検索エンジンを部品とするエージェントの構成について

中藤, 哲也
九州大学情報基盤センター

大森, 敬介
九州大学大学院システム情報科学府

廣川, 佐千男
九州大学情報基盤センター

出版情報：Proceeding of JAWS2004, 2004
バージョン：
権利関係：

検索エンジンを部品とするエージェントの構成について

中藤　哲也[†a)]　　大森　敬介[††]　　廣川佐千男[†]

Search Engines as Components of Constructing an Agent

Tetsuya NAKATOH[†a)], Keisuke OHMORI[††], and Sachio HIROKAWA[†]

Abstract.
検索エンジンの機能は非常に単純だが，最も広く利用されている実用的エージェントといえる．特定の対象に特化した検索エンジンを利用する場合，同じ検索エンジンを連続して反復的に利用することが多い．例えば，文献検索システムでは，検索結果の中の共著者や論文タイトル，あるいは重要語を新たなキーワードとして更に検索を続けることが多い．しかし検索サイトは，ブラウザ経由での利用しか想定されていないので，このような操作の自動化や，アプリケーションからの直接利用はできない．あるいは，ある検索サイトの結果から得られたキーワードを使って別の検索サイトで検索を行うには，キーワードをコピーし別に開いた検索サイトの画面で入力しなければならない．我々は，検索結果の HTML ページから検索項目を抽出するラッパーを使うことで，Web 上で公開される専門的検索サイトを部品として結合し，新たな検索機能を構成するための枠組を提案する．本発表では，文献検索についてのプロトタイプを通じてシステム概略を述べる．

Keywords. 検索エンジン，ラッパー

1. ま え が き

Google などの一般検索サイトや特定の情報に特化した専門検索サイトを使って特定のテーマについての調査を行う場合，一つのキーワードによるただ１度だけの検索で作業が終ることは稀である．例えば，ある地域のレストランリストを検索し，次に各レストランのメニューや価格に関する情報を集める，あるいは販売中の中古車の一覧を検索し，そのうちの幾つかの車の詳細情報を集めて比較する場合のように，１度の検索だけで終らないことが多い．得られた数十件の検索結果を見ることで，重要な人名や関連するキーワードを学習し，より広範な検索やより精度の高い検索を行う．適切なメモを残しながらこのような操作を繰り返し，納得できる検索結果のリストを構築する．それらの検索は同一の検索サイトで引き続き行う事もあれば，異なる検索サイトで行う事もある．

このように検索エンジンを反復的に利用する場合でも，一般の検索エンジンの利用と専門検索エンジンの利用では状況が大きく異なる．一般の検索エンジンでは，検索結果は多様なページなので再検索のための新たなキーワードを獲得する一般的な手立てはない．一方，専門検索サイトの検索結果は，そのサイトの背景にあるデータベースの同質データであることが多い．例えば，文献検索のサイトでは，人名やキーワードを与えて得られる検索結果は 単なる Web のページではなく，著者，タイトル，雑誌名，ページ，出版年などの項目からなる文献データである．網羅的に文献検索を行う時には，１回目の検索結果が得られてもそれで終りでなく，そこで得られる情報をもとに更に検索を続けることが多い．一つの論文を見つけると

- 著者や共著者が他にどのような論文を書いているか
- その論文はどのような論文を引用しているか
- その論文がその後，どのように引用されているか
- 関連研究で重要なキーワードはなにか
- 著者らのホームページはどこか
- 関連するプロジェクトがあるか

などを繰り返し調べることが多い．つまり，専門検索サイトを使って反復的に検索を行う時には，文献データという構造情報から著者やタイトルという部分的情報を抽出して利用することが多い．DBLP [16] や CiteSeer [14] のような文献検索のサイトでは，文献

† 九州大学 情報基盤センター，〒 812-8581 福岡市東区箱崎 6-10-1
†† 九州大学大学院 システム情報科学府，〒 812-8581 福岡市東区箱崎 6-10-1
a) E-mail: nakatoh@cc.kyushu-u.ac.jp

リストを検索結果として返すだけでなく，このようなユーザーの操作を先取りし，著者ごとに分類したページやそのようなページを動的に生成する URL へのリンクが提示されるので，効率よく関連研究の調査を行うことができる．

