Towards A Unified View of Service-Oriented Web

Pushpa Kumar[1], Guanglei Song[2], Kang Zhang[3]

***Abstract* - This paper discusses the current landscape of the services available over the Web and proposes a unified view for the future service-oriented Web. Apart from the current Web services accessible by Web applications, other services are directly provided to end users through Web pages, among which the "deep Web" provides online queries through Web pages for users to access the hidden databases. This paper presents our categorization examples to illustrate this concept, as well as the methodology used to categorize these hidden Web services. We illustrate how computational Web services can improvise supply chain management as well as delivery logistics.**

***Index Terms* - categories, executable web, hidden web services, service-oriented Web**

I. Introduction

Web services have attracted much research effort and gained great progress. With the advance of Web service techniques, a growing amount of such services are registered in UDDI. These registries provide a list of all registered Web services in a certain category. UDDI provides a convenient mechanism for users to find Web services [21].

In recent years, the Web has been "deepened" by online databases that support Web query forms [4], and these query forms supported by back-end databases is defined as the so-called Deep Web. Apart from the Web query interfaces supported by databases, many services are presented through Web pages, such as currency conversion and language translation. Our preliminary study through current Web search engines suggests that when provided with a keyword, resultant Web query forms that accept user inputs and produce outputs could be computation-based and not necessarily supported by databases. Similar to the Deep Web, these Web pages also provide services for users through forms, though the services are provided through some computation logic. For example, consider the keyword "bmi calculators" that determines the body mass index of a person as depicted in Figure 1. The popular search engine "google" returned a total of 224,000 responses for this search. On sampling the first 100 responses (the first 10 pages with 10 results on each page), 83 of these presented Computation based Web query forms. As such, statistics (83%) for this example shows the significant amount of non-database oriented services.

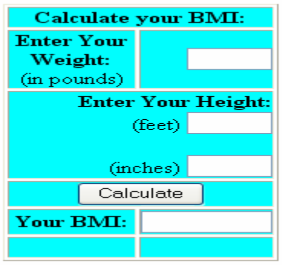


Figure 1. BMI calculator query form

Computations in these services occur behind the scenes and hence are also invisible to search engines. Search engine spiders access the surface Web and cannot go inside the computation logic that is literally "invisible" to them. In this paper, we refer to those computation-based services as executable Web, or "X Web" in short. They, together with the Deep Web, make up the hidden Web services, which unlike the Web services; have no program-oriented interfaces or service descriptions like WSDL. Without machine-readable descriptions and interfaces, these hidden Web services are yet to be discovered and utilized. We attempt to discover and categorize these hidden Web services, aiming at eventually providing common program interfaces accessible to all the services on the Web.

As an example of its application, we illustrate how X Web can contribute to supply chain management (SCM). SCM is the process of optimizing the shipment of goods and services from supplier to customer. Since X Web provides Web services through forms, external customers can utilize this mechanism for establishing a logistics trace through the supply chain. Each product order and delivery can be tracked through confirmation numbers generated by the transaction system automatically. The customer has added advantages of the convenience to track from an online computer without human intervention, and having more overall control.

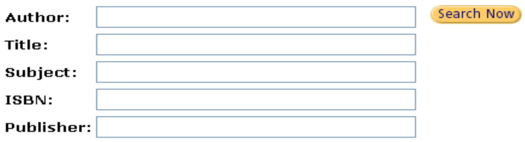
Consider a customer trying to purchase a book online through amazon.com's web service shown in Figure 2. Once the user selects a particular book of her choice, and makes payment for the book as well as mail delivery (UPS), she receives a confirmation/tracking number. The provided tracking number serves as input for UPS.com's Web service shown in Figure 3. This example illustrates a delivery

[1] Pushpa Kumar is with the Department of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688, USA (phone: 469-878-8026; fax: e-mail: pkumar@utdallas.edu).

[2] Guanglei Song is with the Department of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688, USA (e-mail: gxs017800@utdallas.edu).

[3] Kang Zhang is with the Department of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688, USA (e-mail: kzhang@utdallas.edu).

logistics trace.


