特　集　「WWW 上の情報の知的アクセスのためのテキスト処理」

WWW 上のテキスト情報の知的統合

Advanced Integration of Multiple Text Information on the World Wide Web

難波　英嗣　広島市立大学情報科学部
Hidetsugu Nanba　School of Information Sciences, Hiroshima City University.
nanba@its.hiroshima-cu.ac.jp

Keywords: relationships between texts, integration, automatic summarization.

1. はじめに

WWW 上のテキスト情報の知的統合とは，関連するテキストを WWW 上から自動的に収集し，それらをまとめ，提示することで，ユーザの効率的な情報へのアクセスを可能にする技術である．本稿では，WWW 上のテキスト情報の統合技術，および関連する研究やシステムを紹介する．

近年，クローリング技術の発達，計算機の高速化および大容量化，検索技術の発達とともに，Web 検索エンジンなどを用いて，膨大な WWW 中の情報を検索することが可能になった．しかし，現在の Web 検索エンジンは，検索結果を検索キーワードとの適合度などに応じて順序づけされたリストとして提示するのが主流であり，検索結果が膨大な場合，そこからユーザが目的の情報を探し出すのは困難であることが多い．このような問題が生じる原因の一つは，多くの Web 検索エンジンが汎用的な目的でつくられているということと関係する．汎用的な Web 検索エンジンは，さまざまな形式，多様なトピックのテキストを処理する必要性から，テキストの内容に十分に踏み込んだ処理や，特定のテキスト形式に特化した処理が取りづらい．このような状況で，近年では，すべての WWW ページをカバーするのではなく，特定のトピックや目的に限定して，利用者の情報ニーズの種類ごとにきめ細かく対応しようというアプローチが生まれている[古関 01]．現在，ショッピング，ニュース，旅行，学術情報など，ある目的に特化したさまざまな検索エンジンが存在しており*1，一部のシステムは，運営を自動化している．本稿で扱う WWW 上のテキスト情報の知的統合は，このような目的特化検索エンジンの一部であると考えられる．自動化された目的特化検索エンジンの中には，HTML のテーブル解析技術などを用いて，さまざまなサイトから数値データなどを抽出し，比較するといったものもあるが，本稿では，テキスト処理に主眼を置いている．特に，関連する複数のテキスト間でテキスト情報を関連づけ，そこからユーザにとって有用な情報を抽出し，わかりやすく提示するための諸技術および関連システムについて述べる．

本稿の構成は以下のとおりである．2 章では，計算機上で WWW 上のテキスト情報の統合手順および要素技術について述べる．3 章では，いくつかの関連するシステムや研究を紹介する．4 章では，この分野の今後の展望について述べる．

2. テキスト情報統合の手順

WWW 上に存在する特定トピックやジャンルのテキスト情報を統合するには，まず，関連するテキストを WWW 上から網羅的に収集する必要がある．次に，テキスト中からトピックと関連する情報のみを抽出しなければならない．なぜならば，収集したテキスト中の情報がすべて特定のトピックに関するものであるとは限らないためである．さらに，抽出された情報を列挙するだけでは，ユーザにとってその中から必要な情報を探すのが困難なため，何らかの観点から分類・整理し，わかりやすく提示する必要がある．

以上をまとめると，テキスト情報の統合手順および要素技術は，大まかに以下に示す四つの段階が必要となる．なお，これらの手順は，目的や観点によって多少異なる場合もある．

(1) 特定トピックやジャンルのテキストの収集

あるトピックに関連するテキストを，Web 検索エンジンなどを利用して収集する．ここで述べるトピックとは，例えば，用語，人物名，組織名，製品名，イベントなどを指し，テキスト収集の際には，これらの語句をキーワードとして利用することができる．

[要素技術] Web 検索エンジンを用いた情報検索が代表的な方法であるが，このほかにも情報フィルタリング，特定のトピックや分野のテキストのみを収集するフォーカストクローリング[Chakrabarti 99]などの技術の利用も考えられる．

*1 例えば，SearchEngineWatch のページにさまざまな目的特化検索エンジンが紹介されている．
http://www.searchenginewatch.com/links/article.php/2156351

