Accessing the Deep Web: A Survey

Bin He, Mitesh Patel, Zhen Zhang, Kevin Chen-Chuan Chang

Computer Science Department

University of Illinois at Urbana-Champaign

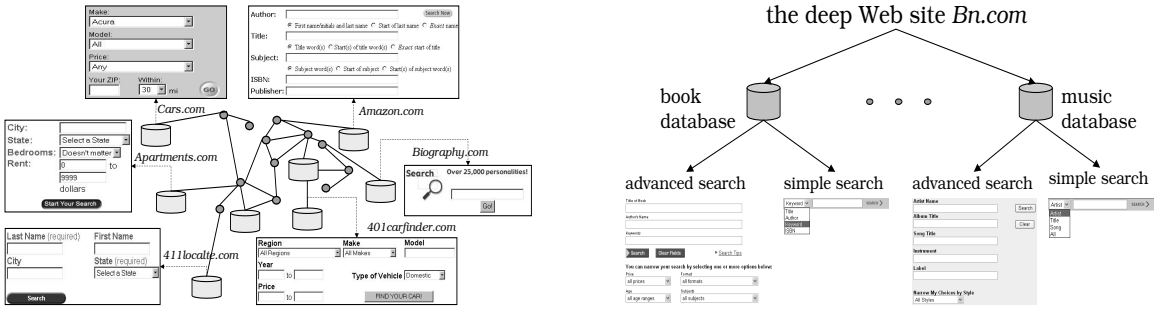
{binhe,mppatel2,zhang2,kcchang}@uiuc.edu

Introduction

Today's search engines do not reach most of the data on the Internet– The Web has been rapidly "deepened" by massive databases online [1]: While the *surface Web* has linked billions of static HTML pages, it is believed that a far more significant amount of information is "hidden" in the *deep Web*, behind the query forms of searchable databases, as Figure 1(a) conceptually illustrates. Such information may not be accessible through static URL links– They are assembled into Web pages as responses to queries submitted through the "query interface" of an underlying database. Because current search engines cannot effectively "crawl" databases, such data is believed to be "invisible," and thus remain largely "hidden" from users (thus often also referred to as the *invisible* or *hidden Web*.). Using overlap analysis between pairs of search engines, a July 2000 white paper [2] estimated 43,000-96,000 "deep Web sites" and an informal estimate of 7,500 terabytes of data– 500 times larger than the surface Web.

With its myriad databases and hidden content, this deep Web is an important yet largely-unexplored frontier for information search– While we have understood the surface Web relatively well, with various surveys (*e.g.*, [3, 4]), how is the deep Web different? This article reports our survey of the deep Web, studying the scale, subject distribution, search-engine coverage, and other access characteristics of online databases.

We note that, while the 2000 study [2] opens interest in this area, it focuses on only the scale aspect, and its result from *overlap analysis* tends to underestimate (as [2] also acknowledges). In overlap analysis, we need to estimate the



(a) The conceptual view of the deep Web. **(b)** Site, database, and interface.

Figure 1: The deep Web: Databases on the Web.

number of deep Web sites by exploiting two search engines. If we find n_a deep Web sites in the first search engine, n_b in the second and n_0 in both, we can estimate the total number as $\frac{n_a \times n_b}{n_0}$ by assuming that the two search engines randomly and independently obtain their data. However, as our survey found, search engines are highly correlated in their coverage of deep Web data (see question **Q5** in this survey). Therefore, such an independence assumption seems rather unrealistic, in which case the result is significantly underestimated. In fact, the violation of this assumption and its consequence were also discussed in the white paper itself [2].

Our survey took the *IP sampling* approach to collect random server samples for estimating the global scale as well as facilitating subsequent analysis. During April 2004, we have acquired and analyzed a random sample of Web servers by IP sampling. We randomly sampled 1,000,000 IPs (from the entire space of 2,230,124,544 valid IP addresses, after removing reserved and unused IP ranges according to [5]). For each IP, we used an HTTP client, the GNU free software wget [6], to make an HTTP connection to it and download HTML pages. We then identified and analyzed Web databases in this sample, in order to extrapolate our estimates of the deep Web.

