Vol.2010-DBS-150 No.14
Vol.2010-IFAT-99 No.14
2010/8/4

簡易類似文書検索手法「ふわっと関連検索」の予備的評価と分析

高 久 雅 生[†1]　江 草 由 佳[†2]

学術論文との出会いを促すツール「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．本稿では，新聞記事サイトに対する評価実験と検索結果例の分析を示し，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．

Preliminary Analysis and Evaluation of A Simple Similarity Search Tool, "*Fuwatto Search*"

Masao Takaku[†1] and Yuka Egusa[†2]

The authors propose a 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 a document-by-document retrieval capability between any text and a scholarly database. This paper reports a preliminary evaluation results of its effectiveness.

†1 物質・材料研究機構
National Institute for Materials Science
†2 国立教育政策研究所
National Institute for Educational Policy Research

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，鉄道の運行など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)2)]．

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，意識すらしていない市民が，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自動的に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[3)] 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

Vol.2010-DBS-150 No.14
Vol.2010-IFAT-99 No.14
2010/8/4

に提供を開始したサービス J-GLOBAL[4] 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[5]．

「Science and You」[6] 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[7]．

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[8]．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，書籍情報表示画面に関連論文一覧を自動的に提示する．

一方，文書類似度を用いた検索という観点では，高野ら[9] 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」[10] や「IMAGINE・想」[11][12] 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．さらに，丸川ら[13] は，単語空間における類似度を使うことによって文書空間を超えた類似文書検索が実現されるとする，連想検索の役割についても触れられている．

3. 提案手法「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．

(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する．

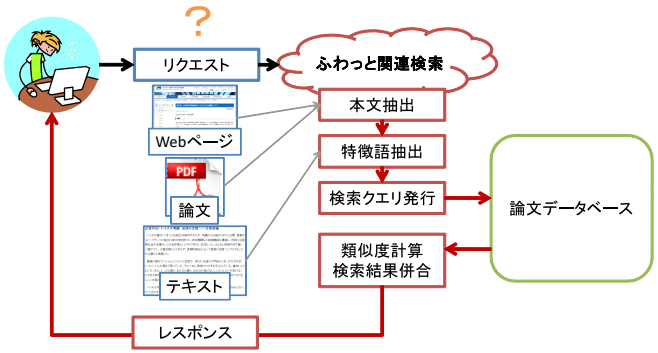

図 1 システムの概要図

(4) 検索結果の提示
前項で得られた検索結果を，詳細度または類似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．

3.1 CiNii API による実装

国立情報学研究所が提供する論文データベース CiNii[3] 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[14] 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[15] 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[16] 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[17] 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．

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，

Vol.2010-DBS-150 No.14
Vol.2010-IFAT-99 No.14
2010/8/4

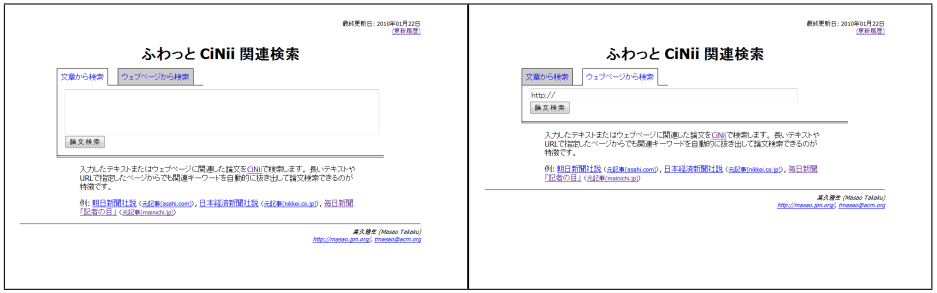

図 2 CiNii API を対象とした論文検索ツール「ふわっと CiNii 関連検索」トップページ（左:「文章から検索」，右:「ウェブページから検索」）

図 3 「ふわっと CiNii 関連検索」の検索結果画面例

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．

4. 評 価 実 験

本節では提案手法の評価実験について述べる．前節で説明した「ふわっと CiNii 関連検索」を対象として，Web 上で提供されている新聞記事内容からの検索結果の評価を行った．