(2) Web文書からの関連情報の抽出

(1) で収集したテキストからトピックに関連する情報（語句，文，段落）を抽出する．

[要素技術] 一般的なテキストを対象にした場合には，情報抽出，パッセージ（部分テキスト）検索などの技術が有用であるが，Web文書の場合は，さらにHTML構造解析技術なども利用できると考えられる．

(3) 情報の統合

(2) で複数のテキストから抽出された情報を統合する．抽出された複数の情報が等価である場合，それらをすべてユーザに提示するのは冗長で効率的ではないため，まとめておく必要がある．

[要素技術] 関連するテキスト情報を分類するには，テキスト（あるいはパッセージ）分類技術が有用であると考えられる．また，テキスト間で類似する箇所を同定する方法に関しては，複数テキスト要約の分野で研究されており[奥村 99，奥村 02]，これらの技術が利用可能であると考えられる．

(4) 情報の提示

(3) で解析された関係をわかりやすく提示する．情報の提示方法は，複数のテキスト情報を要約して，テキスト形式で提示することも可能であるが，このほかに，表や図で提示する方法も考えられる．

[要素技術] 情報の提示方法は，対象によってさまざまであり，特に一般的な方法はない．次章で，いくつか具体的な事例を示す．

3. 事例紹介

これまでに，さまざまな目的や観点から，WWW上のテキスト情報を収集・統合するシステムが開発されている．以下に，いくつか例を示す．

(a) **用語の説明**[藤井 04，桜井 03]：ユーザが用語を入力すると，その用語を説明したWeb文書を収集し，説明箇所を抽出する．用語が多義語である場合は，語義ごとに分類して提示する．

(b) **人物，組織名，地名に関する情報** [McKeown 02，大槻 01，佐藤 01，山本 00]：ある人物，組織名，地名に関する属性情報やイベント情報を収集し，それらを統合する．

(c) **製品などに関する評判**[立石 04]：ある製品に関するさまざまな評判情報を収集し，それらが肯定的であるか，否定的であるかによって分類する．

(d) **WWW上の論文データ**[Lawrence 99，McCallum 99，Nanba 04]：WWW上に存在するPDFやPostscript形式の論文データを収集，それらの引用関係を解析し，論文データベースを構築．ある論文を引用する複数の論文の引用箇所（引用の文脈）を抽出，提示する．

(e) **ある事件（イベント）に関するニュース記事**[Calishain 03，McKeown 02，Radev 00，Radev 01]：WWW上の複数のニュースサイトからニュース記事を収集し，事件やイベントごとに分類・提示する．

本章では，紙面の都合上，(a) ～ (e) の一部を紹介する．システム紹介では，各システムの統合手順を2章で述べた手順別（(1) ～ (4)）に示す．また，各システムの簡単な説明，URL，関連論文なども合わせて提示する．関連文献中に手順の一部が明記されていない場合は省略する．なお，(a) の藤井の研究，(c) の立石らの研究は，本特集の藤井[藤井 04]，立石ら[立石 04]の記事を参照されたい．また，近年，WWW上のテキストを情報源とした質問応答システムが開発されているが，その多くは (a) と (b) に関連するものである．これらの関連システムに関しても藤井の解説[藤井 04]を参照されたい．

3・1 WWW上のニュース記事の統合

WWW上の複数のニュースサイトから記事を自動的に収集・統合し，カテゴリーなどに分類して提示するサービスがいくつかある．ここでは，Google News，MSN Newsbot，Newsblaster，NewsInEssenceを紹介する．

● Google News

説明: WWW上のニュースサイトから収集した情報を統合したニュース検索サイト．

URL: `http://news.google.com`

システム構築の手順:

(1) 約4 500のニュースサイトから定期的に記事情報を自動収集．

(2) 各ページからヘッドラインと写真画像を自動抽出．

(3) ほぼ同時期に報道された複数ニュースサイト上の同一トピックの記事をグループ化．

(4) トップページには，ユーザがページにアクセスした時点で最も注目を集めているニュース（最も多くの記事を含んだグループ）が表示される．これは，最も多くのニュースサイトで取り上げられているニュースは重要であるという考えに基づいている．また，統合されたニュースは，「国際」，「アメリカ国内」，「ビジネス」，「科学技術」，「スポーツ」，「エンターテイメント」，「医療・健康」の7カテゴリーに自動分類される．ユーザは，データベース中のニュースを，カテゴリー検索とキーワード検索することが可能．

