Vol.2014-DBS-160 No.16
Vol.2014-OS-131 No.16
Vol.2014-EMB-35 No.16
2014/11/18

学術論文閲覧支援インタフェースのための頭字語の活用

前野 明子[1,a)]　太田 学[1,b)]　高須 淳宏[2,c)]

概要：iPad 等のマルチメィアデバイスを利用すると，電子文書から必要な情報を素早く参照することができる．そこで我々は，電子文書として学術論文を想定し，その閲覧支援を目的としたインタフェースを開発している．提案するインタフェースは，ユーザが選択した一単語に対して，その出現頻度や重要度等の解析結果，及びそのテキストに関連する Web から抽出した情報を提示する．本稿では，ユーザに提示する重要語として特に頭字語に注目し，学術論文 PDF から頭字語とその実体となる語を抽出し、その活用法を提案する．また，提案インタフェースの概要について述べる．

Use of Acronyms for Interface for Support of Browsing Scholarly Papers

1. はじめに

タブレット端末や電子書籍端末の普及に伴い，紙媒体で読んでいた文書をタブレット端末を用いて読む読書方法が普及した．この読書形態の変化に伴い，紙媒体とタブレット端末での読書方法を比較する研究が多数行なわれている．

柴田ら [1] は娯楽を目的とした読み方と，答えを探すことを目的とした業務での読み方の 2 種類の読み方から，紙媒体とタブレット端末を比較している．娯楽を目的とした読み方では，ユーザがタブレット端末に抵抗なく適応できれば，ページめくりが頻発しない限り，タブレット端末と紙媒体では認知負荷に大きな違いはないことが報告されている．しかし，答えを探すことを目的とした業務での読み方では，複数ドキュメントの移動が発生する場合，電子媒体ではページの行き来に時間を要し，紙媒体のほうが優れていることが報告されている．紙媒体とタブレット端末の比較では，Adler ら [2] も同様の報告をしている．

Eva ら [3] は視標追跡 (Eye tracking) の観点から読み方の比較を行い，電子媒体で読む事の利点として，フォントサイズの拡大を挙げている．これは，10 代から 70 代までの幅広い年齢層で閲覧比較実験を行ったことで明らかになった．特に高齢者ではフォントサイズを拡大することで画面を凝視する時間が短縮された．また Eva らは，紙媒体と読みやすさに差異はないため，電子媒体の端末開発者に対し，単にテキストを表示させるだけではなく，デザイン，また読むための機能を充実させるほうが良いと述べている．このような観点から，一概にどちらが優れているかは判断しにくく，紙媒体と電子媒体は状況に応じて使い分ける必要がある．

学術論文のように専門性の高い文書を読む場合，専門用語等の未知語に遭遇する可能性が高く，その度に専用の辞書を引く，ないしは Web サイトを検索するといった行為は効率が悪い．また，予め論文の重要なフレーズを知った上で論文を閲覧すれば，論文内容を理解するための助けとなる．そこで，マルチメディアデバイスである電子媒体の利点を生かし，論文中の重要語を予め自動抽出し，外部リソースを用いてユーザに提示できれば有用である．

本研究では，テキストの埋め込まれた学術論文 PDF ファイルを用い，一単語に対して，重要度等の論文の解析結果や Web 上の関連情報をユーザに提示するインタフェースを作成する．また，重要語の一つである省略語を含む頭字語に特に注目し，論文中に出現する頭字語とそれを意味する元の語を抽出した．更に抽出した頭字語を分類し，どのような頭字語に解説が必要かを考察した．また，その頭字語の関連項目の語を Wikipedia から抽出し，頭字語と関連項目の語がどのような関係にあるのかを示した．本インタフェースのユーザとしては初学者や，専門外の論文を読む人を想定している．また提案インタフェースはスワイプ等

[1] 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University
[2] 国立情報学研究所
National Institute of Informatics
[a)] maeno@de.cs.okayama-u.ac.jp
[b)] ohta@de.cs.okayama-u.ac.jp
[c)] takasu@nii.ac.jp

のタッチ操作を用いることで，直感的な操作が可能である．

以下に本稿の構成を示す．2 節で本研究の関連研究を紹介し，3 節で提案する学術論文閲覧インタフェースについて述べる．4 節で重要語の抽出方法について記し，5 節で評価実験について説明する．6 節で本稿をまとめ，今後の課題について述べる．

2. 関連研究

2.1 論文閲覧支援システム

阿辺川ら [4] は学術文献閲覧システムのケーススタディとして，脚注表示機能を備えた論文閲覧システムを作成した．このシステムは，連続ページめくり，拡大縮小表示，ブックマーク機能，背景色変更等の論文閲覧の機能だけでなく，論文の本文を解析し，表示されている論文の左右に Wikipedia をリソースとした補足情報を表示する機能を有している．さらに人手で論文中の段落と発表スライドの各ページを対応付け，システムで表示できるようにしている．今後は図表領域，およびタイトルやセクションなどの構成要素の認識を行い，更なる論文閲覧支援を行う予定だと述べている．