表 1 検索対象文書とした新聞記事（抜粋）

記事 ID	掲載紙	掲載面	記事タイトル	掲載日
NP017	産経	社会	通過駅？熊本、九州新幹線に不安　企業アンケ「プラス」は６割	2010 年 4 月 5 日
NP018	産経	政治	消費税論争勃発　その背景は？　民主執行部のバラマキ路線を牽制　「ポスト鳩山」の思惑も	2010 年 4 月 13 日
NP021	産経	科学	山崎さんが琴を演奏　国際宇宙ステーション	2010 年 4 月 12 日
NP028	産経	国際	「軍事対応」は１割以下　韓国艦沈没で世論調査	2010 年 5 月 7 日
NP029	朝日	国際	ＣＯＰ１５政治合意「尊重」で一致　経済国フォーラム	2010 年 4 月 20 日
NP032	朝日	社会	光が織り成す夜の芸術　徳島でＬＥＤフェスティバル	2010 年 4 月 18 日
NP033	朝日	政治	沖縄知事、県内移設反対の県民大会出席へ　普天間問題	2010 年 4 月 23 日
NP035	朝日	科学	巨大氷山衝突、南極の氷河もぎ取る　ＮＡＳＡ撮影	2010 年 3 月 19 日
NP049	朝日	スポーツ	バイエルンが先勝　リヨンに１－０　欧州ＣＬ準決勝	2010 年 4 月 22 日

4.1 課 題 文 書

評価実験に使用する対象文書は，2010 年 3 月 19 日から 5 月 9 日までの約 2ヶ月間に，2 つの新聞社サイト（産経新聞，朝日新聞）に掲載された新聞記事から無作為に抽出した 34 文書を用いる．**表 1** に，今回対象とした新聞記事の例を示す．

これらの対象文書を検索トピックとみなして，新聞記事に適合する論文を提案手法を用いて検索を行い，その結果について評価を行う．提案手法において，いくつかの関連検索アルゴリズムを試行し，それらの検索結果の主題適合性を人手判定した結果を用いて，評価結果を比較考察する．

4.2 適 合 判 定

提案手法の検索結果中の各文書について，一名の判定者が判定を行った．

判定に際しては，判定者に，以下のカバーストーリーを提示し，そのシナリオに準じた検索タスク遂行中における状況を想起してもらい，判定を行った．

> あなたは，ある *1* つの新聞記事（もしくはブログ記事）の内容やできごとを対象として，レポートを書いて提出することになりました．そこで、レポートに参考文献として出すにふさわしい論文を探しています．

検索要求に合致するかどうかの基準に基づく適合度は，A）適合，B）部分適合，C）不適合の 3 段階とした（**表 2** 参照）．

判定の際には，論文の情報として得られるタイトルや掲載誌などの書誌事項に加え，抄録や，判断に迷う場合などは本文も参照して，判断するように指示した．

判定集合は，4.1 節で述べた新聞課題文書 34 件に対して，後述する複数の提案手法の検

Vol.2010-DBS-150 No.14
Vol.2010-IFAT-99 No.14
2010/8/4

表 2 適合度

適合度	説明
A（適合）	検索要求に適合（そのトピックについて述べており，検索要求を満たす情報がふくまれている）
B（部分適合）	検索要求に部分的に適合（関連することがらに言及しているが，それだけでは情報要求のすべてを満たさない．他に材料がない場合には他のものと組み合わせて材料とすることができたり，傍証となる．間接的には役に立つ．情報要求のある一面は満たす）
C（不適合）	不適合（まったく的外れな文献 ）

索結果から上位 50 件までを使用した．判定に際しては，順序による影響を考慮して，結果の提示は論文タイトルの順に並べ替え，検索結果順位とは切り離して判定を行った．各課題文書平均 647 件の論文情報を適合判定した．

4.3 評価対象アルゴリズム

評価実験には，提案手法のいくつかの派生版を加えて，比較評価した．対象とした特徴語抽出およびランキング手法は以下の通りである．

(1) 特徴語抽出と重み付け

- TF: 特徴語の重みとして，文書内での単語頻度を用いる．
$$weight(t) = TF(t)$$
- LogCost: 特徴語の重みとして，MeCab による生起コスト値の対数化し，文書内の出現位置ごとに足しあわせた値を用いる．
$$weight(t) = \sum_i log_2(Cost(t_i))$$
- IDF:算出された特徴語の重みを，対象データベースにおけるヒット件数により補正する．元の重み値に，ヒット件数を対数化して除算した値をその特徴語の重みとする．
$$weight'(t) = weight(t) \times log_2(DF(t))$$

(2) 検索クエリ発行とランキング

- AND: 特徴語の重みの降順に上位 n 件を AND 条件で連結し，クエリとして発行する．m 件以上のヒットがない場合は，最終的に m 件以上の検索結果が得られるまで，$n-1, n-2, \ldots$ とクエリ実行する語数を減らしていき，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る．
- Comb: 特徴語の重みの降順に上位 n 件から，3 語づつの全ての組み合わせを抽出し，それらの AND 条件を，クエリとして全て発行する．
- Rerank: 検索結果に含まれる各結果文書の単語ベクトルと元の特徴語ベクトルとをコサイン類似度で比較し，類似度の高い順に並べ替える．