その他: データベースには過去30日分のデータが蓄積されており，また1時間おきに更新されている．Google Newsのトップページでは，アメリカに関連するニュースを中心にまとめてあるが，オーストラリア，カナダ，フランス，ドイツ，インド，イタリア，ニュージーランド，イギリス各国のニュースを中心にまとめたバージョンのものもあり，`http://news.google.com` からたどって利用できる．

関連論文，解説など：[Calishain 03]

Google News のサイト: http://news.google.com/intl/en_ca/about_google_news.html

● MSN Newsbot

説明: Google News と同様のサービス．Google News との相違点は，ユーザの興味に応じてニュースサイトがパーソナライズされる点である．ユーザが MSN にアカウントを作成（無料）しサインした状態で Newsbot を利用すると，過去の利用状況やほかのユーザのパターンに基づいて，ユーザが興味をもちそうなニュースを薦めてくれる．

URL: http://uk.newsbot.msn.com

システム構築の手順:

(1) WWW 上の約 4 000 のニュースサイトから定期的に情報を収集．

(2) ほぼ同時期に報道された複数ニュースサイト上の同一トピックの記事をグループ化．グループ化されている記事間に成立する関係は，Google News とほぼ同じであると推測される．

(3)「ビジネス」,「スポーツ」といったいくつかのカテゴリーに分類．

その他: フランス，イタリア，スペイン，イギリス各国のニュースを中心にまとめた複数のバージョンが存在する．

● Columbia Newsblaster

説明: コロンビア大学の McKeown らのグループが開発したニュース要約 & 検索システム．Google News や Newsbot が複数のニュースサイト間で同一トピックの記事を同定し，分類することに主眼を置いているのに対し，さらに同一トピックの複数記事の内容を自動的に一つの要約にまとめてユーザに提示する機能をもっている．

URL: http://www1.cs.columbia.edu/nlp/newsblaster

システム構築の手順:

(1) 17 のニュースサイトからニュース記事データを定期的に収集．

(2) HTML タグを考慮してヘッドライン，ニュース記事本文，記事の署名，画像を抽出．

(3) まず，収集した記事をトピックごとにクラスタリングする．次に，関連する複数のイベントクラスタをグループ化し，六つのカテゴリー（「アメリカ国内」,「国際」,「金融」,「娯楽」,「科学技術」,「スポーツ」）に分類する．さらに，グループごとに要約を作成する．要約作成の際は，まず，グループ化されたテキスト集合がどのような性質のものであるのか（例えば「ある事件に関する記事とその続報記事集合」や「ある人物に関する記事」など）を判定し，その後テキストの性質ごとに異なる要約器を用いて要約を作成する．

(4) ユーザは，目的の要約をキーワードで検索するか，六つのカテゴリーの中から探す．

関連論文・解説など: [McKeown 02, 難波 02]

● NewsInEssence

説明: ミシガン大学の Radev らが開発したニュース要約 & 検索システム．Newsblaster と同様，複数の関連記事から要約を自動作成する．Newsblaster との違いは，ユーザが関心のあるトピックを入力すると，複数のニュースサイトからリアルタイムで記事を収集し，要約を作成する点である．上で紹介したほかのシステムは，グループ化される記事がほぼ 1 ～ 2 日の期間内であったのに対し，NewsInEssence では，数日～数週間にわたる比較的長期間の記事をグループ化し，要約を作成している．

URL: http://www.newsinessence.com

システム構築の手順:

(1) 2 種類の方法でニュース記事を収集する．一つは定期的にニュースサイトを巡回して記事を収集する方法であり，もう一つはユーザが関心をもっているニュースについて，リアルタイムでニュースサイトから関連記事を収集する方法である．

