

論文情報ナビゲータの構築

大山 敬三, 相澤 彰子, 後藤田 洋伸, 小陳 左和子, 大綱 浩一¹

{oyama,akiko,gotoda,kojin,otsuna}@nii.ac.jp

国立情報学研究所
(大学共同利用機関法人 情報・システム研究機構)

本報告では国立情報学研究所で構築を進めている論文情報ナビゲータのシステムについて、機能及び構成とコンテンツ形成の概略について述べる。本システムは学術論文情報のポータルとして、国立情報学研究所における今後の学術情報サービスの核となるものであり、研究所および協力機関が構築している各種の論文情報資源を効果的に利用可能とすることを目的としている。独立に重複して作成された大規模な書誌情報を統合し、情報源へのアクセスを可能とするとともに、引用文献を同定しリンクすることにより検索や探索を可能とする。本システムは単に情報サービスを提供するだけでなく、様々な情報学研究のプラットフォームとして活用されることを目指している。

Development of Scholarly Article Information Navigator

OYAMA, Keizo; AIZAWA, Akiko; GOTODA, Hironobu;
KOJIN, Sawako; OTSUNA, Koichi

National Institute of Informatics
Research Organization of Information and Systems

The National Institute of Informatics is now developing a Scholarly Article Information Navigator. In this paper, the authors report on the outline of functions and structure of the system and of compilation of its contents. This system will be an essential component of the scholarly information services at the Institute, aiming at effective utilization of various article information resources compiled by the Institute and by cooperating organizations. Access to the original resources is enabled by integrating the large scale bibliographic databases created independently, and search for citing articles is enabled by linking referenced citations. It is also expected to be used as a platform of various informatics researches.

¹ 現在, 京都大学附属図書館。

1. はじめに

情報技術の発達と共に、学術情報の発信形態も急激に変化しつつある。その結果、論文情報について以下のような変化が生じている。

- 情報発信源が情報生産の上流に移行しつつある。

DB プロバイダ

→DB プロデューサ

→出版社

→著者

- 情報内容が二次情報から全文情報+メタデータに移行しつつある。

現在は、これらの変化の途中であり、情報源が様々なレベルで混在している状況である。

論文情報のエンドユーザである研究者にとって見ると、情報源が多様化し、必要とするより多くの情報を電子的な方法によって入手できるようになり、研究の効率化が図れる潜在的な状況ができてきている。

しかし一方で、本来同一の情報が異なる場所で異なる条件の下に供給されていたり、個々の情報源の規模が小さくなる傾向にあったりするため、必要とする情報が入手可能な場所にあるにもかかわらず、その所在を見いだすのが困難であったり非効率であったりすることも少なくない。

このような状況を改善し、利用可能な論文情報に対して統一かつ効率的なアクセス方法を提供することにより学術論文ポータルの実現を目指し、国立情報学研究所(NII)では論文情報ナビゲータ(現:引用文献情報ナビゲータ CiNii[1])の構築を進めている。

類似の概念や機能を持つシステムとしては、米国 ISI 社の“Web of Science”[2]や NEC 米国研究所・ペンシルバニア州立大学情報科学研究科の“CiteSeer.IST”[3]などがある。

Web of Science は ISI 社が長年にわたって構築し評価を得てきた大規模な引用索引データベースである SCI, SSCI, HCI などを、引用情報をキーとして統合しつつ外部のリソースへのリンクを付加して利用価値を高めたサービスシステムである。

CiteSeer.IST は NEC 米国研究所が構築してきた CiteSeer をベースにしており、インターネット上に散在する科学論文情報を収集してメタデータと引用情報の抽出を行い、リンクを付加すると共にオリジナル情報へのリンクを提供するシステムであるが、様々な実験的機能も提供されている。

NII の論文情報ナビゲータは現状では Web of Science に近いサービスを実現しようとするものであるが、以下のような点において実現上の特徴がある。

- 英文だけでなく日本語論文も同様に扱う。
- 形式が確立した主要な論文誌だけでなく、研究会資料なども同様に扱う。
- 各種機関によって独立に作成された論文情報に基づいて構築する。
- 複数の大規模な論文情報データベースを統合する。