鉢木ら [5][6] は OCR テキストを用いた学術論文閲覧支援システムを開発した．電子化された論文の閲覧においてオンライン化のメリットが十分に生かされていないと考えた彼らは，Web 資源を活用して論文の閲覧支援を行うことを提案した．具体的には文書画像から低コストで作成することができる OCR テキストを用い，二つの機能を実装した．まず論文中から専門用語を抽出し，それらの語についての解説やツールなどの有用なページへのリンクを提供した [5]．さらに抽出した各専門用語で検索される論文集合とそれらに出現する重要語集合の二部グラフを作成し，HITS アルゴリズムを適用することで，関連論文をランク付けして推薦した [6]．

2.2 重要語

内山ら [7] は著者が示すキーワードが出現する年度数，研究領域，文書数について分析した．これは適切な論文を求める初学者に対し，専門用語の専門度を示す客観的な指標を作成するためである．彼らは，著者キーワードはその論文を特徴付ける語であり，一般的な語が含まれにくく，専門用語の候補として適切であると判断した．著者キーワードが，ある一定期間多く出現する場合それは流行している語であり，ピークが収まっても継続的に出現する語は特定分野において専門度の高い語とした．しかし，複数の分野において低頻度であるが継続的に出現する語においては，論文を読む上で重要であるため，頻度以外の情報も必要であるため，引用情報や文脈情報，語彙情報も加味する予定だと述べている．

鈴木ら [8] は短縮形の代表として頭字語を挙げ，括弧書き表記を伴う表現から頭字語とその元となる語を抽出する手法を提案した．例えば，Human Comuputer Interaction (HCI) 等がこれに該当する．まず頭字語に含まれる大文字数を数え，その数に応じた直前の数語を実体語候補とする．次に頭字語候補の頭文字と実体語候補の先頭 1 文字がそれぞれ等しければ出力する，といった方法で頭字語を抽出した．その過程において頭字語の規則性についても言及し，頭字語と実体語の抽出実験では F 値で約 0.8 の精度となり，研究の有用性を示した．しかし，この方法では省略語やいくつかの規則性のある頭字語等を抽出できない．本稿では我々の実装したシステムを鈴木ら [8] の手法と比較し，評価する．

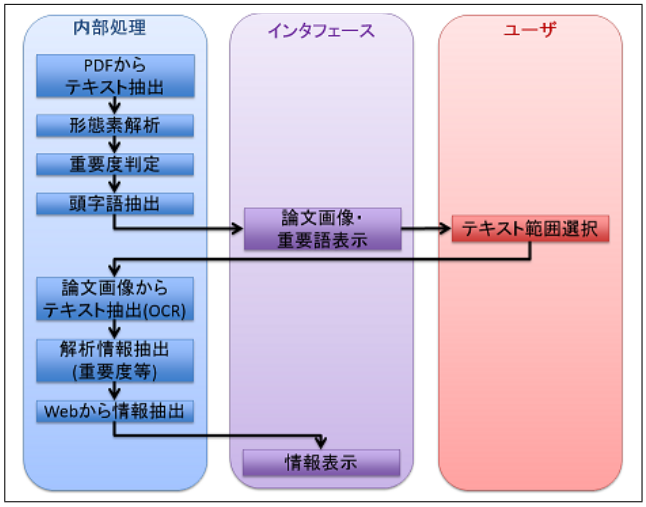

図 1 本インタフェースの動作の流れ

松尾ら [9] は閲覧対象文書だけの情報から，語の共起情報を用いてキーワードを抽出した．閲覧対象文書の頻出語を抽出し，その頻出語との共起頻度を求め，共起頻度がどの程度偏っているかでその語が重要語であるかどうかの指標とした．

湯本ら [10] は出現頻度と連接頻度に注目し，専門用語を専門分野の用語コーパスから自動抽出する方法を提案した．まず単名詞バイグラムを用いて単名詞のスコアを付け，それを拡張し，連続する単名詞のスコアの平均をとることで複合名詞のスコアも定めた．そこから専門用語候補集合における構造の情報，コーパスにおける用語候補となる統計的性質を組み合わせ，複合名詞が単独で出現する頻度も考慮し，専門用語を抽出した．

3. 学術論文閲覧支援インタフェース

3.1 概要

開発中のインタフェースは英語の学術論文を対象とした，学術論文閲覧支援のためのインタフェースである．動作の流れを**図 1** に示す．

本インタフェースは起動時，まず論文 PDF からテキストを抽出し，そのテキストを形態素解析する．次に頭字語

とその元となる語を抽出し，単語ごとのテキストの重要度を判定した後，論文画像を表示する．ユーザが任意にテキストを選択すると，選択範囲のテキストに対し，OCR で論文画像を認識し，予め判定しておいた重要度等の解析結果及び Web から収集した情報，また選択したテキストによっては関連用語を表示する．表示するテキストの解析結果は 3.3 節，Web から収集した情報については 3.4 節，重要度の判定方法については 4.3 節で詳しく述べる．

