セレンディピティを促す論文検索ツール 「ふわっと関連検索」

高久雅生
物質・材料研究機構 科学情報室
〒305-0047 茨城県つくば市千現 1-2-1
TAKAKU.Masao@nims.go.jp

江草由佳
国立教育政策研究所 教育研究情報センター
〒100-8951 東京都千代田区霞が関 3-2-2
yuka@nier.go.jp

概要

学術論文に対する検索体験として，セレンディピティを促すツール「ふわっと関連検索」を提案する．国立情報学研究所が提供する論文データベース CiNii API を対象とした検索ツールを通じて，その有効性を示す．本手法の特長は，類似文書検索機能をもたない従来型の論文データベースに対して，特徴ベクトル抽出と検索クエリ発行方法を工夫することにより，簡易的な類似文書検索を実現する点にある．2010 年 1 月の検索ツールの公開から一ヶ月間でのアクセスは約 1,800 件程度あった．本稿では，利用事例の報告を通じて，論文との新たな出会いを得るための検索ツールの可能性を示す．

キーワード

論文検索，文書類似度，マッシュアップ，Web API

A Serendipitous Scholarly Search Engine — "*Fuwatto Search*"

Masao Takaku
Scientific Information Office,
National Institute for Materials Science
1-2-1 Sengen, Tsukuba, Ibaraki, Japan
TAKAKU.Masao@nims.go.jp

Yuka Egusa
Educational Resources Research Center,
National Institute for Educational Policy Research
3-2-2 Kasumigaseki, Chiyoda, Tokyo, Japan
yuka@nier.go.jp

Abstract

The authors propose a novel search method, *Fuwatto search*, that allows users to find scholarly documents in a serendipitous way. And we also present an implementation of the method as *Fuwatto CiNii Search Engine* for targeting CiNii database service, provided by National Institute of Informatics. Fuwatto search provides psuedo related search capabilities between any text documents and a scholarly database. The Fuwatto CiNii Search Engine has been launched in January 2010, and since then it aquired approximately 1,800 page views. A case study of Fuwatto CiNii Search Engine shows a new possibility for users to encounter with scholarly articles.

Keywords

Document retrieval, Document similarity, Mashup, Web API

1 はじめに

学術論文は，研究者による知的成果の単位として，もっとも重要なもののひとつであり，研究活動そのものの可視化や，研究上の議論の接続点として重要な役割を果たしている．

一方で，学術論文は単に学術研究の世界に閉じているのではなく，一般の市民社会，生活とも結びついている．たとえば，列車に乗って旅をしたり，鉄道関連の話題を好む人々にとっては，列車運行のための待ち行列モデル，列車やレールなどに使われる鋼材の耐久劣性といった学術研究で扱われる話題を，知的な趣味の一部として，学術専門の内容であったとしても十分な価値をもつものとして受容できるだろう．

市民生活の知的メディアとしての，学術論文との出会いの場を創出することは，科学・技術の議論の多様性を確保し，社会における学術のあり方を担保するためにも欠かせないものである [1].

論文との出会いを生むための障害のひとつに用語体系の違いが挙げられる．学術における議論では専門家同士の厳密な定義にもとづく専門用語が使われており，それらの語彙を日常的に使うことがなく，用語を意識していない市民が，潜在的な知的ニーズを持つ専門分野の論文に出会う機会を得ることは，その第一段階に困難があるといえる．

そこで本稿では，このような具体的な語彙を適切に選択することが難しい状況にあっても，比較的容易に検索を行い，さまざまな関連分野の専門的な論文を発見できるように，1）明示的な検索キーワードの入力を必要とせず，2）自らの興味がある任意のテキスト内容から関連文献をひきだすことができる検索方式を提案する．このような検索方式は，ユーザによる学術論文の発見に役立つメリットがあるだけでなく，データベース提供者にとっては，これまでディープウェブの中で発見されてこなかった論文の再発見を促すというメリットもある．「ふわっと関連検索」と名付けた提案手法による検索は，さまざまな分野で蓄積された学術論文により気軽に出会い，アクセスできる環境を提供する．国立情報学研究所が提供する CiNii[2] を対象として，提案手法を実装したツール「ふわっと CiNii 関連検索」を通じて，その有効性を示す．

2　関連研究

以下では，1）科学技術情報の提供と論文情報との接続，2）文書類似度を活用した学術情報検索，という 2 つの観点から，本研究と関連した研究を紹介する．

科学技術情報の提供と論文情報との接続という観点からは，科学技術振興機構が 2009 年に提供を開始したサービス J-GLOBAL[3] がある．J-GLOBAL は「つながる・ひろがる・ひらめく」をキャッチコピーとした，論文検索をふくむ科学技術情報提供サービスであり，その API を通じた機能のひとつとして，科学技術情報のポータルサイト「SciencePortal」上の記事に対して，文書類似度を用いた類似文献や類似特許を自動的に提示する機能を提供している [4].