本システムは一方で、現在 NII が個別に提供している論文情報系のサービスを統合し、統一的なアクセスの窓口を提供するものとしての位置づけもある。

また、将来的には NII が有する、あるいは提携機関から提供を受けた、論文情報以外の学術情報として、以下のような情報との連携も計画している。

- 図書・雑誌の目録・所在情報
- 研究者・著者情報
- 研究プロジェクト情報

一方、本システムは、引用の可視化や分析、情報の同定や統合、効率的な索引方式など、研究のためのプラットフォームとなると共に、研究成果をいち早くサービスシステムに反映する実験環境にもなる。

基幹部分が完成した後は、共同研究などを通じて研究プラットフォームとしても研究者に提供していくことを考えている。

2. 論文情報ナビゲータの構成

本システムの構成の概略を図 1 に示す。主にユーザに対して直接的に情報の提供を行う「論文情

報ナビゲータ」本体と、統合書誌データベースを維持・管理し同一の書誌・引用間の同定・統合の処理を行う「書誌・引用同定・統合システム」、それに本文情報などを提供するNIIおよび外部の各種サービスからなる。

書誌・引用同定・統合システムは、NII あるいは外部の提携機関から情報源の書誌および引用情報を一括あるいは差分で受け取り(受け渡し方は相手により様々である)、統合書誌データベースにロードする。次に、それらを統合書誌データベース中の全レコードに対して同定・統合処理を行う(詳細は 3.3 節で述べる)。その後、差分データを統合書誌データベースから抽出して検索・表示

用データを更新し、インデクスを更新する。

表示用データには、情報源中の情報資源を指示する識別子を持っており、論文情報ナビゲータで書誌情報を表示する際に、これをリンクとして表示する。ユーザはこのリンクを選択することにより複数の異なる情報資源に対するナビゲーション機能を利用することができる。

