社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DBS－142 (5)
2007－ＦＩ－ 87 (5)
2007／5／31

# WebDBにおける出力レコードのメタデータ自動抽出

中藤 哲也† 森 雅生†† 廣川佐千男†

† 九州大学 情報基盤研究開発センター
†† 九州大学 大学評価情報室

**あらまし** ブラウザに表示される入力フォームにおいて，属性ごとにキーワードを指定して検索が可能なWebデータベースが増えている．さらに，このようなサービスをアプリケーションから直接利用する枠組としてWebサービスがある．多数のWebサービスのプールから必要なものを選択し，組み合わせることにより新たなサービスを構築する研究が多くの注目を集めている．しかし，一般に公開されているWebサービスはWebデータベースと較べごく少数である．本稿では，WebデータベースをWebサービスとして利用できるようにするために，検索結果の出力からメタデータを推定する方法を提案する．

# Automatic Extraction of Output Metadata from Web Databases

Tetsuya NAKATOH†, Masao MORI††, and Sachio HIROKAWA†

† Research Institute for Information Technology, Kyushu University.
†† Office of Information for University Evaluation, Kyushu University.

**Abstract** There are increasing number of Web sites which dynamically generate web pages from their data bases according to users request specified with attributes and keywords. On the other hand, Web services are programmable components to provide services via the Web and are gaining much attention due to its composition mechanism. However, the number of available Web services is very small compared to Web databases. In this report, the authors propose a method which transforms a Web database to a Web service by extracting the set of output attributes to the site.

## 1. はじめに

Web上には，静的なページ上のコンテンツだけでなく，検索機能によって動的に生成された（定まったURLを持たない）Webページ上のコンテンツも多く存在する．一般に直接参照する事ができないため，それらのページはInvisible Web [15], [16]，Deep Web [1]，Hidden Web [2], [3]などと呼ばれている．それらは多くの場合，背後に隠されたデータベースを持っており，ユーザの質問に応じてデータベースから情報が取り出され，HTMLとして再構成されてブラウザに出力される．Webのインターフェイスを持つデータベースという意味で，それらを**Webデータベース**と呼ぶ．

Webデータベースの多くは，特定のテーマに限定した質の高い情報やサービスを提供している．例えば，Amazon.com [20]は本のリストを返す．kakaku.com [21]はPCのリストと共にそれらの価格を返す．Travelocity [22]は指定されたエリアのホテルのリストを返す．定まったURLを持ち直接参照可能なWebページよりも，これらのWebデータベースは多くの情報を持つと言われている．このため，それらのデータの自動的な取り扱いは，情報抽出の重要な研究テーマの一つであり，メタサーチや情報統合といった観点で多くの研究がなされて来た．

我々は現在，これらのWebデータベースの情報を，単に統合するだけではなく，それらの連携によって新しいサービスの生成が自由に行える環境の構築を目的に研究を進めて来た．例えば，書籍やCDの推薦システムとお気に入りの通販サイトを連携させる事で，自分専用のショッピングサービスサイトを構築する，あるいは，いつも使うホテルの予約サービスとマイレージカードを持った航空会社の予約サービスを組み合わせて，自分がいつも行なうような出張の手配を素早く行えるようにする事，な

どが可能となる．

近年，Web サービスをユーザサイドで自由に組み合わせて新しいサービスを構築する**マッシュアップ**と呼ばれる活動が盛んに行われ始めている．これまでは，サービス提供プロバイダや企業などが，収益を目的としたり，集客や広告を目的として，多くの人にメタサーチや情報統合システムなどのサービスを提供して来た．それに対してマッシュアップにおいては，主に個人ユーザが自由な発想や感覚で Web サービスを組み合わせる事で，新しい情報統合サービスの構築がなされている．ユーザ自身の利用のため，あるいは楽しみのためにサービスが構築され，多くの人に喜んでもらうために，あるいは自己表現のために，それらを公開されている．マッシュアップの公開目的に運営されているサイト Mashup Feed (注1) には，2007 年 5 月当初においては，1,865 のマッシュアップが登録されている．

しかし，これらのマッシュアップに用いられているのは一般に公開されている Web サービス (API) のみであり，その数は多くない．同サイトに登録されている公開 Web サービスの数は僅か 426 に過ぎない．一方，Web データベースは Web サービスよりも多く存在する事が明らかであるが（我々による収集は実際の Web データベースのごく一部であると考えられるが，それでも既に 121,782 サイトに及ぶ），Web サービスのような情報をやり取りをするための共通のプロトコルを仕組みを持たないためにマッシュアップに利用されていない現実がある．