(2) 手順 (1) で収集された同一トピックの記事集合から，要約を作成．作成時に，記事間で類似する内容の文を検出している．要約作成は，入力された記事集合のトピックとの関連度と各文の記事中での位置などの情報に基づき，各文の重要度を計算し，要約率に応じてスコアの高い順に文を抽出するが，出力される要約の冗長性を考慮し，類似した 2 文はその両方とも要約に含めてしまわないよう配慮している．

(3) トップページには，ユーザが関心をもっている記事の URL や複数のキーワードを入力するためのフォームがあり，ユーザがフォームに記入すると，NewsInEssence は直ちにニュースサイトを巡回して関連記事を収集，要約を作成し，数分程度で結果を E メールで通知してくれる．このようにして作成された要約はサーバ側で保存してあり，他のユーザが閲覧することも可能である．

その他: このシステムで使われている要約器は WWW 上で公開されている．http://www.clsp.jhu.edu/ws2001/groups/asmd/

関連論文，解説: [Radev 00, Radev 01, 難波 02]

3・2 WWW 上の論文データの統合

近年，出版社や学会，あるいは研究者個人の Web ページで，Postscript や PDF といった形式のフルテキスト論文データを公開するケースが増えている．このような論文データを収集して論文データベースを構築し，検索可能にしたサービスがいくつかある．ここでは，CiteSeer, Cora, PRESRI の三つのシステムを紹介する．これら 3

システムは，いずれも引用文献索引データベースであり，また，ある論文がほかの論文から引用されている場合，その論文の被引用論文中における引用箇所（context）を提示できる点が特徴的である．このような引用箇所には，引用論文がどのような研究を行っているのかを手短に述べるだけでなく，どのような問題点があるのか，また他の研究にどのように利用できるのか，といった情報が記述されている．引用箇所は，被引用論文の著者から見た引用論文の一種の要約と考えることができる．したがって，ある論文に関する複数の引用箇所を提示することにより，その論文に関するさまざまな意見や見解をユーザが比較できる．以下，3システムを紹介する．

● CiteSeer（ResearchIndex）

説明: NEC Research Institute の Lawrence らが開発したシステム．WWW 上の英語論文を収集して構築．今回紹介する3システムの中では最も大規模な論文データベースを保持している．

URL: `http://citeseer.ist.psu.edu`

システム構築の手順:

（1）Web 検索エンジンを利用して“publications”，“papers”，“postscript”といった語を含んだ Web ページを検索し，これらのページを出発点としてクローリングして Postscript や PDF 形式の論文を収集する．このほかに HomePageSearch（`http://hpsearch.uni-trier.de/`）なども収集に利用している．

（2）（1）で収集した論文データは，PreScript（`http://www.nzdl.org`）を用いてテキストに変換した後，テキストのヘッダから，タイトル，著者名などの情報を抽出．抽出には，文字装飾やセンタリングなどのテキストのスタイルと，見出し語の情報を用いている．また，各論文データから，参考文献情報の抽出も行っている．参考文献の各書誌情報からは，タイトル，著者，著作年，ページ，引用識別子（[6]や[Gilese 97]などの本文中の文字列）の抽出を行っている．さらに，引用識別子を用いて，本文中の引用箇所も抽出している．

（3）手順（2）で，各論文データから抽出された書誌情報の同定を行い，同一論文はグルーピングする．結果として，論文集合全体の引用関係が解析されることになる．また，この処理により，ある論文とそれを引用する複数の論文の引用箇所が関連づけられるが，これらは，引用論文に関する一種の批評的な要約と考えることができる．ただし，CiteSeer も次に紹介する Cora も，引用の理由の分類までは行っていない．

（4）論文は，キーワードによる検索，論文間の引用関係をたどった検索およびカテゴリー検索が可能．ある論文が複数の論文から引用されている場合，その論文に関する複数の引用箇所が並べて提示される．これを見れば，ユーザはある論文に関するさまざまな見解，評価などがわかる．

関連論文，解説など: [Lawrence 99, 吉岡 02]

● Cora

説明: Just Research の McCallum（現 マサチューセッツ大）らが開発したシステム．WWW 上に存在する Postscript 形式の英語論文ファイル約 52 000 を収集して構築．論文データのクローリング，各論文からの書誌情報の抽出，収集した論文の分類といった主要な処理にはすべて何らかの形で機械学習を用いている．