Author:

Title:

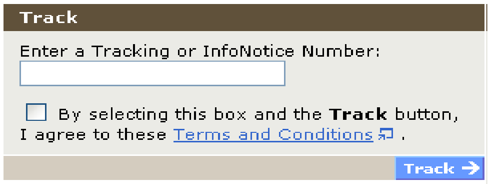
Subject:

ISBN:

Publisher:

Search Now

Figure 2. Amazon.com's Web service


Track

Enter a Tracking or InfoNotice Number:

□ By selecting this box and the **Track** button, I agree to these Terms and Conditions .

Track →

Figure 3. UPS.com's Web service

In order to measure the significance of hidden Web services, we identified keywords for each category and used search engines such as google, and documented a few Web query forms that perform computations when requests are submitted. For the category of "bmi calculators", the form presents various fields used to calculate results. The resultant data is presented solely based on the information the user provides in the query form and is not obtained from any database.

The rest of the paper is organized as follows. Section II discusses the current status of the services available on the Web. Section III presents our vision towards a service-oriented Web. In Section IV, we discuss the methodology used for this research, and present details of category examples. Section V describes possible ways to achieve our vision. Section VI discusses related work followed by conclusions in Section VII.

II. CURRENT SERVICES ON THE WEB

The part of the Web that is served dynamically "on the fly" is far larger than the static documents associated with Web pages. This section overviews the current types of services available on the Web. Figure 4 depicts the existing world of Web services. Current Web services communicate through application program interfaces (APIs) and cannot be accessed by end users. They also interoperate through the APIs. Current Web services are visible to Web applications through their APIs, while hidden Web services, including the Deep Web and X Web, are designed to serve end users rather than Web applications.

Search engine spiders as well as end users can access only the surface Web through static URLs or links, but not the Deep Web or X Web. Promising work has been done on making the Deep Web accessible not only to end users but also to Web applications [6]. Our aim at integrating the Deep Web, the X Web, and existing Web services through common program interfaces is discussed in Section V.

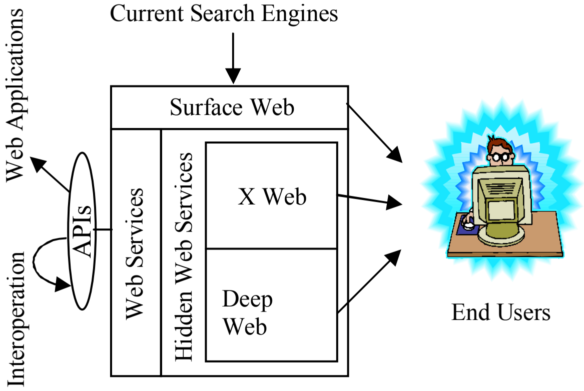


Figure 4. "X Web" in existing world of Web

A. *Existing Web Services*

The simplest and most basic definition of a Web service is an application that provides a Web API that supports application-application communication and interoperation with other Web services [22]. Web services represent a category of resources on the Web and are designed to be accessed only by Web applications or programs, rather than by human users [5] [19]. For example, Amazon offers a Web API for its online catalog that allows its marketing affiliates to easily incorporate Amazon contents and features into their Web sites. This is depicted in Figure 5. The affiliate Web site uses the Amazon Web service to search Amazon's catalog and display the results on its own site, including features such as Amazon reviews and book ratings [19].

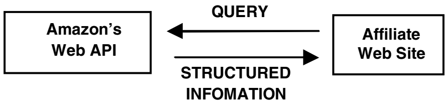
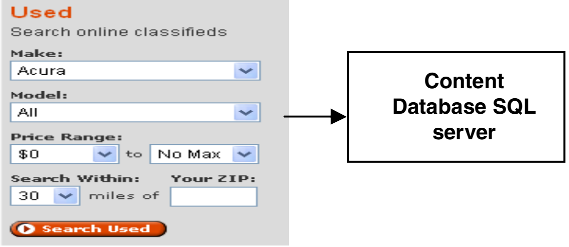


Figure 5. "X Web" in existing world of Web

B. *Surface Web and Deep Web*

The contents in databases accessible on the Web are distinct from static and surface Web pages, which are essentially documents possibly with multimedia files accessed directly by end users and search engines. A significant amount of valuable information on the Web is however generated from databases below the Web surface. It has been estimated that the content of the Deep Web may be 500 times larger than the surface Web [15]. As an example of the Deep Web, cars query form shown in Figure 6 retrieves queried results from the underlying database on cars.