Our survey distinguishes three related notions for accessing the deep Web– site, database, and interface: A *deep-Web site* is a Web server that provides information maintained in one or more back-end *Web databases*, each of which is searchable through one or more HTML forms as its *query interfaces*. For instance, as Figure 1(b) shows, *bn.com* is a deep-Web site, providing several Web databases (*e.g.*, a book database, a music database, among others) accessed via multiple query interfaces (*e.g.*, "simple search" and "advanced search"). Note that our definition of deep Web site did not account for the virtual hosting case, where multiple web sites can be hosted on the same physical IP address. Since identifying all the virtual hosts within an IP is rather difficult to conduct in practice, we do not consider such cases in our survey. Our IP sampling-based estimation is thus accurate modulo the effect of virtual hosting.

When conducting the survey, we first find the number of query interfaces for each web site, then the number of Web databases and finally the number of deep Web sites.

First, as our survey specifically focuses on online databases, we differentiate and exclude non-query HTML forms (which do not access back-end databases) from query interfaces. In particular, HTML forms for login, subscription, registration, polling, and message posting are not query interfaces. Similarly, we also exclude "site search," which many Web sites now provide for searching HTML pages on their sites– These pages are statically linked at the "surface" of the sites; they are not dynamically assembled from an underlying database. Note that, our survey considered only unique interfaces and removed duplicates– Many Web pages contain the same query interfaces repeatedly, *e.g.*, in *bn.com*, the simple book search in Figure 1(b) is present in almost all pages.

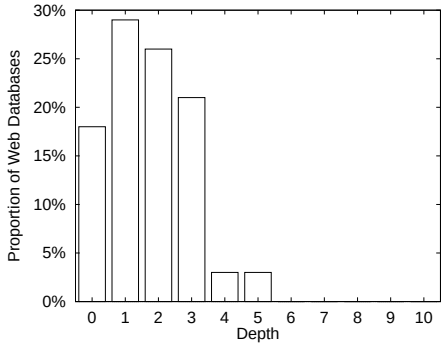
Second, we survey Web databases and deep Web sites based on the discovered query interfaces. Specifically, we compute the number of Web databases by finding the set of query interfaces (within a site) that refer to the same database. In particular, for any two query interfaces, we randomly choose 5 objects from one and search them in the other. We judge that the two interfaces are searching the same database if and only if the objects from one interface can always be found in the other one. Finally, the recognition of deep Web site is rather simple: A Web site is a deep Web site if it has at least one query interface.

Results

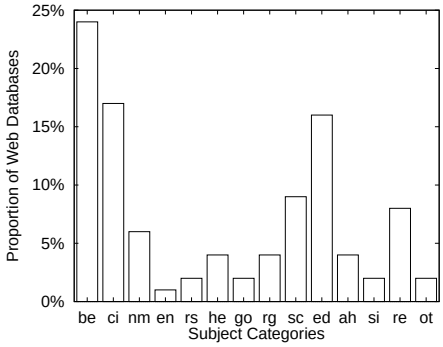
(Q1) ***Where to find "entrances" to databases?*** To access a Web database, we must first find its *entrances*– *i.e.*, query interfaces. How does an interface (if any) locate in a site– *i.e.*, at which depths? For each query interface, we measured the *depth* as the minimum number of hops from the root page of the site to the interface page[1]. As this study required deep crawling of Web sites, we analyzed $\frac{1}{10}$ of our total IP samples, *i.e.*, a subset of 100,000 IPs. We tested each IP, by making HTTP connections, and found 281 Web servers. Exhaustively crawling these servers to depth 10, we found 24 of them are deep Web sites, which contained a total of 129 query interfaces representing 34 Web databases.

We found that query interfaces tend to locate shallowly in their sites– None of the 129 query interfaces had depth deeper than 5. To begin with, 72% (93 out of 129) interfaces were found within depth 3. Further, since a Web database

[1]Such depth information is obtained by a simple revision of the wget software.



(a) Distribution of Web databases over depth.



(b) Distribution of Web databases over subject category.

Figure 2: Distribution of Web databases.

may be accessed through multiple interfaces, we measured its depth as the minimum depths of all its interfaces: 94% (*i.e.*, 32 out of 34) Web databases appeared within depth 3; Figure 2(a) reports the depth distribution of the 34 Web databases. Finally, 91.6% (*i.e.*, 22 out of 24) deep Web sites had their databases within depth 3. (We refer to these ratios as *depth-three coverage*, which will guide our further larger-scale crawling in **Q2**.)