3.2 操作方法

開発中のインタフェースを**図 2** に示す．ここでは例として，NTCIR-9 PatentMT で発表された Jeff らの論文 [12] を表示している．画面には論文 PDF から抽出した論文画像が表示されており，左右にスワイプすることでページが切り替わる．また，ピンチイン，ピンチアウトを行うと論文の拡大，縮小ができる．論文中の単語一語のテキストに対して左上と右下をタップすると選択した箇所が青く網がけされ，情報提示窓が表示される．この情報提示窓の詳細を**図 3** に示す．情報提示窓は最初に出現頻度や重要度等の解析結果（赤）を表示し，スワイプを行うと Wikipedia[*1] の要約（橙），更にスワイプを行うと Weblio[*2] で表示される情報（緑），そして最後には Bing[*3] で検索された上位 3 件の結果（青）を表示する．情報提示窓はタップすると消える．また単語が頭字語の場合，情報提示窓（赤）の下部に関連用語を提示するボタンが表示され，タップすると関連用語が表示される．

図 2 で青く網がけされている部分を上にスワイプすると Wikipedia，右では Weblio，下では Bing のテキストをクエリとした検索結果のページに移動できる．これは情報提示窓に提示できる情報に限りがあるため，ユーザがテキストの抽出元となるページを閲覧できるよう配慮した．また左上に表示されている重要語と書かれているボタンを押下すると，論文中の重要度（後に述べる TF-IDF）の高い単語上位 10 件が表示される．

3.3 表示される情報

情報提示窓の解析結果（赤）の内容について説明する．解析結果は，論文内でのその語の出現頻度，その語が頭字語である場合には頭字語の元となる語，そのテキストの初出ページと周辺文章，TF-IDF に基づく重要度からなる．ここで重要度とは，その語がその文書において TF-IDF の上位から何番目であるかを表す．上位から 10 件を最も重要度の高い語とし，「★★★★★」と表示する．以降は 10 件ごとに一つの★が☆となり，50 件以降は変わらない．

また，ユーザが指定したテキストが頭字語であった場合，関連用語のボタンを表示する．これは，テキストが Wikipedia に存在する場合，Wikipedia の関連項目を抽出し，関連用語を提示する．提示する際に，ユーザが指定したテキストと関連用語の関連度を計算し，関連用語がユーザが指定したテキストの上位語であれば赤，下位語であれば緑で表示する．関連用語の関連度の判定は 4.4 節で詳しく述べる．

```
1: BT
2: \F13 12 Tf
3: 288 720 Td
4: (ABC) Tj
5: ET
```

図 4 フォントに着目した PDF のコード例

3.4 関連ページ検索

本インタフェースでは，Web 上の有力な情報源である，Wikipedia，Weblio，Bing を用いて，ユーザが選択したテキストの関連情報を抽出する．

まず MediaWiki API を用いて Wikipedia の情報を取得し，要約部分を表示する．Weblio からは Web ページの html タグを削除し，プレーンテキストのみを表示する．Bing については，Bing API を用いて検索結果の上位 3 件を取得し，ページタイトル，URL，スニペットを表示する．

4. 重要語抽出

論文 PDF ファイルからのテキストの抽出方法，そのテキストからの頭字語及びその元となる語の抽出方法と重要度評価，また頭字語とその関連語の関連度の算出方法について述べる．

4.1 学術論文 PDF からのテキスト抽出

論文 PDF からテキストを抽出する際，PDF の内部構造を知る必要があるため，PDF の内部構造，特にフォントについて述べる．PDF の内部構造は，Acrobat[*4] で容易に確認することができる．

PDF ファイルは先頭部分，本体部分，末尾部分に分かれる．先頭部分には PDF のバージョン等があり，末尾にはファイルの本体部分に並ぶオブジェクトをどれでも直接呼び出せる，相互参照のコマンドが書かれている．本体部分には複数のオブジェクトが定義されている．本研究で必要となるテキストのオブジェクトは本体部分に定義されている．テキストオブジェクトの PDF コードの例を**図 4** に示す．

図 4 の 1 行目はテキストの始まりを表す．2 行目の Tf はフォントを指定するオペレータを示す．ここでオペレータとは処理方法を表す記号である．よってこれは，ページ内

[*1] http://en.wikipedia.org/wiki/
[*2] http://ejje.weblio.jp/
[*3] http://www.bing.com/
[*4] http://www.adobe.com/jp/products/acrobat.html