より多くの情報源となる Web データベースを自由に利用するためには，ユーザ（人間）に対するインターフェースのみを持つ Web データベースを機能化（Web サービス化）する必要がある．その為には，次の 5 つの機能が求められる．

(A) 入力インターフェースからのフォーム情報取得
(B) 各入力フィールドのメタデータ抽出
(C) 検索結果からの個別データの切り出し
(D) 出力データの各フィールドのメタデータ抽出
(E) Web データベース，Web サービスの連携の記述

我々は，Web データベースの機能化（Web サービス化）の観点から，上記のそれぞれに関する研究を行っている．(A) については，HTML で書かれた Web データベースの入力インターフェースを解析し，実際に Web データベースより問い合わせを行うことができるように形式化した [8], [11], [12]．(B) については，同系統 Web データベースに関する事前知識や統計処理を用いずに，入力フィールドのメタデータを抽出する手法を提案した [9]．(C) に関しては，反復パターンの発見に基づく頑健なデータの切り出し抽出手法を提案した [13]．(E) に関しては，Web データベースの自由な統合による新しいサービスの構築が可能となるようなアーキテクチャを提案している [6], [7]．

本稿では，(D) の出力データの各フィールドのメタデータ抽出を扱う．Web サービスであればスキーマが明示的に与えられているが，Web データベースにおいてはユーザによるブラウザ経由での利用しか想定されていないため，メタデータは明らかではない．従って，Web データベースの統合に際してスキーマを参照したい場合，検索結果の出力のページからスキーマを推定，抽出する必要がある．しかしながら，出力結果のページには推定に必要な情報ですら充分ではない現実がある．我々は本稿において，スキーマを確定させるタイミングを情報が得られる時点まで遅延させ，各フィールドの属性に関して現実的な情報提供を行う手法を提案する．

## 2. 関連研究

Web データベースの出力スキーマの推定・抽出に関する研究は，基本的にスキーマ・マッチングの研究として行われている．これは，出力スキーマの推定に必要が情報が充分では無いこと，またホモジニアスな Web データベースの情報統合が目的となっている事が多いため，同系統の Web データベース間にまたがる共通の情報を有効利用するためである．

これらのスキーマ・マッチングは大きく次の 2 種に分類できる．

一つは入力フィールドや出力要素の周囲の文字列からその要素の属性を見つけ出すものである．この観点からの研究には，複数の Web データベースの共通の構造としての文法を想定するもの [19]，表構造に着目するもの [9]，木構造として解析するもの [18] などがある．しかしこれらは主に入力フィールドのメタデータ抽出を目的としている．

スキーマ・マッチングに関するもう一種の研究は，入力や出力の要素に関して，同系統の Web データベースのインスタンスを収集し，その統計的性質からスキーマのマッチングを図るものである [17]．しかしながら，同系統の Web データベースがない場合，あるいは充分な統計的性質が得られない場合，すなわち我々の目標には，この手法は適応できない．

## 3. メタデータ抽出

Web データベースのメタデータ抽出に関する既存の研究では，同系統の Web データベースに関する事前知識や統計的情報を用いている [17]〜[19]．ところが，我々の目的とする Web データベースのサービス結合において

(注1)：http://www.mashupfeed.com/

は，そもそも同系統の Web データベースが存在しない場合も多く，それに関する事前知識も統計的情報も期待できない．したがって，目的 Web データベースの構成する入力インターフェースの出力 HTML，及び出力結果の HTML に出現する情報のみから，メタデータを抽出する必要が生じる．すなわち，入力項目を説明する文字列や出力データの周囲に出現する文字列をそれらの属性候補とし，何らかの方法で曖昧性を解消することで，属性を決定し，メタデータを構成する必要がある．ところが，それらの情報が充分ではない現状がある．例えば，書籍通販の amazon.com の検索結果には，属性に相当する文字列が存在しない．これは，閲覧するユーザが人間の持つ常識で判断することが可能であるために，特に属性を記載することが必要ではない事が背景となっていると思われる．

ここでメタデータ抽出の目的に立ち返ってみると，我々はユーザ自身による新たなサービスの構築，Web データベースの自由な組み合わせのための環境を作る事を目的としており，その結合ツールを利用するユーザに対して，各フィールドの属性に関する情報を提供するためにメタデータを必要としている．その観点から整理すると，Web データベースに関して次の観察を得る．