3. 提供情報内容の構成

本節では、本システムで統合して提供する論文情報の情報源と、それらを統合するための方法について記す。

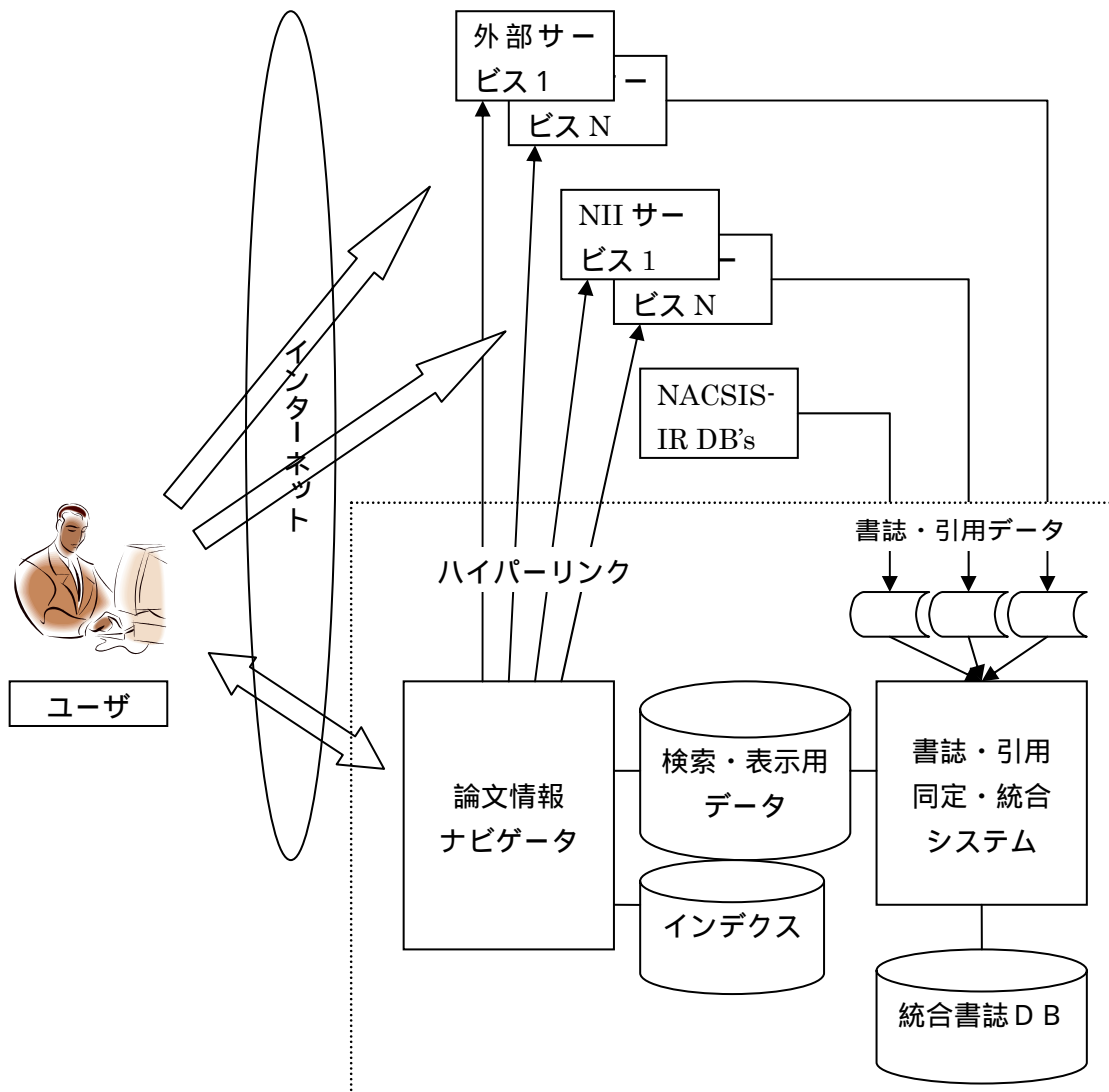


図 1 . 論文情報ナビゲータの構成概要

3.1. 国立情報学研究所の情報源

本システムで統合する国立情報学研究所の情報源には以下のようなものが含まれる。

- 情報検索サービス NACSIS-IR[3]の「引用索引データベース」(CJP) (書誌, 抄録, 引用)
- 電子図書館サービス NACSIS-ELS[4] (ELS) (書誌, 抄録)
- 電子ジャーナルリポジトリ NII-REO[5] (REO) (書誌, 抄録)
- 研究紀要ポータル[6] (KIYO) (旧, 学術雑誌目次速報) (書誌, 抄録)

そのほか, NACSIS-IR で提供している論文系のデータベースについても, 条件が整ったものから統合する。

これらのうち, NACSIS-IR[7,8]で提供しているCJP などは, 元々, 書誌・引用などの二次情報だけからなるものであるため, 引用情報ナビゲータに統合することにより全サービスを移行する。

一方, NACSIS-ELS[9]は本文画像情報を含むサービスであるため, 引用情報ナビゲータに書誌・抄録情報を統合して検索機能を移行することにより, 学術論文のアーカイブという位置づけを明確にし, 本文情報提供機能のみを持つものとなる。

また, NII-REO は電子ジャーナルリポジトリとして構築されたものであり, この位置づけは変更なく, 現在のシステムにおいて引き続き収録雑誌内での検索機能と本文情報の提供を行う。

研究紀要ポータルが提供する論文情報には, 書誌情報だけのもの, 本文情報を持つもの, および, 大学等にある本文情報へのリンクを持つものがある。本文情報を持つものについてはその提供機能をNACSIS-ELS に統合する。