Vol.2014-DBS-160 No.16
Vol.2014-OS-131 No.16
Vol.2014-EMB-35 No.16
2014/11/18

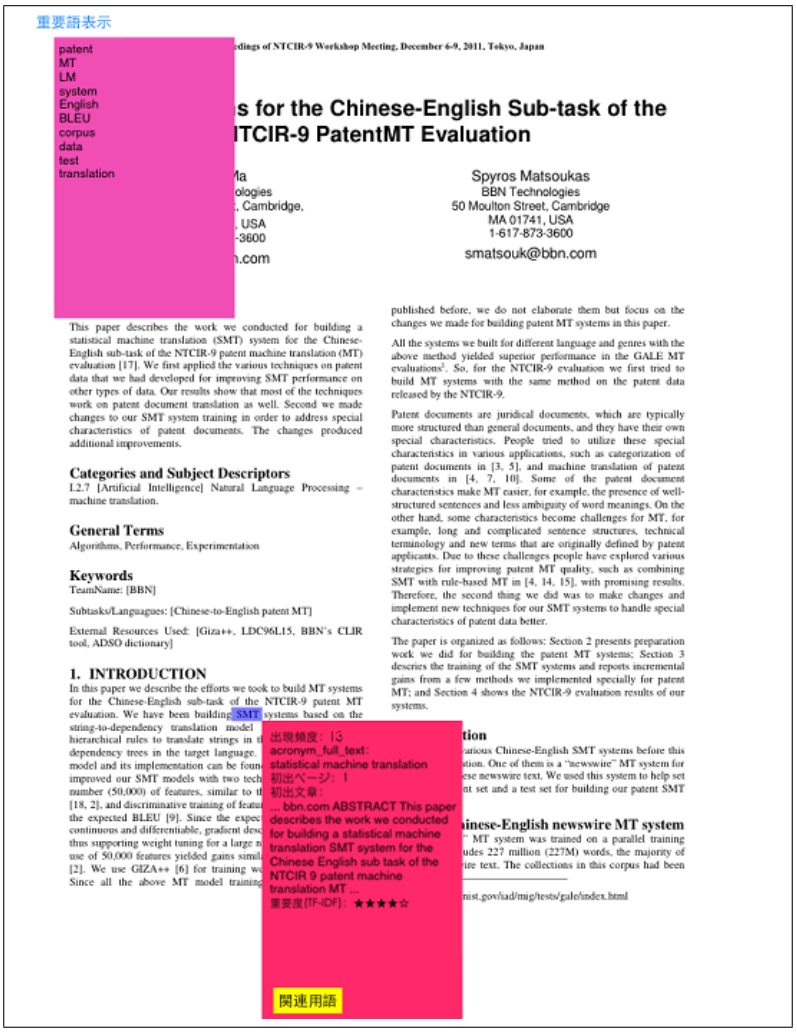

図 2　学術論文閲覧支援インタフェース

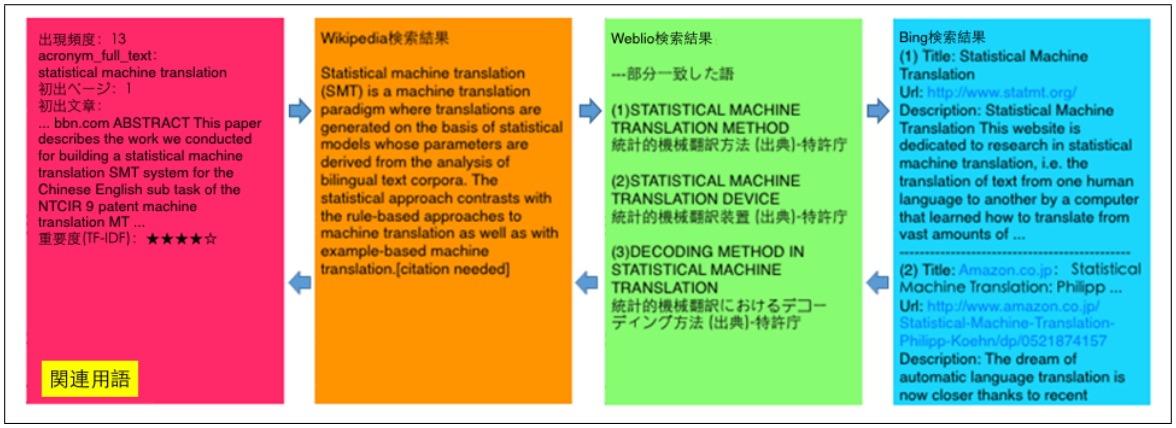

図 3　情報提示窓

にの情報が存在する \Resources で定義された F13 という名のフォントを用い，12pt の大きさで表示することを示す．3 行目の Td はフォントの開始位置のオペレータを示す．つまり，左から 4cm(4×72=288)，下から 10cm(10×72=720) が開始位置になることを示す．4 行目の Tj はテキストのオペレータを示す．つまり，テキスト文字列「ABC」が内包されていることを示す．5 行目はオブジェクトの終わりを示している．

本インタフェースでは，まず論文 PDF からテキストデータを示す Tj をトリガとして登録しておいたコールバック関数を呼び出し，テキストデータを抽出する．図 4 の例では，「ABC」をテキストとして抽出する．