「Science and You」[5] は北海道大学科学技術コミュニケーター養成ユニット (CoSTEP) が運営するサービスで，ブログ記事の内容と科学技術情報に関するトピック記事とを結び付ける類似文書検索機能を実装し，ブログパーツとして提供している．

寸田は，宮崎大附属図書館 OPAC における JuNii+ 論文検索結果の自動表示機能を提案している [6]．この OPAC システムでは，図書蔵書検索で得られた書籍情報をクエリとして JuNii+ に発行し，自動的に関連論文の一覧をその書籍情報表示画面で提示している．

一方，文書類似度を用いた検索という観点では，高野ら [7] は，キーワードと文書群から得られる用語間の連想関係を提示する連想検索の重要性を提唱するとともに，「Webcat-plus」[8] や「IMAGINE・想」[9][10] といった検索サービスを提供して，その有効性を示している．

3　ふわっと関連検索

「ふわっと関連検索」は，任意のテキストを対象に，類似文書検索を実現する手法である．図 1 に，システムの概要図を示す．以下では，この手法について説明する．

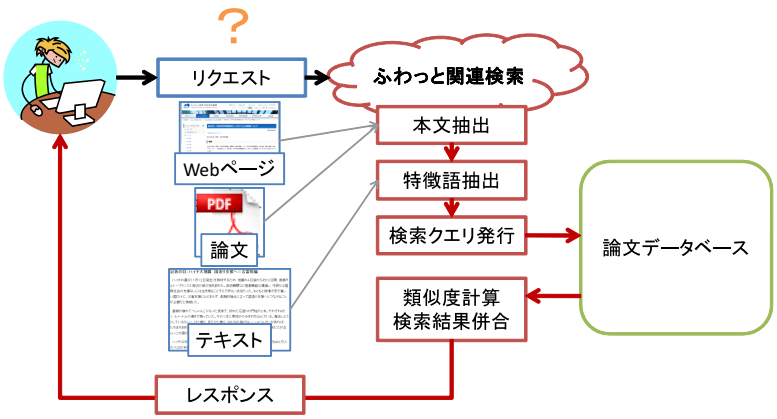

図 1　システムの概要図

1. 本文抽出
   入力に PDF などのプレインテキスト以外の形式が指定された場合には，当該データの取得とテキストの抽出を行う．また，対象テキストとして Web ページが指定された場合は，当該ページを取得し，本文テキストを抽出する．この際，HTML タグを除く等の処理を行う．
2. 特徴語抽出（テキスト中のキーワードの重み付け）
   入力されたテキストまたは抽出した本文テキストをもとに，単語分割を行い，各単語それぞれのテキスト中での出現回数（tf）と，その単語の生起確率をかけあわせて，重み付けを行い，その重みを特徴語スコアとする．
3. 検索クエリの発行
   前項の処理で得られた特徴語群ベクトルから上位 n 件の単語を，論文データベースに検索クエリとして発行する．この際，検索結果がゼロヒットとなる特徴語は除外し，スコア順に下位の特徴語を順次，論文データベースに問い合わせ，データベース中から検索結果が得られる，n 件の特徴語リストを求める．次に，n 件の特徴語をすべて含む AND 条件式をクエリとして，論文データベースに検索をおこなう．この検索結果がゼロヒットとなる場合には，特徴語スコアの最下位 1 件を除外したうえで，$n-1$ 件の特徴語リストを AND 条件式として，クエリ発行する．以下，同様に，最終的に m 件以上の検索結果が得られるまで，漸次的に特徴語スコアの低い語から，AND 条件式に用いる特徴語を減らしていく．
4. 検索結果の提示
   前項で得られた，AND 条件式による検索結果を，詳細度の高い順に併合していき，最終的な検索結果ランキングとして提示する．

