Vol.2009-DBS-149 No.14
2009/11/21

Web 資源を利用した学術論文閲覧支援システム

鉢木 稔浩[†1] 太田 学[†1] 高須 淳宏[†2]

本稿では電子図書館の学術論文文書画像と Web を連係させることで，学術論文のオンラインでの閲覧を支援するシステムを提案する．特に学術論文には多くの専門用語が出現するため，本研究では学術論文の OCR テキストを用いて，これらの専門用語の解説ページ等を Web 検索を利用して提供する．具体的には論文中の専門用語を抽出し，解説等の有用なページを検索し，それらへのリンクを自動生成する．実験では専門用語の抽出精度，リンク先ページの適切性を評価した．

A Browsing Support System of Research Papers Using Web Resources

Toshihiro HACHIKI,[†1] Manabu OHTA[†1]
and Atsuhiro TAKASU[†2]

This paper proposes an online-browsing support system of research papers by linking document images in digital libraries with the related Web resources. Using the OCRed text of research papers, we search the Web for explanatory pages of the technical terms which we encounter while reading research papers. To be concrete, the system extracts technical terms, searches for useful pages such as those explaining the terms, and automatically generates links to the pages. We evaluated the accuracy of the technical term extraction and how appropriate the generated links were by the experiments.

†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

1. はじめに

現在の電子図書館の多くは Web 経由でアクセス可能なものが多いが，その文書の閲覧においてオンラインであるメリットが十分に生かされているとはいい難い．例えば，検索などのサービスが一つの電子図書館内で閉じていることが多く，検索できるデータベースが限定される．そこで我々は，電子図書館と Web を連係させることでより充実したサービスが提供できると考えた．

一方学術論文を読む場合，大学生のように必ずしもその分野に精通していない読者は，専門用語になじみがないことは珍しくない．しかし，そのたびに辞書や Web で調べながら読み進めていくのは効率が悪い．そういった語句については，解説や関連するページへのリンクがあれば論文の内容理解が容易になるが，これは Web 資源を利用した学術論文の閲覧支援といえる．

そこで本研究では学術論文の OCR テキストを利用し，専門用語に対して Web 上の有用な情報源への適切なリンクを提供する．ここで OCR テキストとは，文書画像を OCR 処理して得られる認識誤りを含む生のテキストのことで，文書画像があれば比較的低コストで作成できる．文書画像を提供する電子図書館では，この OCR テキストの有効利用が非常に重要な課題である．本研究では，この論文 OCR テキストから専門用語を抽出し，解説等の有用なページを検索する．さらにそれらのページへのリンクを自動生成し，論文閲覧時に表示することでオンラインでの閲覧を支援する．

本稿は 2 節で関連研究，3 節で提案する論文閲覧支援システムについて述べる．さらに 4 節で実験と評価について説明し，5 節でまとめと今後の課題を述べる．

2. 関連研究

論文の表題や著者名といった書誌要素を，学術論文文書画像のタイトルページから自動で獲得する研究が行われている[1)–3)]．これは論文検索では通常書誌要素を属性として指定するため，この書誌要素が特に重要となるからである．例えば薬師らは CRF[4)] を用いて，さまざまな学術論文誌におけるタイトルページから重要な書誌情報を高精度で自動抽出する手法を提案した[3)]．本研究では，薬師らが書誌要素を自動抽出した OCR テキストを利用して実験を行った．

Web 上にある一般的な文章から自動的に専門用語を抽出する研究は，今まで多くの研究が行われている．例えば久光らは語の話題性や分野特定性を求める指標を提案した[5)]．これ

はある語と文章内共起する語の集合の偏りから，語の重要度を測るものである．また湯本らは出現頻度と連接頻度を用いた専門用語抽出を行った[6]．単語に連接する語，つまり単語バイグラムの出現頻度からその単語のスコア付けをする．複合語の場合，構成する単語のスコアの平均をとる．これに単語または複合語自身の出現頻度も考慮して抽出する方法を提案した．本研究においても専門用語を抽出するが，出現頻度情報のみを用いたスコア付けを行っている．

キーワードに関連するページを自動抽出する研究もある．中谷らはリンク元ページに関連するページへのリンクを，キーワードとページ間の類似度を用いて自動生成した[7),8]．これはリンク元ページの文章を，ある程度の意味のまとまりをもつ部分に分割し，その部分ページとリンク先ページとの類似度を求める．最も関連したリンク元の部分ページに，リンク先ページへのリンクを構築する．この手法は分割した部分ページから関連ページへリンクをはるが，本研究ではキーワード，つまり専門用語からその関連ページへのリンクを生成する点が異なる．