提案インタフェースでは，テキストオペレータの Tj 内に存在するテキスト文字列をテキストとして用いる．

4.2 頭字語抽出

本研究で言う頭字語とは，省略語の一種であり，複数の単語からなる語の頭文字を繋げて作られた語を指す．例えば，Human Computer Interaction は一般に HCI と表記され，この HCI を頭字語と呼ぶ [*5]．多くの頭字語がこれに当てはまるが，前置詞のため頭文字が頭字語の一文字として使用されない場合や，特定の単語において単語を数字に置き換える場合等の例外が存在する．本稿で抽出する頭字語の種類は 4.2.2 項において詳しく定義し，頭字語とその頭字語の元となる語を抽出する．

4.2.1 抽出対象

英文において頭字語を記述する際，頭字語の後に頭字語の元となる語が括弧書きとして書かれる場合，もしくはその逆がある．その例を以下に示す．以降では鈴木ら [8] の記述を参考に，頭字語の元となる語を実体語と記す．

(I) 頭字語（実体語）の場合
　例：HCI (Human Computer Interaction)
(Ⅱ) 実体語（頭字語）の場合
　例：Human Computer Interaction (HCI)

本研究ではこの 2 種類を抽出対象とし，頭字語とその実体語を抽出する．

4.2.2 抽出する頭字語の種類

以下に本研究で扱う頭字語の種類を示す．

(1) 単語の頭文字を繋げて作られた頭字語
　例：HCI (Human Computer Interaction)
(2) 実体語中に前置詞等の小文字の単語が存在する頭字語
　例：ADC (Analog to Digital Converter)
(3) 実体語が ex-で始まる頭字語
　例：XOR (Exclusive OR)
(4) 記号による単語分割を伴う頭字語
　例：CLIR (Cross-Lingual Information Retrieval)
(5) 複数形の s を含む頭字語
　例：ISPs (Internet Service Providers)
(6) 実体語の 1 単語中に頭字語に使用する文字を複数含む頭字語
　例：RaDAR (Radio Detection And Rangin)
(7) 頭字語と単語を組み合わせた頭字語
　例：XSLT (XSL Transformations)
(8) 頭字語の文字数と実体語の単語数が一致しない頭字語
　例：HTK (Hidden Markov Model Toolkit)
(9) 数字が共通の文字数を表す頭字語
　例：W3C (World Wide Web Consortium)
(10) 前置詞を数字に置き換えた頭字語
　例：P2P (Peer to Peer)

[*5] 本稿では，いわゆる acronym と initial word を併せて頭字語と呼ぶ．

なお，鈴木ら [8] が対象とした頭字語は，上記の (1)〜(5)，(7)，(9) である．

4.2.3 抽出手法

4.2.1 項で説明した頭字語及びその頭字語に該当する実体語を抽出する．

「頭字語 （実体語）」と表記された頭字語及び実体語を抽出する場合，括弧内に必ず実体語が含まれている．一方，「実体語（頭字語）」と表記された頭字語及び実体語を抽出する場合，実体語の範囲を特定する事が難しい場合がある．このような違いはあるが，頭字語と実体語の頭文字との比較方法は基本的に同じである．以下にそれぞれの抽出方法を示す．

共通の前処理
(i) “ (” と “) ” の間の単語列を抽出する．
(ii) 括弧内に単語が複数ある場合は (I) の実体語候補，一単語のみの場合は (Ⅱ) の頭字語候補とする．
(iii) (I)，(Ⅱ) に当てはまる頭字語かどうか判定する．

(I) 頭字語（実体語）の場合
(i) “ (” の一つ前の単語（頭字語候補）の最後の文字と，抽出した括弧内の単語列（実体語候補）の最後の単語の頭文字を比較する．
(ii) (i) で比較した文字が一致すれば，頭字語候補のその前の文字と実体語候補のその前の単語の頭文字を比較する．
(iii) 頭字語候補の文字数分，頭字語候補の文字と実体語候補の頭文字が一致すれば，頭字語とその実体語として抽出する．

(Ⅱ) 実体語（頭字語）の場合
(i) “ (” の一つ前の単語（実体語候補）の最後の単語の頭文字と，抽出した括弧内の単語（頭字語候補）の最後の文字を比較する．
(ii) (i) で比較した文字が一致すれば，頭字語候補のその前の文字と実体語候補のその前の単語の頭文字を比較する．
(iii) 頭字語候補の文字数分，頭字語候補の文字と実体語候補の頭文字が一致すれば，頭字語とその実体語として抽出する．

(I)，(Ⅱ) のいずれの方法でも，4.2.2 項で述べた (1) と (4) 以外を抽出することができないため，さらに以下の例外処理を行う．なお，(4) のように，実体語中に「-」等で単語分割を伴う場合，形態素解析の際記号で繋がれた一つの単語が別々の単語に分割されるので，(I)，(Ⅱ) の方法で頭字語とその実体語を抽出できる．