3.2. 外部の情報源

本システムで統合する外部の情報源には以下のようなものがある。

- 国立国会図書館の「雑誌記事索引」[10] (書誌)。現在は個別の記事情報へ直接リンクすることはできないので, 当面は NDL

OPAC[11]へのリンクを表示する。

- 専門分野の論文提供サービス(書誌)。複数について交渉中。
- 学会等が提供する電子ジャーナルサイト(書誌, 抄録, 引用)。

また, 当面は引用文献の同定・統合の目的に限定して, 複数の大規模文献データベースの提供を提携機関から受ける。これらについても将来の統合提供を指向して提供機関との協議を行っている。

3.3. レコード統合方法

前 2 節に記した情報源を統合するにあたり, 同一論文に対応する書誌レコードが複数の情報源に含まれるため, これらを同定して統合する必要がある。

ここで, 「同定」とは, 任意の 2 つの書誌レコードが同一の論文に対応するものであることを評価・確定することであり, 「統合」とは, 同定された書誌レコードをデータベース上で一体化することである。

引用データに関しても, 各インスタンスを個別の書誌データであると見なして, それぞれから書誌レコードを作成する。これを他の書誌レコードと同様に同定・統合することにより, 全体として引用関係で対応づけられた統合書誌データベースが構築できる。

同定においては特に以下の点に注意する必要がある。

- 情報源ごとの書誌単位の差異。例えば, 一つの論文が複数の雑誌に並行して掲載されている場合や, 特集号の巻頭言などの扱い。
- 元となる情報自体の誤り。特に引用情報には著者によると思われる記入誤りが頻出する。
- 情報源作成時のデータの誤り。印刷物を元に入力された情報源で発生するが, 誤りの頻度は情報源によって大きく異なる。
- 誤同定の影響。本来同定すべきでない書誌を同定・統合し, そこに引用リンクが付加されてしまうと, 誤りの訂正に伴う引用リンクの付け

替えに大きなコストを要する。

以上のような点を考慮し、同定は以下の段階を経て行う。なお、(1)～(3)は計算機処理による一括処理、(4)は計算機を用いた対話処理として行う。

- (1) 新規にロードされた書誌レコード(書誌情報, 引用情報の両方に対応する)に対し, 再現率重視で効率的に同定対象候補を収集する。
- (2) 同定対象候補のそれぞれについて精度重視で同一性の判定を行う。
- (3) 同一性の確信度や候補数に応じて, 自動同定, 人手チェック, 非同定に振り分ける。
- (4) 人手チェックにより同定, 非同定を確定する。

ここで, 誤同定の可能性を極力低減するため, (3)においては精度優先で振り分けを行う。しかし, 非同定誤り(漏れ)も一定程度に抑える必要があるため, 人手チェックに振り分けられる候補数が多くなる傾向がある。この結果, 同定において時間的・経費的にもっともコストがかかるのは(4)の段階になる。

そこで, 人手チェックの対象候補数を減らすように各種の手法を組み合わせ(2), (3)の処理を行うと共に, 方式を工夫して(4)のチェック自体の効率を高める必要がある。

このために現在は, ヒューリスティクスと統計的手法に基づいて雑誌, 巻号, 論文の階層関係を利用する効率的な同定方式[12]と, 情報検索と機械学習を用いて個別論文を直接扱う漏れの少ない同定方式を併用している。これらの詳細については別の機会に報告する予定である。

なお, 引用対引用の同

定に関しては人手チェックにおいて高い精度を確保することが難しいため, 完全一致に近い機械的な同定のみを行う。

4. ユーザインタフェース

論文情報ナビゲータの利用者インタフェースの概要を以下に示す。

ユーザは通常, まず閲覧を希望する論文を探す必要があるが, これには通常の電子ジャーナルシステムなどと同様に以下のような手段を利用できる。

- (1) キーワード等を用いた簡易検索
- (2) 特定のデータ項目中出现するキーワードや分野, 出版年などを組み合わせた詳細検索
- (3) 出版者, 雑誌名, 巻号などを順に指定するディレクトリ探索

図 2 に本サービスの試行版として公開中の CiNii のトップページ(URL: <http://ci.nii.ac.jp/>)を示す。上記(1)と(2)の両方がこのウィンドウから利