URL: 現在サービスを停止しているが，Dr. McCallum のページ（`http://www.cs.umass.edu/~mccallum/code-data.html`）から，Cora で使われていたプログラムやデータの一部が入手可能．

システム構築の手順: テキストの収集にフォーカストクローラを用いる点はほかのシステムと異なるが，ほかの手順は基本的に CiteSeer と同じであるため，ここでは省略する．

関連論文，解説など: [McCallum 99]

● PRESRI [*2]

説明: 著者と東京工業大学奥村 学研究室のメンバが共同で開発したシステム．WWW 上の日英論文データを収集して構築．また，WWW 上の論文データと，ほかの論文データベースとの統合も可能．検索結果をグラフィカルに提示できる．データベースは，約 18 000 件の英語論文と約 2 000 件の日本語論文から構成される．

URL: `http://www.presri.com`

システム構築の手順: 基本的な手順は，CiteSeer や Cora とほぼ同じであるので，ここでは省略する．ただし，手順（3）で，引用箇所の情報からどのような理由で論文が引用されているのか（引用タイプ）を自動的に解析している点はほかのシステムと異なる．また，手順（4）で，検索結果を提示する際，論文間の引用関係をグラフィカルに提示できる点も異なる（図 1，口絵参照）．

関連論文，解説など: [難波 99，Nanba 04]

その他: 以下に，ほかのシステムにはない PRESRI の新しい機能について詳しく説明する．

（1）**論文検索機能**

PRESRI では，まずキーワードにより論文を検索する．検索結果はリスト表示される．検索された各論文の先頭にチェックボックスが表示される．ユーザは，関心のある論文をいくつか選択し，“Showing citation graph”のボタンを押すと，図 1 に示すような画面が表示される．図 1 において，ドットは一つの論文を，矢印は論文間の引用関係を示している．

---

*2 このシステムは，IPA（情報処理振興事業協会）の支援による未踏ソフトウェア創造事業の成果である．

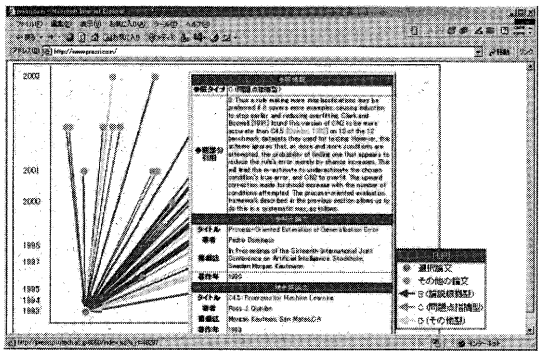

図1 引用関係のグラフィカルな提示（PRESRI）

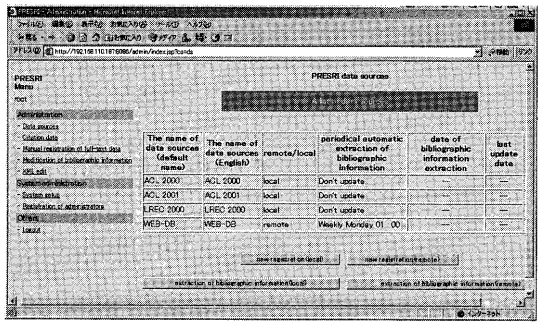

図2 複数論文データベースの統合（PRESRI）

キーワード検索過程でユーザが選択した論文は“Selected papers”として赤いドットで示される．また，PRESRIは，“Selected papers”と関連のある論文を論文データベースから自動的に収集し“Other papers”として一緒に提示する．矢印は引用タイプ（C，B，O）ごとに色分けされて表示される．図中のドットにカーソルを重ねると，各論文のタイトルや著者名などがポップアップウィンドウ内に表示される．また，矢印にカーソルを重ねると，論文中の引用箇所がポップアップウィンドウ内に表示される．図では，[Quinlan 93]に関する引用箇所が表示されている．このように，引用タイプを提示することで，「ある論文の成果がどのように利用されているのか（type B）」，「ある論文はどのような問題点を指摘され，それがどのように解決されているのか（type C）」といった情報がわかる．また，論文間の関係をグラフィカルに表示することで，その分野の研究の変遷や，個々の論文の分野の中での位置づけが一目でわかる．