用語説明が記述されている Web ページを検索し，該当箇所を抽出する研究では，藤井らが「CD-ROM 世界大百科事典」の用語説明文から説明表現を半自動的に収集した[9]．その中で頻繁に共起する文節に基づいて，用語説明抽出のための 18 種類のテンプレートを作成した．さらに HTML タグを手掛かりとして，テンプレートにマッチした文が出現する段落や見出し語，リンク先に続く一定の範囲内の複数文を用語の説明文として抽出した．また土橋らは，用語説明文のテンプレートの自動生成を行い，それらを用いて用語説明文の抽出を行った[10]．土橋らはまず，用語辞典「imidas2004」に記述されている説明文と意味内容が一致する文を Web 上から収集した．さらに，その中で説明文によく用いられる文中表現 6 種類と文末表現 16 種類を獲得し，それらを組み合わせて計 96 種類のテンプレートを作成した．テンプレートにより抽出された文と直後の文の関連度を求めて，閾値以上なら両方の文を，閾値以下なら前の文のみを説明文として，説明文の集合を抽出した．本研究ではこれらの説明文で使われる表現を参考にテンプレートを作成した．

3. 提案する学術論文閲覧支援システム

3.1 概　　要

提案システムでは学術論文中に出現する専門用語のように，解説等の付加的な情報がある方が望ましい語句（以下用語）に対して，用語解説等の有用ページの検索を行う．有用なページとして本研究では，解説ページとツールページを提案する．

提案システムでは閲覧支援のため以下の処理を行う．

- 用語抽出
  OCR テキストから用語の候補となる特徴語の集合を抽出する．これら特徴語の重要度を算出し，重要度が大きいものを用語として抽出する．
- 解説ページ検索
  用語について解説しているページを Wikipedia[11] と Web の両方で検索し，リンクを生成する．
- ツールページ検索
  用語に関するツールページを検索し，リンクを生成する．ここでツールページとは用語の技術や概念を利用したツールを公開したり，紹介したりしているページとする．

3.2 用 語 抽 出

学術論文 OCR テキストを，形態素解析器 Sen[12] を用いて形態素解析し，以下のルールに従い用語の候補となる特徴語を抽出する．

(1) 品詞情報が「名詞」のものおよびカタカナ，漢字，英数字のそれぞれのみで構成される「未知語」を特徴語として抽出する．
(2) (1) で抽出した語が連続する場合は連結して一つの特徴語とする．
(3) 不要語を除去する．不要語はひらがなや数字のみで構成される語，1 文字の語，除外語リストの語とした．この除外語リストは「あらまし」のように論文の構成上必ず出現する語や,「2 種類」のように用語として不適当と考えられる語を含む．

一方 OCR テキストには，文字の認識誤りが含まれる．そこで特徴語に含まれる認識誤りを修正するために，Yahoo!ウェブ検索[13] を利用する．Yahoo!ウェブ検索では綴りを誤った語で検索すると,「…ではありませんか？」のようにシステムが正しいと推測した語を提示してくれるサービスがある．これを利用し，検索質問の語を提示された語に訂正する．例えば「スケーラピリティ」が特徴語として抽出される．この語を検索質問として Yahoo!で検索をすると「スケーラビリティではありませんか？」と正しい語が提示されるので,「スケーラピリティ」を「スケーラビリティ」に変換する．しかし正しい特徴語で検索しても，別の語が提示されることがある．そこで誤りを含む特徴語の場合，誤りがない特徴語に比べ検索結果件数がかなり少ないという事実があるので，検索結果の総数が 1000 件を超えないときにのみ，この訂正を行う．

次に抽出した特徴語に TF・IDF 法により，重み付けを行う．本研究では特徴語 t_i の重要度 $tfidf_i$ を以下のように定義する．

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

ここで特徴語 t_i が抽出された論文文書中におけるその t_i の出現頻度が tf_i，全文書集合における出現文書数が df_i，全文書数が num である．この重要度を利用して，特徴語 t_i のスコアを以下の式で定義する．

$$score_i = \frac{tfidf_i}{\sum_k tfidf_k} \quad (2)$$

論文から抽出したすべての特徴語をこのスコアに基づいてランク付けする．ランキング上位の特徴語から順にスコアを累積スコアに加えていき，初めて累積スコアが閾値を超えたとき，そのときの特徴語集合を用語として抽出する．