- 入力フィールドは通常，属性となる情報を文字列として持っている．入力フィールドについての属性が示されないと，ユーザは何を入力すべきか判断できないためである．
- 検索結果の出力ページ中に，各フィールドの属性に関する情報が何ら記述されていない Web データベースがある．
- 同じメタデータのサブセットである為，入力フィールドの属性は出力結果のフィールドの属性にも出現する場合が多い [10], [14].
- 入力クエリーと出力結果は基本的に関連する．したがって，入力クエリー自体が検索結果の同じ属性項目中に値として出現する場合が多い．
- ユーザ（人間）はインスタンスを見ることで，多くの場合その属性を想定できる．
- 異なる Web データベースのフィールドであっても，ユーザが結合したフィールド同士は，同じ属性を持つと期待できる．

これらの観察から，本稿では，遅延評価と属性情報の伝播によるメタデータ抽出手法を提案する．提案のポイントは次の 5 点である．(1) 入力メタデータ，出力メタデータ共に，属性候補の絞込みが行えない場合には，無理に絞り込まず候補集合のままそれらを保持する．(2) 生成された出力メタデータの候補の他に，ユーザへの提示情報として実際に出力されたインスタンス（の一部）を用いる．(3) 入力メタデータ候補と出力メタデータ候補の間でのマッチングを行い，曖昧性の解消（候補の削減）が可能である場合にはそれを行う．(4) 入出力メタデータのマッチングが不調の場合，あるいは一方もしくは両方のメタデータ自体が得られていない場合，インスタンス中の同一項目の検索によるフィールド同士のマッチングを行う．(5) ユーザによって Web データベースの結合が行われた際には，その結合フィールド同士で属性候補に関する曖昧性の解消を行う．

以下に処理の詳細を示す．

### 3.1 属性名候補集合の生成

Web データベースの入力インターフェースのスキーマ，すなわち各入力要素の属性名の抽出に関しては，TABLE タグの解析に基づく方法を提案している [9]．この提案においては，複数の属性名候補からなる曖昧性をあえて絞り込まずに残す事が可能である．

他方，Web データベースの検索結果出力のスキーマ，すなわち各出力データのフィールド要素の属性名の抽出に関しては，表からの属性の自動抽出に関する幾つかの既存研究が流用可能である．しかしながら，複数の表にまたがる統計的な手法が使えず，また属性に相当する文字列が出現しない例もある事から，結果スキーマに関しても属性名の候補集合の抽出に留める（Recall 重視のスキーマ抽出を行う）．

以上で入出力フィールドの属性候補の集合が得られる．その集合から正しい属性を選択する為には，3.3 節，3.4 節，3.5 節に示す方法を用いる．

### 3.2 出力データのインスタンスの保持

Web データベースの検索結果の出力には属性名に関する情報が含まれない例もあり，その場合にはラベルとなる文字列を発見する手法に基づく方法は無力である．但し，属性名が得られない場合であっても，ユーザ（人間）は多くの場合，そのインスタンスによって属性が概ね判別可能である．したがって，ユーザによる Web データベースの結合を目的とする場合，接続するフィールドを特定するための情報として提示するのは，必ずしも正しく確定した属性名である必要はない．属性名が不完全である場合にも，ユーザはインスタンスを補助情報とすることで，多くの場合正しい属性を判別できる．

以上の観点から，ユーザに提示する情報の一部として，出力データの各フィールドのインスタンスの実例を保持する．これらを必要に応じてユーザに提示する事で，ユーザは Web データベース同士の結合を正しく記述でき，その結合の情報により 3.4 節に示す方法による属性名の取得が可能となる．

### 3.3 入出力メタデータにおける属性候補の絞り込み

入力フィールドのメタデータと出力データのメタデータは，共に背後にあるデータベースのメタデータに関連付けられており，多くの場合，そのデータベースのメタデータのサブセットである．それ故，各入力フィールドの属性と出力データの各フィールドの属性は，共通のものが現れる可能性が高い．そのため，双方に共通する候補を属性名をして選択できる．属性名の選択と同時に，入力と出力のフィールドの対応付けも得られる．

また，一般に出力されるフィールド数の方が入力フィールドの数よりも多いため，属性名の決定に必要な参考情報を得られない出力フィールドも多い．しかしながら，幾つかの出力フィールドの属性名に関する曖昧性が解消されている場合，それらと同じ位置にあった属性名候補を選択する事で適切な候補が選べる可能性が高い．

