度を用いているので, 画像検索で取得された結果には Web ページ内の文書で重要度を評価し, 出力される際には画像が表示されるという形式になっている.

(1) 紙面レイアウト

本研究ではあらかじめ紙面レイアウトを作成し, 書き出す範囲や場所を固定している. 文章はフォントが明朝体でサイズが 10.0 ポイント, 縦方向 12 文字, 横に 6 段という形として記事が書き出される. 記事紙面の 1 枚の大きさは A4 の用紙と同じサイズである. 新聞紙面のレイアウトを図 3 に示す.

Web ページのタイトルである文章は, Web 検索で取得された記事の場合は水色の枠内に青色の文字, ニュース検索で取得された記事の場合は黄色の枠内に緑色の文字で表記した. 見出しと仮定された文章は赤文字で表記した. また, 右下の赤枠の部分には, 入力した検索キーワードと紙面に書き出された Web ページのタイトルをヘッドラインとして書き出した.

画像検索で取得された記事を書き出す際には画像の表示を行う. 今回は表示する画像のサイズを固定しており, 元の画像のサイズを考慮していないので, 画像によっては違和感のある表示がされることがある. 画像の下にはその画像のタイトルを横書きで表記した. 文が長い場合は三点リーダで省略している.

図 3　紙面レイアウト

(2)　出力ファイルの構造

PDF ファイル内に複数のページを作り，Web 検索とニュース検索と画像検索の全ての検索結果が表示される総合面，ニュース検索の結果のみが表示されるニュース検索面，Web 検索の結果のみが表示される Web 検索面，画像検索の結果のみが表示される画像検索という，複数の紙面を持った新聞記事の作成を行った．これは PDF を閲覧する画面の左上にあるしおりから各紙面へ移動することができる．

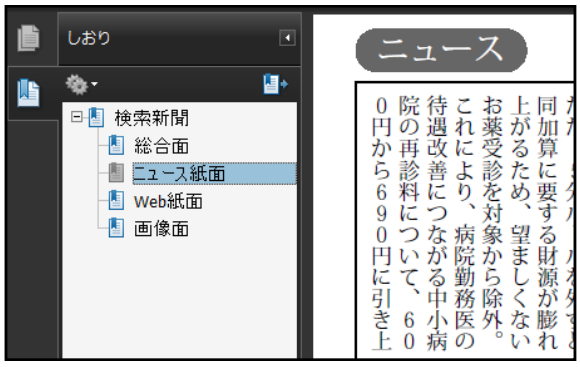

図 4　PDF 内のしおり

4.　実験

4.1　重要度の評価

重要度の尺度に関して有効性を確かめるために，TF*IDF を用いた場合との比較実験を行った．TF*IDF の手法では，検索キーワードの TF*IDF 値が高い記事を重要とした．

被験者 11 名に対して，『JAL』『オバマ』『オリンピック』『iPhone』の 4 種類のキーワードで作成された記事を評価してもらった．それぞれの記事は，あるジャンルに関連する記事を取得するようにした．『JAL』は「経済」に関して，『オバマ』では「社会」，『オリンピック』では「国際」，『iPhone』では「科学」に対応する．TF*IDF で取得した記事には，このジャンルの情報が付与されていない．また，今回の実験では，TF*IDF との比較の為に，それぞれの重みをできるだけ均等にして実験を行っている．「専門性」が 3，「難易度」が 3，「新規性」が 4 である．ある検索キーワードに対し，提案手法によって抽出された記事と，TF*IDF によって抽出された記事について，以下の 4 項目で比較してもらった．それぞれの結果を以下の図 5 から図 8 に示す．

1. どちらの記事が良いか(全体的な評価)
2. どちらがより専門的な内容を含んでいるか(専門性)
3. どちらが読みやすい記事か(難易度)
4. どの程度話題の情報を含んでいるか(新規性)