Used

Search online classifieds

Make: Acura

Model: All

Price Range: $0 to No Max

Search Within: 30 miles of Your ZIP:

Search Used

Figure 6. "Deep Web" example

C. X Web

The computation logic that occurs behind the scenes of Web interface forms is hidden from the user and is also invisible to search engines. An example of this type of Web is the "bmi calculator" illustrated in Figure 7. The computation logic may not be driven by an underlying database. Comparing to the Deep Web, the X Web generates an infinite number of content pages according to users' inputs. The next section will show the significant presence of this type of services available on the Web.

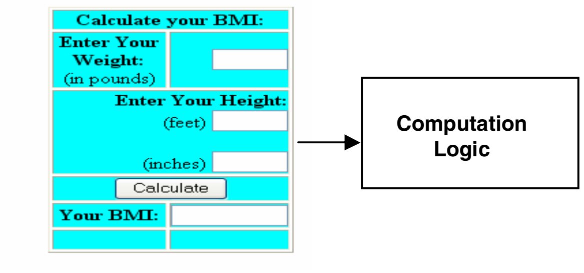


Figure 7. "X Web" example

III. Towards the Service-Oriented Web with Common Program Interfaces

Figure 8 illustrates our vision on the future Web service infrastructure, which we will simply call the *service-oriented Web*. Current Web services can communicate through program interfaces, and do not provide access to end users, while hidden Web services serve end users through Web user interfaces. Our ultimate aim is to provide common program interfaces to hidden Web services as well as the existing Web services. Automation of Web analysis, service profiling, and search can then be achieved. A service discovery engine will be able to find desired services through the common program interfaces.

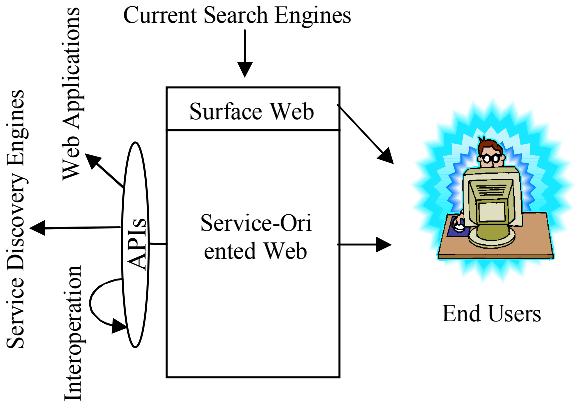


Figure 8. Service-oriented Web

IV. Categories

Fundamental to our work on hidden Web services is the determination of categories representing this classification of Web services through an experimental survey. Section A presents the methodology used to perform this experiment. Section B provides detailed examples including the obtained statistics.

A. Methodology

Most query forms seem to follow a concerted structure and appear to be "modularly" constructed upon a small set of building blocks [24]. Such query structures allow us to categorize different types of services. Within each category, this observation steers us towards the prediction, classification and ultimately discovery of the services according to their query forms.

We utilized Web service repository to find hidden Web services. Using information on categorization present in the registries, keywords were identified for use in the existing search engines to find related Web sites. The Web page containing query interfaces was downloaded and the title, url, domain information were recorded in an XML file. The following example illustrates this methodology. We used Microsoft UDDI (http://uddi.microsoft.com) to locate the "finance" category. Then we browsed through the Web service entries returned by UDDI and identified the Web service, "Calculate IT - Web Service to Calculate the Income Tax" for this category. After identifying the key words "Calculate the Income Tax", we used these key words as input to Google. As we browsed through the returned results, we tried to accept user inputs and produce results through some kinds of computations The Web page at "http://www.dinkytown.net/java/Tax1040.html" met this criterion. We recorded the service in an XML file and saved related Web pages. We also recorded its title (Financial Calculators) and domain (Finance). Some of these categories were also determined by their frequent usage such as financial calculators.