3.3 解説ページ検索

本節では抽出した用語について解説，説明を行っているページの検索について説明する．解説ページへのリンクと共に，提案システムでは用語説明の一部をサーチエンジンの検索結果のスニペットのように提示するため，検索した解説ページから用語説明文も抽出する．また検索対象は Wikipedia および Web で，まず Wikipedia から検索する．

3.3.1 Wikipedia 検索

Wikipedia 検索では，Wikipedia に存在するそれぞれの用語に関する記事を検索する．提案システムでは Wikipedia API[14] を使用する．まず用語を検索質問として Wikipedia を検索し，その検索結果を取得する．Wikipedia API は検索質問と見出し語の部分一致検索を行うため，複数の検索結果を返す場合がある．部分一致で検索される記事には用語の解説として不適切なものが多いため，見出し語と検索質問が完全一致する記事のみを取得する．また，取得する記事のサマリを用語の説明文として抽出する．

用語について書かれた Wikipedia の記事が存在する場合，記事中で最も TF が高い語をその用語の共起語として取得する．この語は用語との関連が深いと考えられるため，後述の Web 検索で検索質問に追加して，解説ページ検索の精度向上を図る．共起語の抽出手順は 3.2 節に示した特徴語の抽出手順と同様であるが，さらに以下のものを除外語リストに追加する．

- 記事の見出し語に文字列として含まれる語
- 記事の見出し語を文字列として含む語
- Wikipedia に頻出する語（例： 編集，リンク，ウィキペディア）

3.3.2 Web 検索

Wikipedia の記事では内容が不十分であったり，用語によっては記事自体が存在しない場合がある．そこで本研究では Yahoo!API[15] を用いて，Web 上で用語を解説しているページを検索する．

まず予め解説ページらしさを判断するために，説明文でよく使われる説明表現を集めたテンプレートを用意する．用語説明文は特徴的な表現で記述されていることが多い．本研究では藤井らが作成したテンプレートを参考に 23 種類の説明表現を作成した．例えば，用語 X に対し「X とは―である」や「X を―と定義」などの表現を含む．このテンプレートを用いて以下の処理手順により，用語の解説ページを検索する．

(1) 用語に関して記述された Web ページを，Yahoo!API を利用して 30 件取得する．このとき用語 X の解説ページを検索するために「X とは」という検索フレーズを用いる．またこの用語に Wikipedia の記事が存在しているときには，用語と 3.3.1 節で説明したその共起語の AND 検索により，ページ群を取得する．

(2) 得られた Web ページ群から解説ページを選別する．具体的には取得 Web ページの HTML タグを取り除き，前述の説明表現テンプレートとマッチする文を探し，それに続く二文とともに計三文を説明箇所として抽出する．三文抽出するのは Web 上に存在する用語説明文は，通常複数の文で構成されているからである[10]．

このようにして Web 検索では解説ページを最大三つ取得する．

3.4 ツールページ検索

ここでは用語のツールページの検索について説明する．3.3 節で検索した用語の解説ページに加えて，ツールページを提供することで，用語に対する理解を深めるだけでなく，その専門的な技術がより利用しやすくなる．そこで本研究では，用語の技術や概念を利用したツールを紹介しているページを Web から取得する．

この処理は以下の手順で行う．

(1) Yahoo!API を利用して，ツールページの候補となる Web ページを 30 件取得する．その検索質問は用語を X とすると「X AND (ツール OR tool OR ソフト OR software)」とする．

(2) (1) で取得した Web ページから，ダウンロードまたは download という文字列がアンカータグ ⟨a⟩⟨/a⟩ に囲まれているページを取得する．さらにアンカータグの alt 属性にこれらの文字列が含まれる場合もページを取得する．取得対象となるツールページには，ツール名やツールの説明文などと共にアンカーテキストの「ダウンロード」