一つの専門検索サイトを反復的に利用できるのは，出力情報データの属性に入力として使える情報があることによる．複数の専門検索サイトの統合検索（メタサーチ）が考えられるのは，それらの入力データと出力データの構造が類似しているからである．ある検索サイトの出力データの属性として人名が含まれれば，人名を検索キーワードとする他の検索サイトの入力と結合して利用することが考えられる．

本発表では，このように専門検索サイトを入力データ構造と出力データ構造で規定される抽象的部品として捉え，それらを結合することにより新たな検索エージェントを構成する方法を提案する．本方式の実現可能性を検討するため，情報処理学会，電子情報通信学会，人工知能学会，及び日本ソフトウェア科学会の各学会の論文検索システムを対象とするプロトタイプを実装した．

2. 専門検索サイトの部品化とその効果

一般に専門検索サイトは，内部データベースが持つ情報を扱うため，データベースの持つフィールドの一部を入力をして受け取り，それに一致するレコードのリストを出力するものが多い．例えば，図 1 に示す図書検索サイトでは，入力フィールドとして書籍に関する複数のフィールドがあり，それらの一部を指定する事で，一致する書籍データの一覧をユーザに提示する（図 2）．

この例にあるように，多くの検索サイトの機能は，入力項目と出力項目のペアとして表現できる（図 3）．更に各検索サイト専用のラッパーを組み合わせる事により，他と組み合わせ可能な部品としての基本機能を持たせることができる（図 4）．我々はこれまでに 2,880 件の専門検索サイトを収集している [10]．それらのうち，幾つかの検索サイトについての入出力項目をまとめたものを，例として表 1 に示す．このように，Web 上には非常に多くの多彩な専門検索サイトが存在し，それらの組み合わせによってより高機能な検索が期待される．

検索サイトを入力と出力の組として捉えると，(1) 入力の統合，(2) 出力の統合，(3) 入力と出力の結合，の

図 1　専門検索サイトの例

図 2　専門検索サイトの検索結果の例

図 3　検索サイト機能の模式図

3 通りの組み合わせ方法がある．従来のメタサーチエンジンは (1) の入力の統合だけを実装したものであり，各々の検索結果は単純に（あるいはランキング付で）リスト化されているだけである．(2) の出力の統合では出力結果の各フィールドの意味を使って，例えば価格の比較が可能な，一覧表として検索結果が表示できる。(3) の例としては，求人情報検索で得られた企業について，その企業の業績や動向を株価の検索で調べ

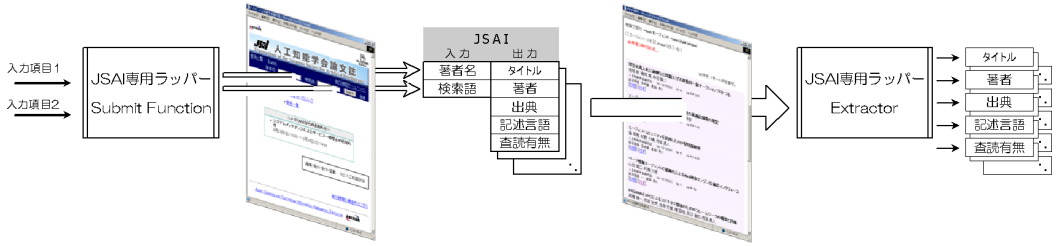

図 4　検索サイトの部品化

表 1　専門検索サイトの入出力の例

サービス名	入力項目	出力項目
図書検索	タイトル，著者，出版者	タイトル，著者，出版者，ISBN，発行年
施設検索	施設名	施設名称, 施設分類 住所, TEL, 管理
塾 予備校検索	塾・予備校名，駅名，出版者	学校名, 教室名 指導形態, 対象, 沿線駅名
野菜生理障害事例検索	キーワード	病名
特許検索	検索語句	出願番号, 発明の名称
最寄りのお風呂屋さん	最寄り駅	浴場名, 路線・最寄り駅名, 道順, 住所, 営業時間 定休日, 特徴, URL