このような手順を採用した理由は，1）最終的な検索結果がゼロヒットとなる確率を減らしてできるだけ多くの論文情報との出会いを生むことと，2）詳細な文書類似検索が実装されていない論文データベースに対しても単純なキーワード検索だけで簡易的な類似論文検索を実装することを意図したためである．

4　CiNii API を対象とした事例

4.1　概要

国立情報学研究所が提供する論文データベース CiNii[2] を対象として本手法を実装した「ふわっと CiNii 関連検索」について述べる．

「ふわっと CiNii 関連検索」では，CiNii を対象として，検索問い合わせには OpenSearch プロトコルによる Web API[11] を利用した．本文抽出には Web ページからの本文抽出モジュール extractcontent.rb[12] を用いて，できるだけ自然な本文部分の抽出を行うようにした．特徴語抽出には形態素解析ツール MeCab[13] を用い，ノイズを減らすために名詞・形容詞の自立語のみを対象とし，英単語の場合にはストップワード [14] を用いて，不要な抽出語が含まれないようにした．また，特徴語の生起確率としては，MeCab および mecab-ipadic がテキスト解析時に出力する単語生起コストを対数化した値を用いた．関連検索のためのパラメータとしては，特徴語リストから使用する語数には $n = 10$ を用い，検索結果件数としては $m = 20$ を設定した．

図 2 に「ふわっと CiNii 関連検索」のトップページを示す．利用者が検索する方法には，テキストを直接入力する「文章から検索」と，指定 URL をもとに検索する「ウェブページから検索」の 2 種類をタブ型インタフェースにより用意している．また，どのような検索がおこなわれるかイメージしやすいよう，朝日新聞，日本経済新聞の 2 紙の社説記事と毎日新聞のコラム記事「記者の目」を使った検索例を試すリンクを付けている．さらに図 3 に，検索結果画面例を示す．検索結果一覧に表示された論文タイトルからは，CiNii 上の当該論文へリンクが張られている．

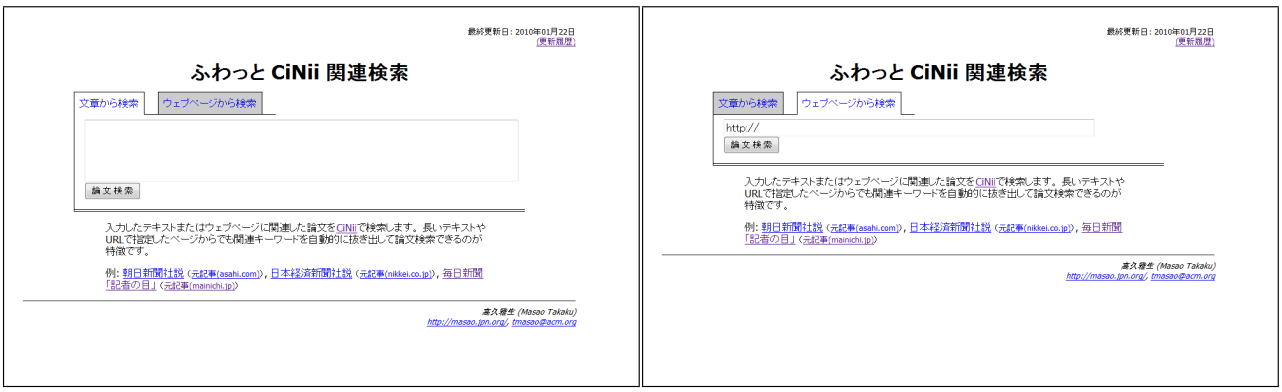

図 2　CiNii API を対象とした論文検索ツール「ふわっと CiNii 関連検索」トップページ（左:「文章から検索」，右:「ウェブページから検索」）

図 3　「ふわっと CiNii 関連検索」の検索結果画面例

4.2　利用事例

「ふわっと CiNii 関連検索」は，2010 年 1 月 10 日に公開し，Twitter 上で公開を宣言した（図 4）．図 5 に，アクセス利用の様子を示す．ページアクセス総数は 1,800 件を超えており，公開直後の数日に高いアクセス数を示し，その後は断続的な利用がおこなわれていることが分かる．公開から 2010 年 2 月 12 日までのアクセス利用は，IP アドレスをもとに計数したアクセス利用者数が 377 名，テキスト検索利用回数が 1,017 回，URL 検索で使用されたユニーク URL が 166 ページであった．