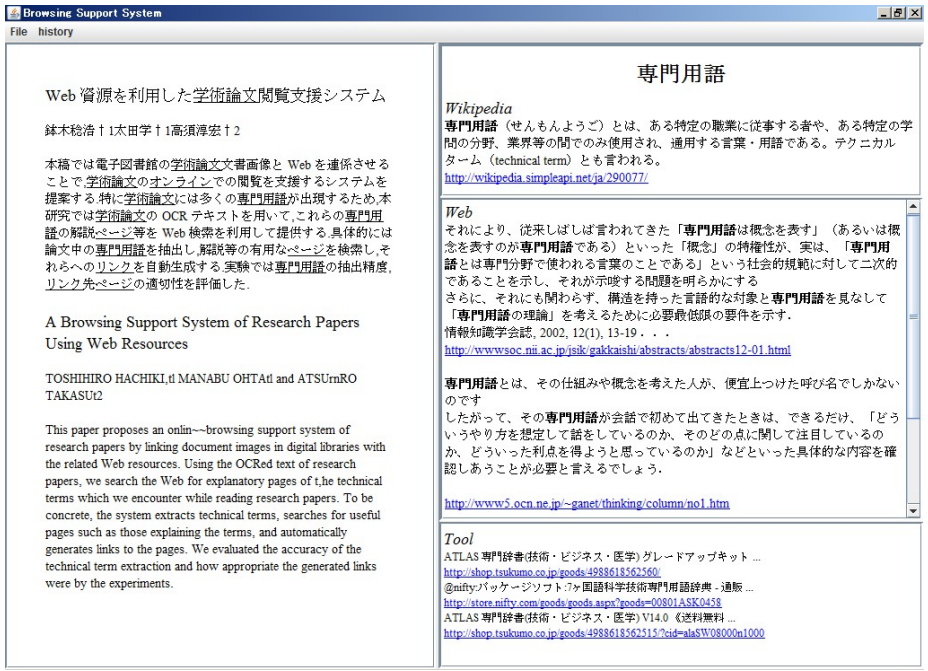

図 1 プロトタイプシステムの実行画面
Fig. 1 A screen shot of the prototype system

という記述があることが多いため，これらの文字列をツールページ検索の手掛かりとした．また alt 属性を考慮することでアンカーが画像であるときにも対応できる．

3.5 実装システム

図 1 に実装したプロトタイプシステムの実行画面を示す．これは便宜上本論文のタイトルページを OCR で認識したものを入力としたときの実行例であり，タイトルページにある和文表題，和文概要について閲覧支援を行っている．画面の左には対象論文の表題や著者名，概要などの主要な書誌要素を表示し，解説ページ，ツールページへのリンクをもつ用語には下線がついている．画面の右には選択した用語に関する Wikipedia の記事，Web 検索による解説ページ，そしてツールページへのリンクが表示されている．なお図 1 のこれらのリンクは，画面の左側で“ 専門用語 ”を選択すると表示される．

表 1 用語の抽出精度
Table 1 Technical term extraction accuracy

閾値 t	0.40	0.45	0.50	0.55	0.60
平均語数	7.0	8.4	10.3	11.9	13.6
再現率 R	0.51	0.59	0.70	0.78	0.82
適合率 P	0.83	0.80	0.78	0.75	0.68
F 値	0.63	0.68	0.74	0.76	0.74

4. 提案システムの評価実験

実験により提案システムの評価を行う．なお実験に使用したデータは，電子情報通信学会論文誌 6 年分の論文，計 713 件である．

4.1 用語抽出精度実験

実験データから無作為に選んだ 10 件の論文について，用語の抽出精度を調べる．抽出する語の数は累積スコアの閾値によるため，閾値の値を変化させて実験を行った．また評価のため，特徴語集合から用語を人手で選び出し，それらを正解としてシステムが抽出した語と比較した．適合率 P と再現率 R，F 値をそれぞれ求めた結果を**表 1** に示す．ここで平均語数とは一つの論文から用語として抽出された語数の平均である．

実験に使用した 10 件の論文より抽出した特徴語の中から，再現率の算出のため人手で選び出した正解の用語は 114 語であった．つまり一論文当り平均 11.4 語の用語が存在することになる．表 1 より閾値 $t = 0.55$ の場合，平均語数が最もこの値に近く，さらに F 値についても最も高い値となっている．よって以下の実験では $t = 0.55$ を用いた．

この閾値の値が大きいほど抽出する語は増えるが，その最適な値はデータやユーザに依存する．よって最適な閾値は想定されるユーザやサンプルデータを用いた予備実験により決定する必要がある．

4.2 解説ページの適切性の評価

4.1 節の実験とは別に実験データから無作為に選んだ 10 件の論文について，提案システムが表示した解説ページの適切性を評価する．なお Wikipedia 検索で共起語が得られる場合，Web 検索時の検索質問にその共起語を追加する．