(2) 複数論文データベースの統合機能

論文データは，WWW上で入手できるもの以外にも数多く存在する．例えば，学会や出版社が保持する論文データベース，国際会議などで予稿集の代わりに配布されるCD-ROMなどがある．PRESRIは，WWW上の論文データとこれらの論文データベースを統合することも可能である．複数データベースの統合などのデータベース管理は，すべてWebブラウザ上で行える．図2はWWW上の論文データ（WEB-DB）と複数のCD-ROMデータを統合する画面である．なお，データベースの著作権の関係上，統合されたデータは，不特定多数のユーザが利用できない場合があるため，複数データベースを統合する際には，この点にも配慮している．

3・3 人物，組織名，地名に関する情報収集・統合

佐藤らはWITプロジェクトにおいて，さまざまな側面からWWW上のテキストデータの統合を試み，システムを構築している[大槻 01, 桜井 03, 佐藤 00, 山本 00]．本稿では，WWWを利用した住所検索に関する研究[佐藤 01]を紹介する．

● WWWを利用した住所検索

説明: 与えられた名称から，その名称に対する住所情報を，WWWを利用して探し出す．名称が同一でも対象が異なる場合，対象ごとに住所情報をまとめて結果を提示．

URL: `http://wit.kuee.kyoto-u.ac.jp/wit/jap/`

システム構築の手順:

(1) Web検索エンジンを用い，入力された名称を含むWebページを収集．

(2) 各ページから，HTMLタグを利用して，住所情報の記述されている領域を抽出し，さらにそこから名称，住所，郵便番号，電話番号，URL（名称がアンカータグによって囲まれている場合）を抽出．

(3) (2)で各ページから抽出されたデータ間の同一性を判定し，同一対象を表すデータをまとめる．データをまとめる際，各ページから抽出された名称に関する属性値を利用する．例えば，二つのデータの電話番号が一致していれば二つは同一であると考えられるが，一つの組織に複数の電話番号が存在することがあるため，電話番号が異なっていても対象が異なるとは限らない．一方，二つのデータの住所が異なっていれば，二つは異なる対象であると考えられるが，同一住所でも組織が異なる場合があるため（例えば同一ビル内の異なる組織），住所が同一でも必ずしもデータが同一のものであるとは限らない．このような属性ごとの識別能力を考慮し，データの同一性判定を行っている．

(4) 結果は，「名称」，「住所」，「電話番号」，「URL」，「住所情報が抽出されたURL」が対象ごとに提示される．

関連論文，解説など：[佐藤 01]

4. おわりに

本稿では，WWW上のテキスト情報を統合する諸技術について述べ，関連システムを紹介した．本章では，今後の展望および課題について簡単に述べる．2章でも述べたが，WWW上の特定トピックのテキスト情報を統合す

るうえで，そのトピックに関連するテキストを網羅的に収集する必要がある．そのためには，情報検索やフォーカストクローリングといった要素技術の改良も必要であるが，今後は多言語化も一つの重要な課題になると考えられる．例えば，ある分野の研究動向を知るためには，さまざまな言語で記述された論文を収集する必要がある．3章で紹介した論文検索システムPRESRIは，こうした多言語化への取組みの一つであるといえる．

この分野の今後の課題について，もう一点述べる．WWW上には，客観的な事実だけでなく，主観的な情報，不確かな情報など，さまざまな情報が混在している．このため，テキスト間の関係も，より多様なものになっていると推測される．3章で述べた関連研究やシステムでも，ある程度，WWW上のテキストやテキスト間の関係の多様性を考慮しているが，それぞれの目的に特化し，独自につくられている部分が多い．今後のこの分野の発展のためにも，ある程度汎用的に利用可能なリソースを整備することが必要になってくると考えられる．

◇ 参 考 文 献 ◇

[Chakrabarti 99] Chakrabarti, S., Berg, M. and Dom, B.: Focused Crawling: A new approach to topic-specific web resource discovery, *Proceedings of 8th WWW Conference* (1999)