@tmasao **年末年始に作った、CiNii API ベースの論文検索ツール「ふわっと CiNii 関連検索」を公開します。何かお気づきの点などありましたら、お気軽にご連絡いただければうれしいです。** http://bit.ly/4JfHh0 #cinii #fuwatto
posted at 07:29:26
@tmasao **いわゆる連想検索風のツールなので、既知文献検索などよりは、テキストなどを元に気づきを得るための検索という使い方に適しているとおもいます。** http://bit.ly/4JfHh0 #cinii #fuwatto
posted at 07:29:57

図 4　Twitter における「ふわっと CiNii 関連検索」の公開アナウンス（2010 年 1 月 10 日）

初出:　http://twitter.com/tmasao/status/7571380397
http://twitter.com/tmasao/status/7571394545

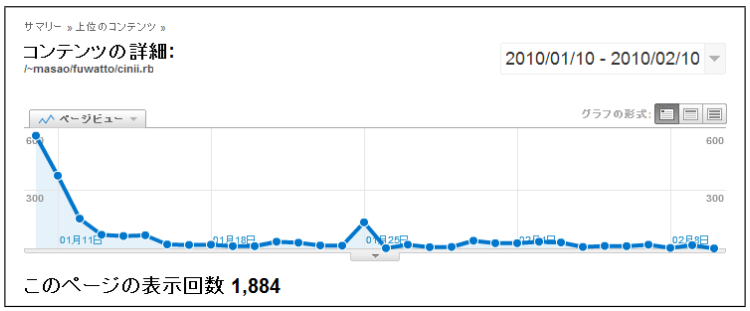

図 5　ふわっと CiNii 関連検索の公開後のアクセス状況（Google Analytics による）

表 1 に，頻繁に検索された URL とそれに対する検索結果の抜粋を示す．表 1 内の例 1 の検索対象ページは，列車運行の自動化と可視化を論じた論文であり，2010 年 2 月 2 日に「プロフェッショナル仕事の流儀」という NHK の番組でとりあげられた鉄道ダイヤ作成担当者が共著者となっていた論文 PDF である．論文は列車運行データの処理を主題とした内容であり，抽出語リストにもその主題を表現する単語が抽出されていることが分かる．実際に検索に用いられた条件式は「列車 遅延 ダイヤ 運行」までの 4 つの特徴語をもとにしたクエリであった．この例では，Web 上にすでにオープンアクセス論文として提供されている内容をもとに，CiNii 上の論文を検索する例を示した．この例は，本システムの公開後にもっとも検索アクセスされた Web ページであり，テレビ番組放映当時に出演者名で Google 検索するなどして発見された論文を，さらに本システムが CiNii 上の他の関連論文とひもづけることによって，一般の視聴者の視点をより専門的研究内容への誘導に成功した事例ではないかと思われる．

表 1 内の例 2 の検索対象ページは，百科事典サイトウィキペディア上の一記事であり，戦国時代の日本に渡来したイタリア人宣教師をとりあげた内容である．実際に検索に用いられた条件式は「日本 ヴァリニャーノ イエズス 巡察」までの 4 つの特徴語をもとにしたクエリであった．検索結果には，日本史学やキリスト教史学の分野における学術論文が挙がっており，百科事典による歴史上の人物に関する概要説明から，より詳細な細分化された話題ごとの学術論文につながる事例となっている．

4.3　考察: 実利用における応答速度

3 節で述べたように，本システムは内部に検索インデックス等は一切保持せずに，CiNii API に検索クエリを発行することにより実現されている．このような設計の場合，対象となる論文データベース（今回は CiNii）との間のネットワークを経由した検索クエリの発行回数が，ユーザから見えるシステムの応答速度に反映されることとなる．また，提案手法では特徴語ごとの組み合わせにより検索クエリを生成するため，検索に使用する抽出語数を示すパラメータ n と，検索結果として提示する論文書誌件数 m によって，検索クエリの発行回数が変化することとなる．

例 1 および例 2 に示した URL 指定検索において，検索結果を提示するまでの実行時間は，それぞれ 2.25 秒，2.40 秒である（実行時間は 5 回試行の平均値）．この際，例 1 および例 2 における HTTP を通じたリクエスト送出回数は，いずれも 20 回である．

図 6 に，n を，10 から 1 まで変化させた場合の実行時間の変化の様子を示す．パラメータ n の値が減るにしたがって，実行所要時間も減ることが分かる．たとえば，$n = 5$ を用いた場合は，当初の場合（$n = 10$）と比べて，事例 1 で 0.69 秒（31%），事例 2 で 1.01 秒（26%）も実行時間を減らすことができる．