B. Detailed examples

Using the above methodology, we identified several categories that perform computations via Web query forms. The results of services are documented in Table 1. The recorded title and URL information are depicted. There are a total of 10 categories in Table 1, with about two or three types of services for each category. Categories cover many fields like health, food and finance.

We also determined sampling and estimation rates through the search engine Google for these categories. Sampling was performed on the first 100 results, i.e. 10 pages each with 10 results. This is documented in Table 2. The statistics (approximately in the range of 80 to 85%) provides a good validation for considering these categories in hidden Web services for our proposed work. The resultant data is presented solely based on the information the user provides in the query form and is not connected to any database.

Table 1: Various categories under the X Web

CATEGORY	TITLE	WEB SERVICE URL
Mortgage	interest.com	http://mortgages.interest.com/content/calculators/additionalpayment.asp
Calculator	banksite.com	http://www.banksite.com/calcs/mortgagecalc.html
	hsh.com	http://www.hsh.com/calc-amort.html
Unit Converter	digital dutch	http://www.digitaldutch.com/unitconverter/
	institute of chemistry	http://www.chemie.fu-berlin.de/chemistry/general/units_en.html
	measurement unit converter	http://www.people.virginia.edu/~rmf8a/convert.html
Currency	x-rates.com	http://www.x-rates.com/calculator.html
Converter	xe.com	http://www.xe.com/ucc/
	oanda.com	http://www.oanda.com/convert/classic
Finance	Financial Calculators	http://www.dinkytown.net/java/Tax1040.html
	Missouri Department of Revenue	http://www.dor.mo.gov/tax/calculators/incometax/
Food	Internet Pizza Server Ordering Area	http://www.ecst.csuchico.edu/~pizza/pizzaWeb.html
Bmi Calculator	halls.md	http://www.halls.md/body-mass-index/bmi.htm
	preventdisease.com	http://preventdisease.com/healthtools/articles/bmi.html
Life Expectancy	msn	http://moneycentral.msn.com/investor/calcs/n_expect/main.asp

Categorization results for the Deep Web are depicted in Table 3. There are a total of 10 categories shown which also cover various fields like travel, books and license renewal. This presents a general idea of the differences between the Deep Web and X Web. For example, consider "amazon.com" under books category. When provided with a query for book search with a title "Unix", it returns a list of books with the complete title, author name, publisher name and price that are retrieved from the database. This example represents Deep Web. Under X web category, consider "halls.md" bmi calculator. In the query form, when a user enters weight in pounds and height in feet/inches, it returns a BMI index number in kg/m^2 units in the result field. This is a result of some computation behind the scenes.

Analysis on the fields of the Web query forms revealed similarities in the form structure for various categories in the X Web. For example, consider the category "currency converter". Figures 9 and 10 depict two separate query forms that are returned from their respective links. As we can see, they have three fields to enter "amount" value, "from" currency and a "to" currency for conversion to take place. This similarity in the form structure can be used for category mining and prediction, which we discuss as proposed work in the next section.

Table 2: Sampling results for the X Web

CATEGORY	ESTIMATE%	TOTAL RESULTS
Mortgage Calculator	82	7,888,000
Unit Converter	85	6,860,000
Currency Converter	80	21,800,000
Finance	81	1,030,000
Bmi Calculator	83	224,000
Life Expectancy	81	120,000
Insurance	85	7,780,000