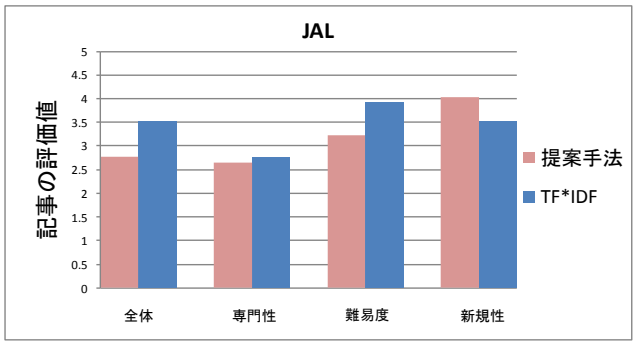

図 5　『JAL』で検索を行った結果

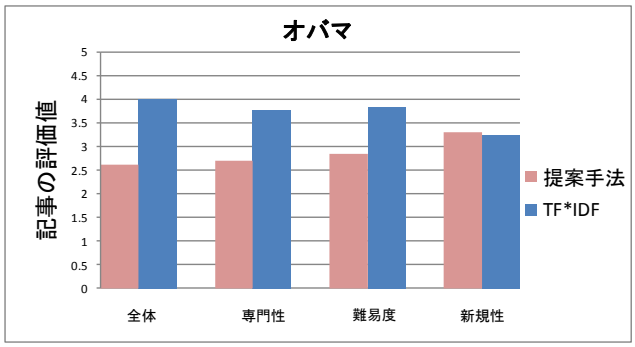

図 6　『オバマ』で検索を行った結果

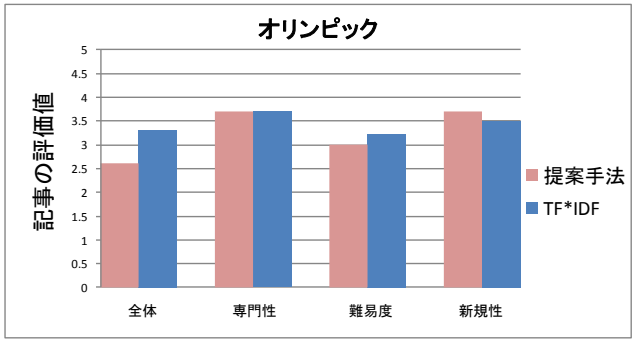

図 7　『オリンピック』で検索を行った結果

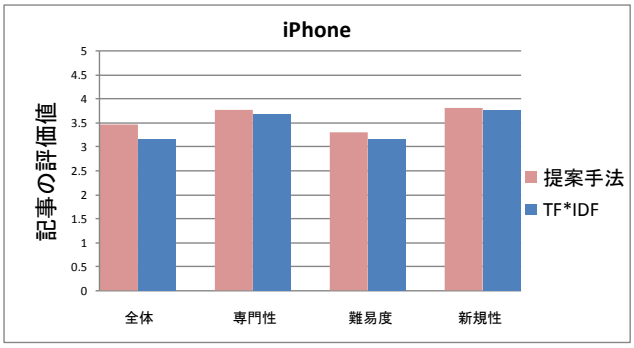

図 8　『iPhone』で検索を行った結果

4.2 結果・考察

図 8 にある通り，検索キーワードを『iPhone』にして行った時は，4 項目全てにおいて良いという評価を得た．また，その他の検索キーワードでも，「新規性」の評価は同等以上の結果を得ることができた．これは記事の更新時間に関する情報を利用することが有益であり，また，多くの人が話題にしている情報というのは，検索新聞のユーザにとっても興味が湧く情報であると言える．

しかし，それ以外のキーワードに関しては，優位性は確認できなかった．特に検索キーワードが『オバマ』であった場合に顕著である．「専門性」の場合，オバマ大統領に関係する情報で，かつ「社会」に関する情報を取得するようにしているが，政治に関する情報から話題が転換している，文章量が多すぎたため削られ，重要な情報が表示されないなどの問題があった．