るような場合がある．福田ら [2] は，入力情報，及び出力情報についてのオントロジーを使ったサービス連携の自動化方式を提案しているが，Web サービスの代わりとして検索サイトを仮に利用している．

また，検索サイトの部品化により，部品の結合に関して一般的プログラミングを考える事が可能となる．例えば,「二つの図書館に対する検索を実現する」ためのスクリプトを書くことを考える．従来提案されているラッパーシステム，例えば MetaCommander では図書館 A，図書館 B それぞれについて，CGI にどのようにパラメータを渡すか，出力の HTML から本の情報をどのように抽出するかを，プログラム中に直接埋め込む必要がある．一方，本稿で提案する方式ではこの問題を，

(a)　図書館 A，図書館 B を入出力データ構造の組として捉え，

(b)　二つの図書館情報検索機能の結合方式を記述し，

(c)　統合したシステムの入出力のインターフェースをマッチングさせる

という 3 つの部分に分離して解決する．(a) のためには，それぞれの図書検索サイトについて個別にラッパーを構築する必要がある．しかし，対象とする図書館が変わったとして，(b), (c) の部分は変更する必要はない．(b), (c) の部分は「複数の Web サービスをどのように組み合わせるか」という一般的なプログラムとして，より抽象的に構成することができる．

3.　論文検索システム

本節では，検索サイトを部品とし，それらを組み合わせる事で，情報収集を行う検索エージェントのプロトタイプについて説明する．本プロトタイプは，次に示す各学会の論文検索サイトを対象に，論文の情報を収集する事を目的としている．

- 情報処理学会電子図書館(注1)
- 電子情報通信学会 和文論文誌(注2)
- 電子情報通信学会 英文論文誌(注3)
- 人工知能学会論文誌(注4)
- 日本ソフトウェア科学会 J-STAGE(注5)

本システムは主に三つの機能から成り立っている．それらは，(1) 複数の検索サイトに対して同時に検索を行い，結果を統合してユーザに提示する機能，(2) 結果中の著者名を抽出し，リスティングする機能，(3) リスティングされた著者名をキーとした次のステップの検索を提供する機能，である．

(1) は，いわゆるメタサーチの機能である．個々の検索サイトに対するラッパーにより入出力の違いを隠蔽し，得られた複数の結果を組み合わせてユーザに提示する．(2) は，出力結果のページの解析により著者名及び共著者名を抽出し，それらを一覧表としてユーザに提示する．これは (3) の機能へのポインタともなっている．(3) は，得られた情報を元に繰り返し検索を行う機能である．得られた論文一覧中の著者名をクリックする事で再び新たな検索を行い，その著者に関する

(注1)：http://www.bookpark.ne.jp/ipsj/, 会誌，英文誌，研究報告，論文誌 (ジャーナル)，欧文誌，論文誌 (トランザクション) を含む
(注2)：http://search.ieice.org/jpn/search-j.html
(注3)：http://search.ieice.org/search.html
(注4)：http://tjsai.jstage.jst.go.jp/ja/
(注5)：http://www.jstage.jst.go.jp/browse/jssst/-char/ja/

論文情報を提示する．

この３つの機能のうち，(2), (3) は文献検索システム DBLP で用いられているものと同等である．検索結果に対するこのような処理を含む機能は，利用しているDB の直接的アクセスが必要なので，DBLP のように通常システム中に組み込まれなければならない．一方，我々の提案する方式では，独立した文献検索システムを統合するだけでなく，この (2), (3) の機能をそれぞれのシステムの外部に構成することができる．

これらの機能のデータ結合の模式図を図 5 に示す．

本システムの基本動作をみよう．先ず最初に，著者名による検索か，キーワードによる全文検索を行う（図 6）．本システムは入力された条件（キーワード or 著者名）を各検索サイトの要求するフォームに変換する．そのフォームを各検索サイトへ送り，それぞれ検索を行う．得られた結果は，各検索サイト毎のラッパーでフィールド単位に分割し，全てのサイトからの結果を一つの表にまとめてからユーザに提示し，同時に次の検索へデータを渡すためのリンクを生成し，各著者名に関連付ける（図 7）．