(2) のように，実体語候補の中に前置詞が含まれる場合，実体語候補の前の単語も比較する．

(3) のように，頭字語に X が含まれる場合，ex を予約語

として扱い，頭字語候補の文字と実体語候補の頭文字を比較する際，ex ではなく x として比較する．

(5) のように，「大文字＋ s」からなる頭字語は，大文字の部分と複数形を表す s を分け，大文字で表される部分のみを比較する．

(6)，(7)，(8) のように，実体語候補である単語の頭文字と頭字語候補の文字が一致しない場合は，実体語候補の頭文字の次の文字も比較する．これを，該当の文字が見つかるまで頭字語候補の文字と実体語候補のその前の文字を比較する．

数字が頭字語候補に含まれる場合，まず数字が共通の頭文字の数を表すのか，前置詞を数字で表す語であるのかを判定し，その結果に従って (9) 及び (10) の種類の頭字語として扱う．

(9) の場合，予め頭字語の頭文字の種類ごとの数を数えておき，数字の分だけその頭文字に置き換えて比較する．

(10) の場合，to と 2，for と 4 の置き換えを予約語として登録しておき，数字を単語に置き換えて比較する．

(9)，(10) のいずれでもない場合は，通常の数字を比較する．

ただし，以下は頭字語に含めない．

(I) 頭字語（実体語）の場合
 (a) 括弧内が一単語で表されるもの

(Ⅱ) 実体語（頭字語）の場合
 (a) 括弧内が全て小文字
 (b) 括弧内が全て数字
 (c) 括弧内が一文字

(I)-(a) や (Ⅱ)-(a) では，括弧内の単語は副詞である場合や，値の説明等が書かれている場合がある．(Ⅱ)-(b) では，参考文献文字列等の年号が抽出される場合がある．(Ⅱ)-(c) では，頭字語として成立するものもあったが，それほど重要ではないと判断した．例として，Yes (Y) 等が挙げられる．よって，これらを頭字語として抽出しても，頭字語には含めない．

4.3 重要度の判定

我々は先行研究で頻出語に注目し，それらに対して有用なリンクを生成することを提案した [13]．しかし，単に頻出と言うだけでは不十分であったため，4.1 節の方法で抽出したテキスト情報を形態素解析し，TF-IDF により単語ごとの重要度を算出する．なお，形態素解析には Objective-C の NSLinguisticTagger class を用いた．重要語の候補とした単語の品詞は，名詞，形容詞，動詞，未知語の 4 種類である．ただし，数字だけからなるもの，記号，英字 1 文字で形成される単語は除外した．単語 t_i の TF-IDF 値の算出式を以下に示す．

$$tfidf_i = tf_i * \log(\frac{num}{df_i}) \tag{1}$$

ここで，tf_i は論文文書中における単語 t_i の出現頻度，$num = 16,831,499$ であり，これは CiNii における論文の総収録件数（2014 年 6 月 17 日時点），df_i は CiNii において t_i を検索した時の検索結果数（論文数）を表す．

4.4 関連用語の判定

佐藤ら [11] はサーチエンジンのヒット数に着目し，関連用語の抽出を行った．用語 t，x の関連性を，以下の式で定めた．

$$H(t) = \text{用語 } t \text{ が現れるページ数} \tag{2}$$

$$H(t \wedge x) = \text{用語 } t, x \text{ が共に現れるページ数} \tag{3}$$

$$a(x \to t) = \frac{H(t \wedge x)}{H(x)} \tag{4}$$

$$a(t \to x) = \frac{H(t \wedge x)}{H(t)} \tag{5}$$

ここで，$a(x \to t)$ は「x が現れるページにどれくらい t も出現したか」を表し，$\mathrm{a}(t \to x)$ は「t が現れるページにどれくらい x も出現したか」である．これら 2 つの指標を比較すると以下のような傾向が見られる．

(1) $a(x \to t) \ll \mathrm{a}(t \to x)$: x は t の上位語である場合が多い
(2) $a(x \to t) \gg \mathrm{a}(t \to x)$: x は t の下位語である場合が多い

本稿ではこの関係を用い，頭字語と Wikipedia にあるその頭字語の関連語についての上位語，下位語を決定する．

5. 頭字語抽出実験

5.1 実験概要

予め人手で頭字語の正解データを作成し，正しく頭字語が抽出できるかどうかを再現率と適合率，F 値で評価した．実験で使用した論文は，情報検索等の評価型ワークショップ NTCIR-10 の論文である．全 104 件の内，正しく PDF からテキストを抽出できた 31 件を実験に用いた．一論文あたりの平均単語数は 3,010.2 語 (93,315/31)，平均ページ数は 4.7 ページ (145/31)，頭字語の種類は 109 種類であった．また各評価指標の算出式を以下に示す．

$$\text{再現率 } (R) = \frac{\text{抽出された頭字語の正解数}}{\text{論文中の頭字語の数}} \tag{6}$$

$$\text{適合率 } (P) = \frac{\text{抽出された頭字語の正解数}}{\text{抽出された頭字語の数}} \tag{7}$$

$$\text{F 値} = \frac{2R \cdot P}{R+P} \tag{8}$$

Vol.2014-DBS-160 No.16
Vol.2014-OS-131 No.16
Vol.2014-EMB-35 No.16
2014/11/18

表 1 頭字語の抽出実験結果

再現率	0.773 (82/106)
適合率	0.953 (82/86)
F 値	0.854

5.2 実験結果

頭字語の抽出実験結果を**表 1**，抽出に成功した頭字語の例を**表 2**，失敗例を**表 3** に示す．なお，表 2，表 3 の分類とは 4.2.2 項で説明した (1)〜(10) を指す．