また，現在は記事の書き出す位置や範囲は固定されており，本当にユーザにとって見やすい位置であるかなどの考慮はされていない．新聞紙面の見た目によって理解のしやすさが下がってしまうという意見も被験者から頂いた．文の内容だけではなく，文のレイアウトも考慮してユーザへの理解を高める必要がある．

その他に，重要な記事を選別できていても，重要な文が推定できていないこともあった．例として，キーワードを『民主党』として作成した紙面を図 9 に示す．ここでは一番初めに，意味のわかる文章とは言い難い言葉の羅列が書き出されてしまっている．この Web ページ自体は民主党の概要や経歴などの情報が列挙されており，重要性は高いと思われる記事であったものの，重要である文をうまく見つけることができなかった．この対策として，箇条書きや表など文でない構成要素の検出やそれらの書き出し方法の検討などが考えられる．

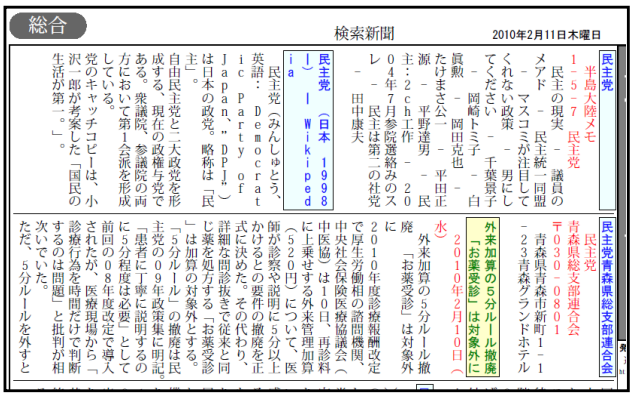

図 9　検索キーワード「民主党」の例

5. まとめ・今後の課題

ユーザが入力した検索キーワードから収集した Web 文書に対して，新聞形式による要約結果を自動的に生成し，PDF を作成して提示する検索新聞システムを提案した．重要度の推定には 3 つの特徴量を用い，ユーザが設定した重みによって重要度の判断基準を変化させた．この推定手法と TF*IDF 値を用いる手法とで比較実験を行った．その結果，Web 検索の結果表示手法の新たな可能性と本研究により作られたシステムの基盤を活用して，新たな情報検索のサービスシステムを提供できる可能性を示すことができた．

今後の課題として以下の内容があげられる．

1. **重要度の推定手法の改良**
   TF*IDF 値の手法と提案手法とで大きな差が見られなかった．ジャンルの推定や適切な重みの調整に加え，新たな尺度を含めた重要度推定手法を検討し，より精度を上げる必要がある．
2. **紙面レイアウトの検討**
   ユーザに見せやすくするにはどうしたらよいのか，記事の配置やレイアウトなどを動的に変化させる方法について検討し，新聞紙面で表示することの利点を生かす必要がある．
3. **システムのさらなる展開の検討**
   ユーザに対して，「このシステムがどういったことに使えそうか」，「将来，情報検索に必要なことは何か」といった意識調査を行い，システムの方向性を検討する．さらに，ユーザが重要と考える指標について調査し，システムの完成度を高めていく必要がある．

参考文献

[前田] 前田朗 : 専門用語自動抽出用 Perl モジュール TermExtract http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

[工藤] 工藤拓 : 形態素解析エンジン MeCab http://mecab.sourceforge.jp/

[佐藤 2007] 佐藤理史ら : 多項ナイーブベイズ分類を用いた日本語テキストの難易度判定手法の検討，言語処理学会第 13 回年次大会発表論文集 pp.534-537,2007.

[佐藤 2008(a)] 佐藤理史ら : 全教科を収録対象とした日本語教科書コーパスの構築，言語処理学会 第 14 年次大会発表論文集 D3-2, 2008.

[佐藤 2008(b)] 佐藤理史ら : 教科書コーパスを用いた日本語テキストの難易度判定手法の検討，言語処理学会 第 14 年次大会発表論文集 D5-5, 2008.

[Yahoo] Yahoo! API, http://developer.yahoo.co.jp/

[kizasi] kizasi.jp, http://kizasi.jp