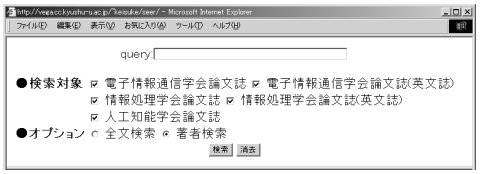

図 6　著者検索システム（プロトタイプ）

図 7　検索結果の例

ユーザは，参照したい著者名をクリックするだけで，順次関連情報を検索して行く事が可能である．我々は本システムを，`http://vega.cc.kyushu-u.ac.jp/guruguru` にて公開している．

4. 関 連 研 究

従来の手法 [17], [18] では，各 WebDB の詳細情報を開発元からされること，あるいは共通形式のデータへの変換プログラムが提供されることを想定している．本論文での手法では，各検索サイトの Web インターフェースだけから必要な情報を得るものであり，各サイトの開発，運用システムとは完全に独立に実現できる．

北村ら [7] は，WWW より情報を抽出し統合するスクリプト言語 MetaCommander を実装し，HTML ページから希望するデータを抽出する為の手順をスクリプトとして記述する事で目的のデータを入手するシステムを示している．しかし，タグや文字列として表された HTML 文書にどのようなデータ構造が含まれているかをスクリプトを書くユーザーが考え，そのデータ構造の表現形式をタグや文字列として記述し，抽出したり変換したりする処理を，タグや文字列の細かい処理として表現しなければならない．例えば，「出力される本の情報から著者名前の項目を抜き出す」というような意味的な記述は，MetaCommander ではできない．

情報融合のエージェントについての関連研究としては，Knoblock らによる ARIADNE [8] がある．これは一般の Web(Visible Web) を対象に，学習に基づいた情報抽出エージェントを容易に構築するための枠組みと，それらを組み合わせるための枠組みを与えているが，本稿のように入出力データの組という一般的枠組ではない．

本稿で検索サイトとよびエージェント構築の部品とみなすページは，単純で静的な HTML ページではなく，検索結果として動的に得られるものであり，**Invisible Web** [11], [12]，**Deep Web** [1] あるいは **Hidden Web** [5], [6] と呼ばれる．検索サイトを入出力データ構造の組として捉え，プログラムから直接利用できるようにするためには，入出力データのメタデータを推定しそのサイトに対するラッパーを作らなければならない．ラッパー自動生成については本稿の対象ではないが，人手で生成したサンプルに対し機械学習を適用しラッパーを自動生成する Kushmerick ら [9] や，繰り返しパターン発見に基づく田口ら [13] の方式など，従来から多くの研究があり今後の研究も期待できる．

5. ま と め

本発表では，専門検索サイトを入力データ構造と出

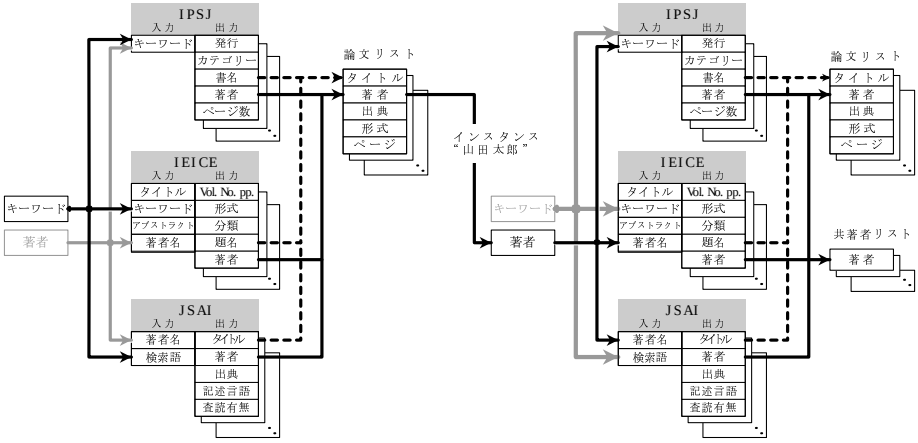

図5　データ結合

力データ構造で規定される抽象的部品として捉え，それらを結合することにより新たな検索エージェントを構成する方法を提案した．本方式の実現可能性を検討するため，情報処理学会，電子情報通信学会，人工知能学会，及び日本ソフトウェア科学会の各学会の論文検索システムを対象としたプロトタイプを実装した．