この実験結果を**表 2** に示す．提案システムが抽出した全 104 語のうち，用語は 70 語であった．よって一論文当り用語を 7 語抽出していることになるが，これは一論文当り 3 語以上不適切な語を抽出しているともいえる．また通常論文は専門性が高いので，一般的な語

表 2 システムが抽出した用語数
Table 2 The numbers of extracted technical terms

抽出した語数	104
抽出した語のうち用語の数	70
リンクが存在する語数	70
リンクが存在する語のうち用語の数	52

表 3 解説ページの適切性
Table 3 Appropriateness of explanatory pages

表 2 の用語 52 語がもつ総リンク数	168
上記のうち適切なリンク数	105
適切なリンクがある用語数	43

表 4 Wikipedia 記事の共起語の有無による Web 検索結果の比較
Table 4 Web search results with and without co-occurrence terms of Wikipedia articles

	共起語無	共起語有
総リンク数	89	88
上記のうち適切なリンク数	59	66
適切なリンクがある用語数	25	29

の出現頻度は低い．例えば「商品」や「教室」といった語は論文中に出現することが少ないため出現した場合，式 (1)，(2) で提案した TF・IDF 法ではスコアが高くなり，用語として抽出されることがある．さらに OCR の認識誤りにより，語としては不適切なものが抽出される場合もある．より正確な用語抽出のために，特徴語の抽出手順やスコア付けの方法に改良の余地があると考えている．

次にリンクが存在する用語 52 語（表 2）について，リンク先の解説ページの適切性を評価した（**表 3**）．ここで，適切なリンクがある用語とは，用語のもつ各リンクのうち少なくとも一つに適切な解説ページへのリンクが存在するものである．つまり，用語 52 語に対して適切なリンクが少なくとも一つ存在する用語が 43 語あった．この実験結果より，用語に対し適切なリンクが存在するのは 43/70，すなわち 61 %，全リンクに対する適切なリンクの割合は 105/168，すなわち 63 %であった．解説ページ検索を誤る原因の一つには，「X とは異なり〜である」のように説明表現テンプレートにマッチするが，用語解説ではない文を含むページを検索してしまうことが挙げられる．これを改善するには，説明表現テンプレートを改良する必要がある．

4.3 Wikipedia の記事から抽出した共起語の Web 検索における効果

Wikipedia 検索の際に得られた共起語を，Web 検索に利用したときの効果を調べる．実験ではシステムが抽出した用語のうち，Wikipedia に記事が存在した 30 語を選び，共起語を用いた場合と用いない場合に分けて，Web 検索により得られるリンク先解説ページの適切性を評価する．

実験結果を**表 4** に示す．ここでリンク数とは，システムが Web 検索を利用して用語に対してはった解説ページへのリンクの数である．提案システムは，Web 検索において最大三つのリンク先をもつため，このリンク数の最大値は 90 である．また適切なリンクがある用語数とは，Web 検索による解説ページへのリンクのうち，少なくとも一つが適切なページへのリンクとなっている用語の数である．実験結果より，共起語を用いた方が総リンク数は一つだけ少なくなるものの，適切なリンク数は多くなっている．また共起語がある場合，30 語中 29 語の用語に適切なページへのリンクが存在している．表 4 は，Wikipedia の共起語が Web 検索精度の向上に有効であることを示している．

提案システムでは，例えば「NP」という用語に対して，「多項式」という共起語が Wikipedia の記事から得られる．共起語を用いない場合には，NP がナース・プラクティショナー（nurse practitioner）の頭字語として使われているページへリンクがはられていた．一方共起語をクエリに追加すると，非決定性多項式時間（non-deterministic polynomial time）という語を検索していることが明確になり，適切なページへのリンクがはられるようになった．ただし逆に精度が悪くなる場合もある．「パイプライン処理」という用語では，共起語に「ID」が得られた．Wikipedia の記事ではこれは Instruction Decode の意味で使われているが，Web 検索では ID が一般的によく使われる身分証明の意味で用いられているページを検索してしまう．このように異なる意味の同一の頭字語などでは，誤って不適切なページへリンクをはることがあった．

4.4 ツールページの適切性の評価

ツールページ検索により得られるリンク先ページの適切性を評価する．実験のため提案システムが抽出した用語から，「XML」や「データマイニング」といったツールが存在する語を 20 語選んだ．それらの用語についてツールページ検索を行い，取得したページについて適合判定を行った．本実験における適合性の判断基準では，リンク先ページが用語に関するツールを紹介するページ，またはツールをダウンロードできるページのとき適切なツールページとした．