表 1 の結果を見ると，適合率が高く，抽出したほとんどの場合は頭字語とそれに対応する正しい実体語であった．しかし再現率が 0.773 であり，24 の頭字語が正しく抽出できなかった．

表 2 の抽出成功例では，多くの頭字語が 4.2.2 項で定義したパターンにあてはまり，特に，(1)，(2)，(4)，(6) のものが多かった．また，実験に用いた論文の中に (3)，(5)，(9)，(10) のパターンに該当する頭字語は出現しなかった．

表 3 の抽出失敗例の中には PDF からのテキスト抽出が不完全で，文字として抽出できなかったものが四つあり，これが抽出漏れの原因の一つであった．例として，*r*ough *s*et *b*ased *m*odel (RSBM) が挙げられる．これは実体語の頭文字がイタリック体で表記されており，フォントの異なる r，s，b，m の部分を抽出できていなかった．

表 3 の他の失敗例を見ると，頭字語の文字をそのままではなく，置き換えて使用する例が存在した．例えば実体中の at を@，and を&等，記号で代替するものも存在した．これらに対応するためには特定の語を予約語として登録しておき，該当する語を逐一比較する必要がある．

参考文献欄の記述で多く見られたのは学会名等を省略する場合に，実体語中には存在しないが，頭字語には年数が追加されるものである．例として，European Association for Machine Translation (EAMT-05) が挙げられる．この場合，学会名とは別に年数を抽出するなどして，比較する必要がある．

頭文字の並びと抽出する実体語の順番が異なるものも存在した．例として，simplified Chinese (CS) が挙げられる．これに対処するには，順番を入れ替える等して抽出を試みる必要がある．

また，頭字語ではない略語を抽出しようとし，失敗している例があった．例として，1CLICK (One Click Access) が挙げられ，これも適合率を下げる一因となった．

データセットが異なるため直接比較はできないが，鈴木ら [8] の同様の実験においては，再現率 0.728，適合率 0.901，F 値 0.805 と言う結果であった．鈴木ら [8] はまた実体語 (頭字語) のパターンで抽出に失敗した例として，16 例を挙げている．その内，9 例は本手法で抽出可能である．

この 9 例の内 6 例は 4.2.2 項で定義した (6)，1 例は (8) の例である．(6) の例として，BMI (Broadcast Music) や input method library (IMLIB)，(8) の例として，Universal Coded Character Set (UCS) がある．ただし，BMI の正式名称は BMI (Broadcast Music Inc.) であると考えられる．

残りの 2 例は，頭字語の文字と実体語の頭文字の比較順序によって，抽出できたりできなかったりするものである．鈴木ら [8] は，頭字語を抽出する際，頭字語の文字と実体の頭文字の比較を左から行う．つまり，Human Computer Interaction (HCI) と言う語を抽出する場合，まず Human の H と，HCI の H を比較する．一方，本手法は右から行う．つまり，Interaction の I と HCI の I を比較する．XML Path Language (XPath) を例にとると，左から比較を行うと Language を抽出することができない．このような比較順序の違いも，より多くの頭字語を抽出できる一因となった．

提案手法でも鈴木らの方法でも抽出できない例としては，先ほど述べた頭文字の並びと抽出する実体語の順番が異なるものや，実体語中に合致すべき文字がないものである．これらに対処する場合，例外的な扱いが必要となる．

5.3 抽出された頭字語の考察

抽出された頭字語の主な種類を以下に示す．

(1) 専門用語
(2) 学会名
(3) 機関，大学名
(4) 手法等論文内で定義された固有名詞
(5) ソフトウェア・辞書等の既存のリソース名
(6) 研究分野名
(7) 長い一般名詞

この中で初学者が解説を求める用語は (1) や (4) のような語である．(1) や (5)，(6) には Wikipedia や Weblio 等の外部リソースから解説が得られれば，一定の有効性がある．(2)，(3) には公式ホームページがあればその URL にリンクを生成するのも有用である．これらは本インタフェースで対応しており，一語であればこれらの情報を提示できる．(4) では論文内の情報を使用し，解説された文章を提示する必要がある．多くの場合最初出の箇所で定義され，解説されているので，今後このような語にも対応したい．

6. まとめ

本稿では，学術論文閲覧支援インタフェースのための頭字語活用方法について述べた．実験では，論文からの頭字語抽出性能を再現率と適合率，F 値で評価した．その結果，再現率 0.773，適合率 0.953，F 値 0.854 となった．頭字語抽出に失敗した例にも規則性が見られるため，それに対応すれば精度向上が見込まれる．