本システムの提供にあたっては，できるだけ多くの抽出語を用いることによって関連論文のヒット漏れを減らすことを念頭に置き，かつ，CiNii が高速性・頑健性に優れている [15] ことから，$n = 10$ を設定している．しかし，ネットワーク上での HTTP アクセスに時間がかかる場合や，対象データベースが応答に時間がかかる場合などには，これらのパラメータの調整により，実行所要時間を制御して，エンドユーザに対する応答速度を改善することができる．さらには，対象データベースに適するように，もしくは，入力テキストごとにパラメータを調整することによって，適切な応答時間と検索性能を満足するような方策も考えられるが，このようなパラメータの最適化に関しては本稿の範囲を超えるため，今後の課題として残されている．

表 1　「ふわっと CiNii 関連検索」による検索結果例

URL・タイトル・抽出語	検索結果（上位 5 件までの抜粋）
例 1 http://www.tomii.cs.it-chiba.ac.jp/kashikaJRAIL.pdf タイトル: 列車運行実績データの可視化 抽出語: 1. 列車 2. 遅延 3. ダイヤ 4. 運行 5.（実績） 6.（データ） 7.（表示） 8.（発生） 9.（計画） 10.（問題）	1. 第 93 回運輸政策コロキウム 都市鉄道の運行ダイヤ過密化に伴う列車遅延の波及に関する研究 仮屋崎, 圭司; 岩倉, 成志; 森地, 茂 運輸政策研究, 2009/Win. 2. スイス連邦鉄道における接続を重視した新しい運行管理手法 : 戦略的施策から実際の運営の場に至るまでの余裕時分の活用手法 (〈特集〉鉄道のスケジューリング問題) Laube,Felix; Luthi,Marco; 富井, 規雄 オペレーションズ・リサーチ : 経営の科学, 2008-08-01 3. スプレッドシートを用いた協調推論型知識調整方式 : 列車運行予測の高精度化へ向けて 江口, 俊宏; 鶴田, 節夫 全国大会講演論文集, 1991-02-25 4. 列車運行予測のための協調推論型知識調整方式 鶴田, 節夫; 江口, 俊宏; 松本, 邦顕 全国大会講演論文集, 1990-09-04 5. 2-E-12 列車ダイヤ遅延時の乗務員スケジュール修正問題 (スケジューリング (2)) 三浦, 礼; 今泉, 淳; 森戸, 晋; 福村, 直登 日本オペレーションズ・リサーチ学会春季研究発表会アブストラクト集, 2007-03-28
例 2 http://ja.wikipedia.org/wiki/ア　レッサンドロ・ヴァリニャーノ タイトル: アレッサンドロ・ヴァリニャーノ – Wikipedia 抽出語: 1. 日本 2. ヴァリニャーノ 3. イエズス 4. 巡察 5.（ヨーロッパ） 6.（月） 7.（インド） 8.（中国） 9.（宣教） 10.（編集）	1. ヴァリニャ-ノ「日本巡察記」—イエズス会総長への機密報告 (外国人の見た日本・日本人〈特集〉) — (江戸時代以前の日本論) 東光, 博英 国文学解釈と鑑賞, 1995/03 2. イエズス会東インド巡察師アレッサンドロ・ヴァリニャ-ノと「日本の上長 (じょうちょう) のための規則」 高橋, 裕史 基督教學, 1994/07 3. 日本・キリシタン音楽教育の原点 : 南蛮文化との出会い; イエズス会士 A. ヴァリニャーノによるミッション教育の軌跡の探訪 進藤, 務子 久留米信愛女学院短期大学研究紀要, 2007-07-00 4. 日本及び中国におけるイエズス会の布教方策 : ヴァリニャーノの「適応主義」をめぐって 狭間, 芳樹 アジア・キリスト教・多元性, 2005-03 5. 十六世紀スペイン語文書の日本記述における「権力・空間」イメージについての一考察—A・ヴァリニャーノの『日本諸事要録』(一五八三年) を中心に 椎名, 浩 イスパニア図書, 2008/秋

※抽出語リスト中において括弧書きの語は検索条件に加えられなかった語を示す.

5　おわりに

本稿では，検索キーワードの明示的な入力を必要としない「ふわっと関連検索」を提案し，CiNii API を対象とした実装とその利用事例を通じて，その有用性を示した.