図 2 . 論文情報ナビゲータトップページ

用可能である。検索結果は出版年の昇順・降順、著者名順、雑誌名準、および引用数順に表示することができる。

なお、(3)は現在開発中であり、実装時には左側のメニューから選択可能にする予定である。

図 3 にキーワード「NACSIS-IR」で簡易検索を実行した結果の論文一覧ページを示す。著者名、論文名、雑誌名、巻号ページの他に、収録データベースのアイコン、書誌詳細表示へのリンク、引用・被引用件数が表示されている。収録データベースのアイコン(図中では「ELS」)はオリジナルの情報サービスへのアンカーとなっている(グレイアウトされたアイコン(図中では「CJP 書誌」)などは、情報は統合済みであるが対応する情報サービスがないことを示す)。

論文名を選択すると図 4 に示すような書誌詳細ページが表示される。書誌詳細情報の他に、利用可能であれば抄録、引用文献リスト、および被引用文献リストを、タブから選択して表示することができる。

後述する可視化などの表示機能も、タブを追加することにより本ユーザインタフェースに組み込む予定である。



図 3 . 検索結果表示

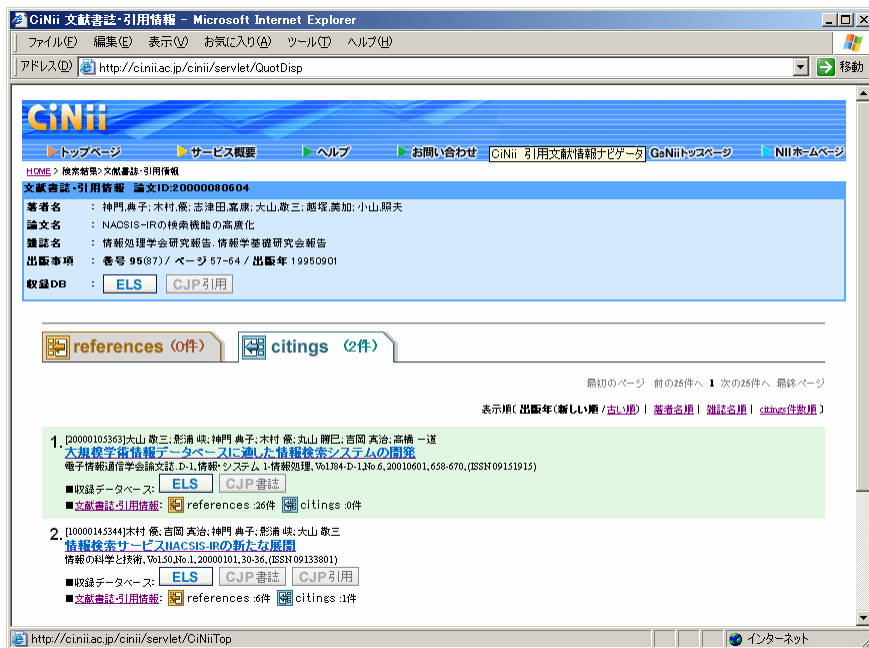


図 4 . 書誌詳細表示

5. 論文情報ナビゲータの拡張と活用

本システムでは、継続的に機能の改善とコンテンツの充実を図ってゆく。

特にコンテンツについては、学会、大学、出版

社、学術情報提供機関などとの多角的な連携を図ることにより、できるだけ広範な学術領域と多様な出版形態の論文情報資源へのナビゲーションを提供し、研究の効率化と深化に貢献できるようにする。

一方、本システムは単に利用者に論文情報を提供するだけでなく、情報学研究との相互的な関係において、機能の拡張・活用を目指している。

具体的には、以下のような課題を掲げて取り組んでいる。

- (1) 引用関係や引用統計の可視化
- (2) 同定の確信度に基づく検索・表示
- (3) ユーザのデスクトップツールとの連携
- (4) 大量文書画像からの引用情報の抽出と書誌同定[13]