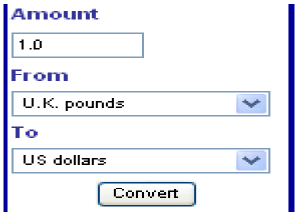

Amount: 1.0
From: U.K. pounds
To: US dollars

Figure 9. Currency converter form 1

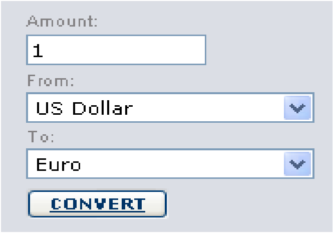

Amount: 1
From: US Dollar
To: Euro

Figure 10. Currency converter form 2

V. PROPOSED WORK

Our proposed work includes two stages. We will first discover and understand hidden Web services, and then cluster and integrate these hidden Web services like the current Web services. Our ultimate goal is to provide unified invocation mechanism towards hidden and existing Web services, so that end users and programs access the Web as a collection of services, which interoperate through the Internet.

Discovering and exploiting these hidden Web services presents great challenges to the research community. Hidden Web services are presented to end-users through Web query forms or scripts, and their Web interfaces have no machine friendly description, like WSDL, for programs to retrieve and utilize the underlying service information. Understanding these Web interfaces has been a research topic in the Deep

Table 3: Various categories under the Deep Web

CATEGORY	TITLE	WEB SERVICE URL
Books	amazon	http://www.amazon.com/books/
	barnes&noble.com	http://www.barnesandnoble.com/index.asp?r=1&popup=0
Travel	travelocity	http://www.travelocity.com
	expedia.com	http://www.expedia.com
Weather	the weather channel	http://www.weather.com
	cnn.com	http://www.cnn.com/WEATHER/
Prescription Refills	kmart	https://pharmacy.kmartcorp.com/KMRx?step=Refill
	randalls	http://www.randalls.com/RxRefill.asp
Cars	cars.com	http://www.cars.com/go/index.jsp
	startribune.com	http://www.startribune.com/cars/
Pets	yahoo!pets	http://pets.yahoo.com/pets/
	petfinder.com	http://www.petfinder.org/
Stock Quotes	lycos	http://www.quote.com/qc/default.aspx
	pcquote.com	http://www.pcquote.com/

Web community, and similar principles can be applied when extracting and understanding Hidden Web services. These techniques include syntax extraction and matching of query forms, and may produce useful results that can be adapted to hidden Web services. For example, the unique characteristics of hidden Web services, such as their extensive and growing vocabulary, need to be considered when adapting the techniques that have been successfully applied to the deep Web.

With the help of text mining and classification, the second stage of investigation involves clustering and integration of hidden Web services. Considered a categorical data clustering problem, service clustering has been a research focus in recent years. Relatively large and still growing vocabulary of Hidden Web services however presents great challenges to existing categorical data clustering algorithms. We are working on an incremental data clustering algorithm for hidden Web services that is scalable to handle a large vocabulary. Natural language processing and ontology techniques are considered complimentary to the existing algorithms.

Web input query forms reveal a common structure within a category. This observation can be used to perform category prediction. When presented with a new Web query form, a service's category can be determined using different techniques, such as correlation methods. As mentioned in Section IV, that data corresponding to various categories is stored in an XML file, we plan to use this file as input for our purposes. This xml file has a well-defined structure with nodes representing the data like <title> and <url>. We plan to develop a system that can automatically extract this data. Categories can thus act as input data as shown in Figure 11. Thanks to an automatic form extraction tool that is available from the UIUC group [24], we will extract various Web forms and then use our SGG graph grammar formalism [16] to perform transformations, automatic analysis and verification. Visualization techniques can be used to complement clustering algorithms for category prediction.

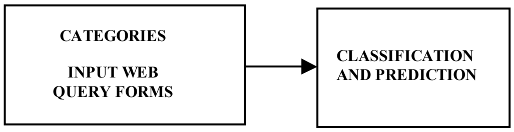
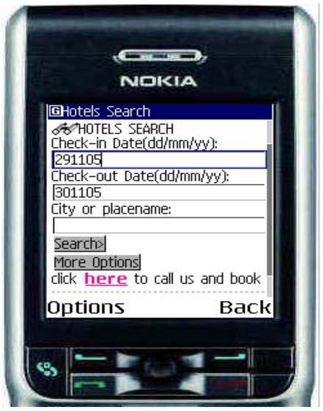