今後は，1）情報ニーズと主題トピックに対する検索結果の適合性評価，2）特徴語抽出および類似度計算手法の改良，3）ふわっと関連検索 API の提供，4）CiNii 以外の論文データベースを対象としたサービスの

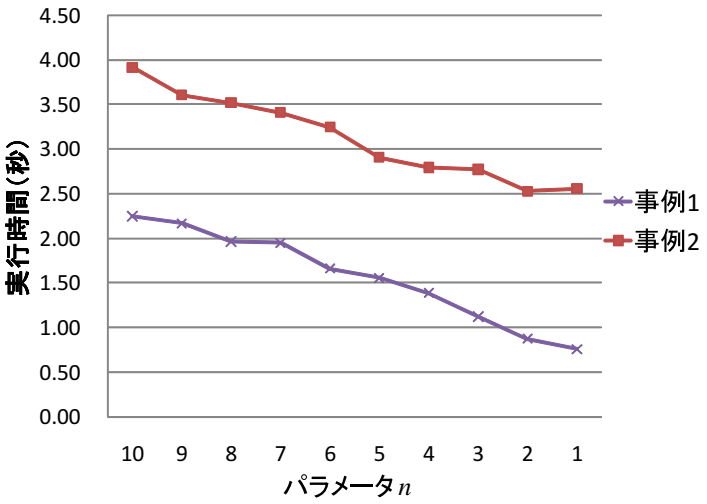

図 6　パラメータ $n = 10, ..., 1$ における実行時間の比較

提供，を予定しており，論文との出会いを広げるツールの完成を目指す．

参照文献

[1] 第 3 期科学技術基本計画, 2006. 平成 18 年 3 月 28 日閣議決定.

[2] 国立情報学研究所. CiNii - NII 論文情報ナビゲータ. http://ci.nii.ac.jp（アクセス日 2010 年 2 月 10 日）.

[3] 科学技術振興機構. J-GLOBAL. http://jglobal.jst.go.jp（アクセス日 2010 年 2 月 10 日）.

[4] 松邑勝治, 黒沢努, 関根基樹, 矢口学, 植松利晃, 加藤治. 「J-GLOBAL」試行版（β版）の構築と今後の展望. 情報管理, Vol. 52, No. 3, pp. 150–157, 2009.

[5] 北海道大学科学技術コミュニケーター養成ユニット (CoSTEP). Science and You. http://you.costep.jp（アクセス日 2010 年 2 月 10 日）.

[6] 寸田五郎. 宮崎大学附属図書館における OPAC と JuNii+ のマッシュアップ. 大学の図書館, Vol. 27, No. 7, pp. 145–146, 2008.

[7] 高野明彦, 西岡真吾, 丹羽芳樹. 連想に基づく情報アクセス技術 : 汎用連想計算エンジン GETA を用いて. 情報の科学と技術, Vol. 54, No. 12, pp. 634–639, 2004.

[8] 国立情報学研究所. Webcat Plus. http://webcatplus.nii.ac.jp（アクセス日 2010 年 2 月 10 日）.

[9] 想 — IMAGINE Book Search. http://imagine.bookmap.info（アクセス日 2010 年 2 月 10 日）.

[10] 小池勇治, 西岡真吾, 森本武資, 丸川雄三, 高野明彦. 分散連想計算サーバー群を統合する連想検索システム「想・IMAGINE」. 情報処理学会研究報告 自然言語処理研究会報告, pp. 31–36, 2008.

[11] CiNii - 外部提供インターフェースについて. http://ci.nii.ac.jp/info/ja/if_opensearch.html（アクセス日 2010 年 2 月 10 日）.

[12] 中谷秀洋. Web ページの本文抽出. http://labs.cybozu.co.jp/blog/nakatani/2007/09/web_1.html（最終更新 2007 年 9 月 12 日, アクセス日 2010 年 2 月 10 日）.

[13] 工藤拓. MeCab: Yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net/（最終更新 2009 年 9 月 27 日, アクセス日 2010 年 2 月 10 日）.

[14] Library of Congress. InQuery stopword list for THOMAS. http://thomas.loc.gov/home/stopwords.html（アクセス日 2010 年 2 月 10 日）.

[15] 大向一輝. 学術情報プラットフォームとしての CiNii. カレントアウェアネス, No. 301, pp. 2–4, 2009.