表 5 に実験結果を示す．実験に用いた用語 20 語のすべてにおいて，一つ以上の適切なリンクが存在した．また生成したリンクの精度は 124/149，すなわち 83 %であった．

ツールページへのリンクとして不適切だったものについてエラー解析を行った結果，誤ったリンク先のうち約三割のページが，ページ内の広告中のリンクの文字列にマッチしたものであった．考えられる対処方法としては，検索対象を Web ページの本文のみに限定する

表 5 ツールページの適切性
Table 5 Appropriateness of tool pages

総リンク数	149
上記のうち適切なリンク数	124
適切なリンクがある用語数	20

ことが挙げられる．Web ページを主要部分とそうでない部分に分割する研究には，例えば，テキストに占めるリンクの割合等のヒューリスティクスを利用する方法[16] が提案されており，このような手法の応用も検討している．

その他の誤りには頭字語の曖昧性によるものがあった．例えば「SVM」という用語が，論文中では「Support Vector Machine」の意味で用いられているが，リンク先のツールページでは「Solaris Volume Manager」の頭字語として使われている場合があった．これは 4.3 節の Web 検索の誤りと同様の問題であるが，論文中の用語が出現する前後の文脈などを用いて，曖昧性を解消できる可能性がある．

5. ま と め

本稿では Web 資源を利用した学術論文の閲覧支援システムを提案した．提案システムは，学術論文の OCR テキストから専門用語を抽出し，それらに関連した有用なページを検索し，有用なページへのリンクを生成する．実装したプロトタイプシステムを用いて用語抽出精度，有用なページの適切性評価のための実験を行った．また Wikipedia の共起語を用いることで Web 検索の精度が向上することを確認した．

一方専門用語抽出や有用ページ検索においては，関連研究と比較し有用性の検証を行う必要がある．OCR の認識誤りへの対処についてもさらなる改善を行う予定である．また閲覧支援の対象を，現在のタイトルページのみから，将来的には論文全体まで広げることを考えている．

参 考 文 献

1) Ohta, M. and Takasu, A.: CRF-based Authors' Name Tagging for Scanned Documents, *In Proc. of JCDL'08*, pp.272–275 (2008).
2) Ohta, M., Yakushi, T. and Takasu, A.: Bibliographic Element Extraction from Scanned Documents Using Conditional Random Fields, *In Proc. of ICDIM 2008*, pp.99–104 (2008).
3) 薬師貴之，太田　学，高須淳宏：CRF を用いた学術論文 OCR テキストからの自動書誌要素抽出，情報処理学会論文誌 データベース， Vol.2，No. 2, pp.126–136 (2009).
4) Lafferty, J., M. and Pereria, F.: Conditional Random Fields: Probabilistic Models for Segmenting and labeling Sequence Data, *In Proc. of 18th International Conference on Machine Learning*, pp.282–289 (2001).
5) 久光　徹，丹羽芳樹，辻井潤一：タームの representativeness を測る，情報処理学会研究報告. 自然言語処理研究会報告， Vol.99, pp.115–122 (1999).
6) 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，自然言語処理， Vol.10，No.1, pp.27–45 (2003).
7) 中谷圭吾，鈴木　優，川越恭二：利用者の要求に応じた Web リンク自動生成手法，電子情報通信学会 第 16 回データ工学ワークショップ (DEWS2005) 論文集 (2005).
8) 中谷圭吾，鈴木　優，川越恭二：文書間類似度とキーワードを用いた Web リンク自動生成手法，*DBSJ Letters*, Vol.4, No.1, pp.85–88 (2005).
9) 藤井　敦，石川徹也：World Wide Web を用いた事典知識情報の抽出と組織化，電子情報通信学会論文誌， Vol.85-D-II, pp.300–307 (2002).
10) 土橋惇一，荒木健治：Web 文書を対象とした用語説明文抽出手法における抽出範囲の特定，情報処理学会研究報告， Vol.2006, No.1, pp.37–42 (2006).
11) Wikipedia，http://ja.wikipedia.org/.
12) Sen Project，http://ultimania.org/sen/.
13) Yahoo!ウェブ検索，http://serch.yahoo.co.jp/.
14) Wikipedia API，http://wikipedia.simpleapi.net/.
15) Yahoo!デベロッパーネットワーク，http://developer.yahoo.co.jp/.
16) 鶴田雅信，増山　繁：未知のサイトに含まれる Web ページからの主要部分抽出手法，言語処理学会第 14 回年次大会発表論文集 (2008).