[Calishain 03] Calishain, T. and Dornfest, R.: Google Hacks プロが使うテクニック & ツール100選，オライリー・ジャパン（2003）

[藤井 04] 藤井敦: 百科辞典としてのWWW，人工知能学会誌，Vol.19，No.3, pp. 296-301（2004）

[古関 01] 古関義幸，福島俊一: 新世代検索ポータル技術，2001年情報学シンポジウム講演論文集，pp.59-66（2001）

[Lawrence 99] Lawrence, S., Giles, L. and Bollacker, K.: Digital libraries and autonomous citation indexing, *IEEE Computer*, Vol.32, No.6, pp.67-71（1999）

[McCallum 99] McCallum, A., Nigam, K., Rennie, J. and Seymore, K.: Building domain-specific search engines with machine learning techniques, *Proceedings of AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace*（1999）

[McKeown 02] McKeown, K.R., Barzilay, R., Evans, D., Hatzivassilogtou, V. Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B. and Sigelman, S.: Tracking and summarizing news on a daily basis with Columbia' s Newsblaster, *Proceedings of HLT' 02*（2002）

[難波 99] 難波英嗣，奥村 学: 論文間の参照情報を考慮したサーベイ論文作成支援システムの開発，自然言語処理，Vol.6，No.5，pp.43-62（1999）

[難波 02] 難波英嗣，奥村 学: ここまで来た自動要約，情報処理，Vol. 43，No.12，pp.1287-1294（2002）

[Nanba 04] Nanba, H., Abekawa, T., Okumura, M. and Saito, S.: Bilingual PRESRI — Integration of multiple research paper databases —, *Proceedings of RIAO 2004*（2004）（to appear）

[奥村 99] 奥村 学，難波英嗣: テキスト自動要約に関する研究動向，自然言語処理，Vol.6，No.6，pp.1-26（1999）

[奥村 02] 奥村 学，難波英嗣: テキスト自動要約に関する最近の話題，自然言語処理，Vol.9，No.4，pp.97-116（2002）

[大槻 01] 大槻洋輔，佐藤理史: 地域情報ウェブディレクトリの自動編集，情報処理学会論文誌，Vol.42，No.9，pp.2310-2318（2001）

[Radev 00] Radev, D.R., Jing, H. and Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies, *Proceedings of ANLP/NAACL 2000 Workshop: Automatic Summarization*, pp.21-30（2000）

[Radev 01] Radev, D.R., Blair-Goldensohn, S., Zhang, Z. and Raghavan, R.S.: Interactive, domain-independent identification and summarization of topically related news articles, *Proceedings of Fifth European Conference on Research and Advanced Technology for Digital Libraries*（2001）

[桜井 02] 桜井 裕，佐藤理史: ワールドワイドウェブを利用した用語説明の自動生成，情報処理学会論文誌，Vol.43，No.5，pp.1470-1480（2002）

[佐藤 00] 佐藤理史: ワールドワイドウェブを利用した住所検索，情報処理学会論文誌，Vol.42，No.1，pp.59-67（2001）

[立石 04] 立石健二，石黒義英，福島俊一: インターネットからの評判情報検索，人工知能学会誌，Vol.19，No.3, pp. 317-323（2004）

[山本 00] 山本あゆみ，佐藤理史: ワールドワイドウェブからの人物情報の自動収集，情報処理学会研究報告，No. 2000-ICS-119，pp.173-180（2000）

[吉岡 02] 吉岡真治: ResearchIndex: 新しい形の電子図書館，人工知能学会誌，Vol.17，No.5，pp.649-650（2002）

2004年3月28日 受理

著 者 紹 介

難波 英嗣（正会員）

1996年東京理科大学理工学部電気工学科卒業．1998年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．2001年同大学院大学 情報科学研究科博士後期課程修了．博士（情報科学）．同年，日本学術振興会特別研究員（PD）．2002年東京工業大学精密工学研究所助手．同年，広島市立大学情報科学部知能情報システム工学科講師，現在に至る．テキスト自動要約，Web上のテキスト情報の知的統合に関する研究に従事．言語処理学会，情報処理学会，ACL，ACM各会員．