Figure 11. Categories and Prediction

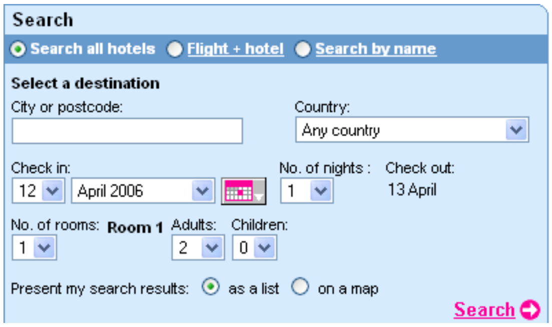
Our aim is to find more Web services, categories and statistics to strengthen our understanding of the characteristics of the X Web. We also aim to extend our research towards the mobile field. Cellular telephones display Web query forms for various categories, but due to limited display screen sizes of the phones, these forms differ from their online counterparts. For example, consider the Web site for "lastminute.com – UK region" (http://www.lastminute.com), which is in the travel category.

Screenshots for the hotels search Web query forms for the mobile and online versions are depicted in Figures 12 and 13 respectively. The basic online version query form has 9 fields, while the mobile version has only 3 fields. We aim to study the structures of these forms for "hidden web services", and also compare mobile and PC interfaces for different categories.



Hotels Search
HOTELS SEARCH
Check-in Date(dd/mm/yy):
291105
Check-out Date(dd/mm/yy):
301105
City or placename:
Search»
More Options
click here to call us and book

Figure 12. Mobile version of query form



Search

● Search all hotels ○ Flight + hotel ○ Search by name

Select a destination

City or postcode: Country: Any country

Check in: 12 April 2006 No. of nights : 1 Check out: 13 April

No. of rooms: 1 Room 1 Adults: 2 Children: 0

Present my search results: ● as a list ○ on a map

Figure 13. Online version of query form

VI. Related work

Much of the related ongoing research is in the area of the Deep Web. This is justifiable since the estimated size of Deep web is 500 billion pages, while the size of surface Web is only around two billion pages [15]. Categorization of hidden-Web databases using both manual classification and automatic methods is being investigated. Probing has been used to classify databases by training the system and then transforming rules to queries for classification [15]. There is a hidden syntax in each Web query interface, a 2P grammar and a best-effort parser had been developed to realize a parsing mechanism for the syntax [24]. According to the syntax, these interfaces can be clustered into groups for further utilization [13][23]. These Web interfaces are considered to have underlying mappings with each other and can be matched to provide an integrated interface for the same group of Deep Webs [9][10]. These proposals have provided a vision for future research direction with the Deep Web, with efficient integrated retrieval framework for the Deep Web [5][23]. Considering similarity between the Deep Web and the X Web, it is highly feasible to take advantage of these proposals for adaptation to the X Web.

VII. Conclusions

UDDI registries provide a convenient starting point in identifying various categories. Current search engines like Google provide sampling estimates in the range of 80 to 85% for the determined categories, and these statistics provide us the basis for further investigation. Similarity in query form structures can help in the prediction of a new category from the form structure. Categories thus act as input data for a data extraction tool. Web services can also help customers in supply chain management. The paper presents an initial finding and inspiration point for further investigation into the composition of the future service-oriented Web.

References

[1] L. A. Adamic and B. A. Huberman, The Web's hidden order, CACM, 44(9), Sep. 2001, 55-60.

[2] M.-L. Antonie and O.R Zaiane, Text document categorization by term association, Proc. IEEE Int. Conf. on Data Mining, IEEE, Dec. 2002, 19–26.

[3] The Deep Web: Surfacing hidden value. Accessible at http://brightplanet.com, July 2000.

[4] K. C.-C. Chang, B. He, C. Li, M. Patel, Z. Zhang, Structured Databases on the Web: Observations and Implications, ACM SIGMOD Record, 33, Sep. 2004, 61–70.