(3) 擬似適合フィードバック

- PRF(α): 適合順に並べた文書リストの上位 k 件を取り出して正規化し，1 つのベクトルとみなして，元のベクトルに併合する．
$$weight'(t) = (1-\alpha)weight(t) + \alpha \cdot weight_d(t)$$

なお，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 $m = 100$ を設定した．また，擬似適合フィードバックに用いる文書数として $k = 20$ を用いた．

4.4 評価指標

各手法の検索結果の評価は，上位 10 位における精度（Prec@10）と，平均精度（MAP: Mean Average Precision）とによって行う．

なお本稿では，多段適合のうち，A 判定および B 判定の両方を適合とみなして，精度を計算した．

5. 結果と考察

5.1 特徴語抽出，重み付け，ランキング手法

表 3 に，手法ごとの評価結果を示す．表 3 において，課題文書とした新聞記事群の全体と，科学面掲載記事とその他の記事とを区別した場合の結果も記載してある．

全体で提案手法を比較した場合，Comb + Rerank ベースの手法は，AND ベースの方式と比べ，Prec@10 で最大 0.11 以上の向上を果たしており，つまり上位 10 件以内で適合文書が 1 件以上多く出現するようになっており，良い性能を挙げていることが分かる．これは，Comb 方式により，多くの適合候補文書をデータベースから取得できていることに加え，リランキングによって，単語ベクトルを通じた類似文書を上位に取得できるようになっていることによると思われる．AND 方式の場合は，特徴語スコアの最上位に位置する語に，検索性能がおおきく依存してしまうが，Comb 方式はこの欠点を改善しているものと考えられる．

特徴語の重み付けに用いた IDF と，TF および LogCost の重み付け手法のあいだでは，さほど大きな差が出ていない．IDF に関して見ると，Prec@10 でおおむね IDF を用いたほうがやや性能が高い結果が出ているものの，MAP での結果ではさほど変わらないか，逆に性能がやや落ちる結果となった．この性能の違いや類似については，十分に精査できておらず，これらの特徴語重み付け手法による影響の精査は今後の課題となっている．

Vol.2010-DBS-150 No.14
Vol.2010-IFAT-99 No.14
2010/8/4

表 3 評価結果

手法	Prec@10			MAP		
	全体	科学	その他	全体	科学	その他
AND + TF	0.0794	0.1105	0.0400	0.0638	0.0559	0.0739
AND + TF + IDF	0.0941	0.0789	0.1133	0.0577	0.0203	0.1101
AND + LogCost	0.1059	0.1158	0.0933	0.0600	0.0500	0.0809
AND + LogCost + IDF	0.0765	0.0737	0.0800	0.0577	0.0202	0.1052
AND + Rerank + TF	0.1206	0.1474	0.0867	0.0459	0.0668	0.0195
AND + Rerank + TF + IDF	0.1382	0.1421	0.1333	0.0462	0.0428	0.0504
AND + Rerank + LogCost	0.1000	0.0947	0.1067	0.0381	0.0434	0.0313
AND + Rerank + LogCost + IDF	0.1324	0.1579	0.1000	0.0437	0.0460	0.0407
Comb + Rerank + TF	0.2176	0.2579	0.1667	0.1669	0.2066	0.1165
Comb + Rerank + TF + IDF	0.2324	0.2737	0.1800	0.1668	0.2149	0.1059
Comb + Rerank + LogCost	0.2324	0.2526	0.2067	0.1679	0.2011	0.1259
Comb + Rerank + LogCost + IDF	0.2500	0.2789	0.2133	0.1698	0.2140	0.1138
PRF(25) + Comb + Rerank + LogCost + IDF	0.2353	0.2632	0.2000	0.0315	0.0411	0.0898
PRF(50) + Comb + Rerank + LogCost + IDF	0.1971	0.2263	0.1600	0.1032	0.1322	0.0666
PRF(75) + Comb + Rerank + LogCost + IDF	0.1059	0.1263	0.0800	0.0527	0.0713	0.0291
PRF(100) + Comb + Rerank + LogCost + IDF	0.0618	0.0789	0.0533	0.0170	0.0374	0.0196
PRF(200) + Comb + Rerank + LogCost + IDF	0.0059	0.0105	0.0000	0.0050	0.0084	0.0007