我々は既に Dnavi に登録されている 2,880 件の専門検索サイトについて分析を行っている．例えば，そのうちの 1,541 件のサイトでは複数の入力項目を要求する構造的な検索エンジンであることが確認できている．また，Query Probing によるサイトの属性抽出 [3], [4] と組み合わせることで，必要な検索サイトの自動選択が実装できると考えられる．現在，本提案をより一般的なものとするため，各検索パーツの Web Service としての記述を試みている．WWW に存在する非常に多くの検索サイトを Web Service として統一的に扱い，自由に組み合わせることにより，より複雑で有用な情報統合が可能となるものと考えている．

文　献

[1] BrightPlanet, The Deep Web: Surfacing Hidden Value, BrightPlanet White Paper, 2000.

[2] 福田 直樹, 肥塚 八尋, 和泉 憲明, 山口 高平. 連携品質を考慮したオントロジーに基づく Web サービスの自動連携 FIT2004, pp. 307-310, 2004.

[3] S. Hirokawa, S. Watanabe, Y. Koga and T. Taguchi, *Automatic Feature Extraction of Search Sites*, Proc. SSGRR2001(CD-ROM).

[4] P. Ipeirotis, L. Gravano and M. Sahami, *Automatic Classification of Text Databases through Query Probing*, Proc. of the ACM SIGMOD Workshop on the Web and Databases (WebDB'00), 2000.

[5] P. Ipeirotis, L. Gravano and M. Sahami, PERSIVAL Demo: Categorizing Hidden-Web Resources, JCDL2001, 2001.

[6] P. Ipeirotis, L. Gravano and M. Sahami, Probe, Count, and Classify: Categorizing Hidden-Web Databases, ACM SIGMOD 2001, 2001.

[7] Yasuhiko Kitamura, Tomoya Noda, and Shoji Tatsumi, Single-agent and Multi-agent Approaches to WWW Information Integration, Multiagent Platforms, Lecture Notes in Artificial Intelligence, Vol. 1599, Berlin et al.: Springer-Verlag, 133-147, 1999.

[8] Knoblock, C. A., S. Minton, J. L. Ambite, N. Ashish, I. Muslea, A. G. Philpot, and S. Tejada, The Ariadne Approach to Web-Based Information Integration, International Journal of Cooperative Information Systems, vol.10, no.1-2, pp.145-169, 2001.

[9] N. Kushmerick, D. Weld and B. Doorenbos, *Wrapper induction for information Extraction*, IJCAI'97, pp .729-737, 1997.

[10] T. Nakatoh, K. Ohmori, Y. Yamada and S. Hirokawa, *COMPLEX QUERY AND METADATA*, Proc. ISEE2003, pp. 291-294, 2003.

[11] P. Pedley, The invisible web, ASLIB, 2001.

[12] C. Sherman and G. Pric, The Invisible Web, Infomation Today, Inc., Medfore, New Jersey, 2001.

[13] T. Taguchi, Y. Koga and S. Hirokawa, Integration of Search Sites of the World Wide Web, Proc. of International Forum cum Conference on Information Technology and Communication, Vol. 2, pp. 25-32, 2000.

[14] CiteSeer, `http://citeseer.ist.psu.edu/`

[15] 専門検索サイトの動的統合による次世代検索システム DAISEn, Directory Architecture for Integrated Search Engines, `http://daisen.cc.kyushu-u.ac.jp/`

[16] DBLP, `http://dblp.uni-trier.de/`

[17] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. "The TSIMMIS Project: Integration of Heterogeneous Information Sources". In Proceedings of IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.

[18] 菅坂 玉美, 益岡 竜介, 佐藤 陽, 北島 弘伸, 丸山 文宏. 知的エージェント環境 SAGE の EC への適用, 取引フェーズへの適用. 第 6 回マルチ・エージェントと協調計算ワークショップ (MACC), 日本ソフトウェア科学会, 1997 年 12 月.

[19] 国立国会図書館関西館データベース・ナビゲーション・サービス Dnavi, `http://dnavi.ndl.go.jp/`