### 3.4 入出力フィールドのインスタンスによる対応付け

一方，あるいは双方に属性名候補がない場合であっても，共通に出現するインスタンスによってフィールドを対応付ける事が可能である．対応付けられた一方に確定した属性名，あるいは属性名候補があれば，他方にそれを適応することができ，双方に属性名がない場合であっても，3.5節の方法で得られた属性に関する情報を，この対応付けでもって伝播させることができる．

### 3.5 Web データベース結合による属性候補の絞り込み

ユーザは，各フィールドに表示された属性名，属性名候補の集合，あるいはインスタンスを参考に，異なる Web データベースのフィールド同士を結合する事で情報の流れを定義し，新しいサービスを構築する．この際結合されたフィールド同士は，同じ属性を持つと期待される．したがって，双方が属性名候補の集合を持つ場合，その積集合を新しい双方の属性名とする事で，曖昧性の解消が可能である．また，一方のみに属性名に関する情報があれば，それを双方に共通の属性名候補とすることができる．

## 4. ま と め

本稿では，Web データベースの自由な連携，マッシュアップを目的とした，メタデータの抽出手法を提案した．今後，評価実験を行い，その有効性を示す予定である．

## 文　献


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

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

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

[4] 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.

[5] 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.

[6] Masao Mori, Tetsuya Nakatoh and Sachio Hirokawa. Functional Composition of Web Databases. Proc. of the 9th International Conference on Asian Digital Libraries (ICADL 2006), LNCS 4312, pp.439-448, 2006.

[7] 森 雅生, 中藤 哲也, 廣川 佐千男, マッシュアップを簡単に実現するメタ CGI とそのアーキテクチャ. IPSJ SIG Technical Reports, 2007-DBS-142, 2007.

[8] 中藤哲也, 酒井美由紀, 廣川佐千男, 検索サイトのための集合演算子の自動推定, 第 1 回情報科学技術フォーラム (FIT2002), 一般講演論文集第 2 分冊, pp. 9–10, 2002.

[9] 中藤 哲也, 大森 敬介, 廣川 佐千男. WebDB の Query-Form におけるメタデータ自動抽出, DBSJ Letters Vol.5, No.2, pp. 97-100, 2006.

[10] T. Nakatoh, K. Ohmori, S. Hirokawa, Report on Metadata for Web Databases, IPSJ SIG Technical Reports, 2004-ICS-138, pp.95-98. 2004.

[11] T. Nakatoh, M. Sakai, Y. Koga and S. Hirokawa, Generation of Query URL for Search Sites, Proc. of SSGRR2002w (CDROM), 2002.

[12] Tetsuya Nakatoh, Yasunori Koga, Axel Uhl and Sachio Hirokawa. Automatic Estimation of Query Syntax for Search Sites, Proc. of PYIWIT'02, pp.329-332. 2002.

[13] Tetsuya Nakatoh, Yasuhiro Yamada and Sachio Hirokawa. Automatic Generation of Deep Web Wrappers based on Discovery of Repetition, Proc. of the First Asia Information Retrieval Symposium 2004 (poster), pp.269-272. 2004.

[14] 大森 敬介, 中藤 哲也, 原 由加里, 廣川 佐千男. 検索サイトにおける入力項目と検索結果のフィールド名の対応調査 FIT2004, pp. 89-90, 2004.

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

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

[17] Jiying Wang, Ji-Rong Wen, Frederick H. Lochovsky and Wei-Ying Ma. Instance-based Schema Matching for Web Databases by Domain-specific Query Probing. Proc. of the 30th International Conference on VLDB, pp. 408-419, 2004.

[18] Wensheng Wu and Clement Yu and AnHai Doan and Weiyi Meng. An interactive clustering-based approach to integrating source query interfaces on the deep Web, Proc. of the 2004 ACM SIGMOD international conference on Management of data, pp. 95–106, 2004.

[19] Zhen Zhang, Bin He, Kevin Chen-Chuan Chang, Understanding Web Query Interfaces: BestEffort Parsing with Hidden Syntax, Proc. of the ACM SIGMOD International Conference on Management of Data (SIGMOD2004), pp. 107-118, 2004.

[20] Amazon.com, `http://www.amazon.com/`

[21] kakaku.com, `http://www.kakaku.com/`

[22] Travelocity, `http://www.travelocity.com/`