一方，PRF に関しては，フィードバックを行って特徴語ベクトルを修正した結果のほうが，元の Comb + Rerank 手法よりも性能が落ちることとなった．このような性能低下を生じた原因は調査できておらず，今後の課題である．

5.2 文書ジャンル

表 3 において，全課題文書群 34 件のうち，科学面掲載記事は 19 件，科学面以外に掲載された記事は 15 件であった．

これらの対象文書群のジャンルの違いは，性能にも影響を与えている．ほぼ全ての手法において，科学記事に対する検索結果の方が，その他の記事に対するものよりもより良い性能を挙げている．これは，検索元の文書と，対象データベースである論文に使われる用語が比較的近い領域の用語を用いていることにより，特徴語を通じた検索と文書取得に有利に働いていることによるものと思われる．

検索される文書がどのようなジャンルに由来しているかによって，検索性能に影響する場合，関連検索をどのようなシーンで用いるかによって，検索手法そのものを切り替えたりするような適応手法も考えられる．今後の課題として検討したい．

5.3 実行時間

提案手法は，Web API などのマッシュアップを想定した手法であり，Web ページコンテンツや自然文のクエリを受け取ってから，ネットワークを通じてリアルタイムに検索する実行モデルを想定している．このため，リアルタイムの応答性も要求される．手法によって，抽出語数や検索クエリ発行回数に違いがあるため，応答時間に影響する．**表 4** に，表 1 の各文書に対して，AND 手法を用いた場合と，Comb 手法を用いた場合それぞれの実行時間を示す．なお，実行時間は 5 回試行の平均を示す．表中，AND は「AND + LogCost + IDF」を，Comb は「Comb + Rerank + LogCost + IDF」をそれぞれあらわす．

表 4 実行時間

文書 ID	実行時間（秒）	
	AND	Comb
NP017	2.26	29.26
NP018	2.28	19.43
NP021	2.35	10.82
NP028	2.98	15.64
NP029	2.47	13.74
NP032	2.12	10.77
NP033	2.57	10.50
NP035	2.64	14.44
NP049	2.32	11.61

AND 手法は 2〜3 秒，Comb 手法は 10〜30 秒程度，実行に要した．これは，AND 手法の HTTP アクセスは平均 20 回であるが，Comb 手法は平均 131 回の HTTP アクセスを行っており，この差がおおむね全体の実行時間に反映されていた．

5.1 節で示したとおり，Comb 手法は AND 手法に比べて圧倒的な適合性能を示しているが，表 4 に示されるとおり，アクセス回数が増えることによって，その実行時間が増えるというトレードオフの関係にあることが分かる．Comb 方式の場合では一回の検索あたり，長い場合には 30 秒近くかかってしまうことから，リアルタイム応答性能が重視される利用シーンで採用することは難しい．そういった場合，利用者が直接に検索要求を渡すシーンではなく，たとえば，Google AdSense のように Web コンテンツ内の一部として埋めこむなど，間接的な利用を行うシーンで，応答速度が重要とならないケースでは，Comb 手法を採用するなど，使い分ける必要があるものと考える．

6. 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評価実験を通じて，その有用性を示した．

Vol.2010-DBS-150 No.14
Vol.2010-IFAT-99 No.14
2010/8/4

評価実験の結果，新聞記事に対しては上位 10 件までの結果に対して精度 0.25 を示し，平均精度（MAP）でも 0.17 の性能を示した．

今後は，これらの分析結果を踏まえた，特徴語抽出および類似度計算手法のさらなる改良や，他の文書種別に対する分析などを通じて，よりよい関連検索手法の完成を目指す．

謝辞 本研究の一部は，科学研究費補助金若手研究（B）（課題番号: 20700228）の助成による．

参 考 文 献

1) 文部科学省：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) 野村一夫：インターネット市民スタイル: 知的作法編，論創社 (1997).
3) 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）.
4) 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）.
5) 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).
6) 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）.
7) 石村源生：市民の日常的関心と科学技術コンテンツを結びつける Web サービス「Science and You」の開発と運用，PC カンファレンス 2009，p.4p. (2009).
8) 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).
9) 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).
10) 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）.
11) 連想出版：想 — IMAGINE Book Search. http://imagine.bookmap.info （アクセス日 2010 年 2 月 10 日）.
12) 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).
13) 丸川雄三，阿辺川武：横断的連想検索サービス「想ー IMAGINE」: データベース連携が拓く新たな可能性，情報管理， Vol.53, No.4, pp.198–204 (2010).
14) 国立情報学研究所：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）.
15) 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）.
16) 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）.
17) 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）.