また、今後取り組むべき課題として以下のような事項を考えている。

- (5) 論文、図書、著者、プロジェクトなどの異種情報資源間の同定とシステム間連携
- (6) Web上の論文・著者情報等の発見・同定・リンク
- (7) 論文の引用や共著などに基づく研究動向分析

これらについては、システムへの実装や論文という形で公開していく予定である。

さらに、国立情報学研究所は大学共同利用機関として、情報学分野における共同研究を推進しており、本システムの活用のための研究テーマや、本システムの高度化につながる研究テーマ等の提案も歓迎する。

6. おわりに

本報告では、現在構築中の論文情報ナビゲータについて、システムの構成とコンテンツの構築についての概念を述べ、ユーザインタフェースの概要を示した。また、研究プラットフォームとしての今後の拡張と活用についての考え方を記した。

現在、参考文献[1]で公開されているシステムは、本稿で報告したシステムの前身となる実験的なシステムであり、今後、機能上の拡張とコンテン

ツの拡充を段階的に経て、2005年4月から正式なサービスとして提供してゆく予定である。

読者の方々には、ユーザとして、また情報学研究者として、様々な視点からのご提案をいただきたい。ユーザ、サービス、研究の三者による相乗効果により、本システムに関わる全ての者がより多くの利益を受けられるよう努力する所存であるので、ご支援を賜りたい。

参考文献

- [1] 国立情報学研究所: “CiNii Home (NII 引用文献情報ナビゲータ)”, <available from <http://ci.nii.ac.jp/>>, (accessed: 2004-04-15).
- [2] ISI: “Web of Science”, <available from <http://www.isinet.com/products/citation/works/>>, (accessed: 2004-04-15).
- [3] School of Information Sciences and Technology at Penn State University: “Computer and Information Science Papers CiteSeer Publications ResearchIndex”, <available from <http://citeseer.ist.psu.edu/cis/>>, (accessed: 2004-04-15).
- [4] 国立情報学研究所: “NACSIS-IR homepage (J)”, <available from <http://www.nii.ac.jp/ir/ir-j.html>>, (accessed: 2004-04-15).
- [5] 国立情報学研究所: “NACSIS-ELS(J)”, <available from <https://els.nii.ac.jp/>>, (accessed: 2004-04-15).
- [6] 国立情報学研究所: “NII-REO (NII 電子ジャーナルリポジトリ)”, <available from <https://reo.nii.ac.jp/>>, (accessed: 2004-04-15).
- [7] 木村 優, 吉岡 真治, 神門 典子, 影浦 峡, 大山 敬三: “情報検索サービス NACSIS-IR の新たな展開”, 情報の科学と技術, Vol.50, No.1, p.30-36, 2000.
- [8] 大山 敬三, 影浦 峡, 神門 典子, 木村 優,

- 丸山 勝巳, 吉岡 真治, 高橋 一道: “大規模学術情報データベースに適した情報検索システムの開発”, 信学論 D-I, Vol.J84-D-I, No.6, p.658-670, 2001.
- [9] 安達 淳: “メタデータを中心に構成した文書画像の電子図書館システム”, 信学論 D-I, Vol.J84-D-1, No.6, 768-776 2001.
- [10] 国立国会図書館: “国立国会図書館:資料の検索”, <available from <http://www.ndl.go.jp/jp/data/opac.html>>, (accessed: 2004-04-15).
- [11] 国立国会図書館: “国立国会図書館 NDL-OPAC”, <available from <http://opac.ndl.go.jp/>>, (accessed: 2004-04-15).
- [12] 相澤 彰子, 大山 敬三, 高須 淳宏, 安達 淳: “複数書誌データベース統合における重複エントリーの高速検出法”, 情処研報, (2004-FI-75 (14)), 2004-5-13 (掲載予定).
- [13] 岡田 崇, 高須 淳宏, 安達 淳: “SVM/HMMによる引用文献データの同定”, 情処研報, Vol.2004, No.36, p.79-86, (2004-FI-74 (11)), 2004-3-26.