[5] K. C.-C. Chang, B. He, and Z. Zhang, Toward Large Scale Integration: Building a MetaQuerier over Databases on the Web, Proc. 2d Conference on Innovative Data Systems Research (CIDR 2005), Asilomar, CA., Jan. 2005, 44-55.

[6] K. C.-C. Chang, B. He, Z. Zhang, Mining semantics for large scale integration on the Web: evidences, insights, and challenges, ACM SIGKDD Explorations Newsletter, 6(2), Dec. 2004, 67-76.

[7] X.-Y. Chen, Y. Chen; L. Wang, Y.-F. Hu, Text categorization based on frequent patterns with term frequency, Proc. 2004 Int. Conf. on Machine Learning and Cybernetics, Aug. 2004, 1610–1615.

[8] J. Fan, S. Kambhampati, Research articles and surveys: A snapshot of public Web services, SIGMOD Record, 34(1), Mar. 2005, 24-32.

[9] B. He, K. C.-C. Chang, and J. Han, Mining Complex Matchings across Web Query Interfaces, Proc. SIGMOD-DMKD’04, Paris, France, June 2004, 3-10.

[10] B. He and K. C.-C. Chang, Making Holistic Schema Matching Robust: An Ensemble Approach, Proc. KDD’05, Chicago, Illinois, Aug. 2005, 429-438.

[11] B. He, T. Tao, and K. C.-C. Chang, Organizing Structured Web Sources by Query Schemas: A Clustering Approach, Proc. 13th Conf. on Information and Knowledge Management (CIKM 2004), Washington, D.C., November 2004, 22-31.

[12] B. He, Z. Zhang, K. C.-C. Chang, Data integration: Knocking the door to the deep Web: integrating Web query interfaces, Proc. SIGMOD’04, Paris, France, June 2004, 913-914.

[13] H. He, W. Meng, C. Yu, and Z. Wu, “Automatic Integration of Web Search Interfaces with WISE-Integrator”, VLDB Journal, Vol.13, No.3, September 2004, pp. 256-273.

[14] Y. L. Hedley, M. Younas, A. James, M. Sanderson, Query-related data extraction of hidden Web documents, Proc. 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, Sheffield, UK, July 2004, 558–559.

[15] P. G. Ipeirotis, L. Gravano, M. Sahami, Probe, count, and classify: categorizing hidden Web databases, Proc. SIGMOD’01, May 2001, 67 – 78.

[16] J. Kong, K. Zhang, and X. Zeng, Spatial Graph Grammars for Graphical User Interfaces, ACM Trans. on Computer-Human Interaction, 2006 (in press).

[17] J. P. Lage, A. S. da Silva, P. B. Golgher, A. H. F. Laender, Web services and performance evaluation: Collecting hidden Web pages for data extraction, Proc. 4th Int. Workshop on Web Information and Data Management, McLean, Virginia, USA, Nov. 2002, 69–75.

[18] B. Liu, K. C.-C.-Chang, Editorial: special issue on Web content mining, ACM SIGKDD Explorations Newsletter, 6(2), Dec. 2004, 1–4.

[19] A. T. Manes, Web Services A Manager’s Guide, Publisher, Location, Date, Addison-Wesley, Boston, MA, USA, June 2003.

[20] E. Newcomer, Understanding Web Services, Addison-Wesley, Boston, MA, USA, May 2002.

[21] O’Reilly Webservices.xml.com. A Web Services Primer. Accessible at http://Webservices.xml.com, April 04, 2001.

[22] W. L. Oellermann Jr., Architecting Web Services, Apress, Berkeley, CA, USA, 2001.

[23] W. Wu, C. Yu, A. Doan, W. Meng, Research sessions: Web, XML and IR: An interactive clustering-based approach to integrating source query interfaces on the deep Web, Proc. SIGMOD’04, June 2004, 95-106.

[24] Z. Zhang, B. He, K. C.-C. Chang, Research sessions: Web, XML and IR: Understanding Web query interfaces: best-effort parsing with hidden syntax, Proc. SIGMOD’04, June 2004, 107-118.