今後の課題としては，頭字語抽出の精度向上の他に，他の重要語の検討やインタフェースの改良等が挙げられる．重

表 2 頭字語の抽出成功例

パターン	分類	論文中の表記	抽出結果
頭字語（実体語）	(1)	CTTD (Chinese Technical Term Dictionary)	Chinese Technical Term Dictionary，CTTD
	(1)	SCS (standing for Stanford Chinese Segmenter)	Stanford Chinese Segmenter，SCS
	(6)	RITE (Recognizing Inference in TExt)	Recognizing Inference in TExt，RITE
実体語（頭字語）	(1)	Maximal Marginal Relevance (MMR)	Maximal Marginal Relevance，MMR
	(2)	Corpus of Spontaneous Japanese (CSJ)	Corpus of Spontaneous Japanese，CSJ
	(4)	Rule-Based Machine Translation (RBMT)	Rule - Based Machine Translation ，RBMT
	(7)	Extra Binary Class (ExtraBC)	Extra Binary Class，ExtraBC
	(8)	Hidden Markov Model Toolkit (HTK)	Hidden Markov Model Toolkit，HTK

表 3 頭字語の抽出失敗例

パターン	分類	論文中の表記	抽出結果
実体語（頭字語）	(1)	*r*ough *s*et *b*ased *m*odel (RSBM)	-
	(8)	Term Frequency Distribution Feature (TF)	distribution feature，TF
	-	simplified Chinese (CS)	Chinese，CS
	-	the Ministry of Education (MOEDICT)	-
	-	European Association for Machine Translation (EAMT-05)	-
	-	the First Recognizing Textual Entailment Challenge (RTE-1)	-
	-	Precision-at-N (P@N)	-

要語については，学術論文閲覧支援のために重要語としてどのような語が重要か，更に検討を進める．インタフェースの特に表示方法については，単にユーザにテキストで情報を提示するだけでなく，どのようにユーザに提示すれば論文の閲覧支援となるかさらに検討したい．

謝辞

本研究の一部は，科学研究費補助金基盤研究 (B)(課題番号 23300040, 24300097)，科学研究費補助金基盤研究 (C)(課題番号 25330384)，および国立情報学研究所公募型共同研究の援助による．ここに記して深謝する．

参考文献

[1] 柴田 博仁，高野 健太郎，大村 賢悟，“ 電子書籍端末は紙を代替できるか? 電子書籍端末の評価実験にもとづく考察 ”，富士ゼロックス テクニカルレポート，No.21，2012.
[2] Annette Adler, Anuj Gujar, Beverly L. Harrison, Kenton O’Hara，Abigail Sellen, “ A diary study of work-related reading: Design implications for digital reading devices ”, In Proc. CHI ’98， pp. 241-248，1998.
[3] Eva Siegenthaler，Pascal Wurtz，Rudolf Groner, “ Improving the Usability of E-Book Readers ”，Journal of Usability Studies，Vol. 6，Issue 1，pp. 25-38，2010.
[4] 阿辺川武，相澤彰子, “ 内部構造解析機能と脚注表示機能を備えた論文閲覧システム ”，人工知能学会 インタラクティブ，情報アクセスと可視化マイニング第 7 回研究会，pp. 13-18，2014.
[5] 鉢木稔浩，太田学，高須淳宏, “ Web 資源を利用した学術論文閲覧支援システム ”，情報処理学会研究報告，Vol. 2009-DBS-149，No. 14，pp. 1-6，2009.
[6] 鉢木稔浩，太田学，高須淳宏, “ 学術論文閲覧支援システムのための関連論文推 ”，第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM 2011)，F9-4，2011.
[7] 内山清子, “ 論文の論理構造における分基礎用語に関する分析 ”，第 2 回コーパス日本語学ワークショップ，pp. 195-199，2012.
[8] 鈴木 伸哉，劉 連文，桝井 文人 , 河合 敦夫，椎野 努, “ 固有名抽出のための短縮形知識の探索手法 ”，自然言語処理研究会報告 2000(107)， pp. 91-96，2000.
[9] 松尾豊，石塚満, “ 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム ”，人工知能学会論文誌，vol. 17，No. 3，pp. 217-223，2002.
[10] 湯本紘彰，森辰則，中川裕志, “ 出現頻度と連接頻度に基づく専門用語抽出 ”，情報処理学会研究報告 自然言語処理，vol. 10，No. 1，pp. 27-45，2003.
[11] 佐藤 理史 , 佐々木 靖弘 , “ ウェブを利用した関連用語の自動収集 ”，情報処理学会研究報告 自然言語処理, NL-153，pp. 57-64，2003.
[12] Jeff Ma，Spyros Matshoukas, “ BBN ’ s Systems for the Chinese-English Sub-task of the NTCIR-9 PatentMT Evaluation ”，NTCIR-9 Workshop Meeting，2011.
[13] 前野明子，太田学，高須淳宏, “ 学術論文閲覧支援インタフェースの試作 ”，第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM 2014)，E3-3，2014.