(Q2) ***What is the scale of the deep Web?*** We then tested and analyzed all of the 1,000,000 IP samples to estimate the scale of the deep Web. As just identified, with the high depth-three coverage, almost all Web databases can be identified within depth 3– We thus crawled to depth 3 for these 1 million IPs.

The crawling found 2256 Web servers, among which we identified 126 deep Web sites, which contained a total of 406 query interfaces representing 190 Web databases. Extrapolating from the $s = 1,000,000$ unique IP samples to the entire IP space of $t = 2,230,124,544$ IPs, and accounting for the depth-three coverage, we estimate the number of deep Web sites as $126 \times \frac{t}{s} \div 91.6\% = 307,000$, the number of Web databases as $190 \times \frac{t}{s} \div 94\% = 450,000$, and the number of query interfaces as $406 \times \frac{t}{s} \div 72\% = 1,258,000$ (the results are rounded to 1000). The $2nd$ and $3rd$ columns of Table 1 summarize the sampling and the estimation results respectively. We also compute the confidence interval of each estimated number at 99% level of confidence, as the $4th$ column of Table 1 shows, which evidently indicates that the scale of the deep Web is well on the order of 10^5 sites. We also observed the "multiplicity" of access on the deep Web: In average, each deep Web site provides 1.5 databases, and each database supports 2.8 query interfaces.

The earlier survey of [2] estimated 43,000 to 96,000 deep Web sites by overlap analysis between pairs of search engines. Although the white paper has not explicitly qualified what it measured as a "search site," by comparison, it

	Sampling Results	*Total Estimate*	*99% Confidence Interval*
Deep Web sites	126	307,000	236,000 - 377,000
Web databases	190	450,000	366,000 - 535,000
– unstructured	*43*	*102,000*	62,000 - 142,000
– structured	*147*	*348,000*	275,000 - 423,000
Query interfaces	406	1,258,000	1,097,000 - 1,419,000

Table 1: Sampling and estimation of the deep-Web scale.

still indicates that our estimation of the scale of the deep Web, *i.e.*, on the order of 10^5 sites, is quite accurate. Further, it has been expanding, resulting in 3-7 times increase in 4 years (2000-2004).

(Q3) ***How "structured" is the deep Web?*** While information on the surface Web is mostly unstructured HTML text (and images), how is the nature of the deep Web data different? We classified Web databases into two types: 1) *unstructured* databases, which provide data objects as unstructured media (e.g., texts, images, audio, and video), and 2) *structured* databases, which provide data objects as structured "relational" records with attribute-value pairs. For instance, *cnn.com* has an unstructured database of news articles, while *amazon.com* has a structured database for books, which returns book records (*e.g.*, title = "gone with the wind", format = "paperback", price = $7.99).

By manual querying and inspection of the 190 Web databases sampled, we found 43 unstructured and 147 structured. We similarly estimate their total numbers to be 102,000 and 348,000 respectively, as Table 1 also summarizes. Thus, the deep Web features mostly structured data sources– with a dominating ratio of 3.4:1 versus unstructured sources.

(Q4) ***What is the subject distribution of Web databases?*** With respect to the top-level categories of the *yahoo.com* directory as our "taxonomy," we manually categorized the sampled 190 Web databases. Figure 2(b) shows the distribution of the 14 categories: Business & Economy (be), Computers & Internet (ci), News & Media (nm), Entertainment (en), Recreation & Sports (rs), Health (he), Government (go), Regional (rg), Society & Culture (sc), Education (ed), Arts & Humanities (ah), Science (si), Reference (re), and Others (ot).

The distribution indicates great subject diversity among Web databases, indicating that the emergence and proliferation of Web databases are spanning well across all subject domains. While there seems to be a common perception that the deep Web is driven and dominated by e-commerce (*e.g.*, for product search), our survey shows the contrary: To contrast, we further identify non-commerce categories from Figure 2(b)– he, go, rg, sc, ed, ah, si, re, ot– which together occupy 51% (97 out of 190 databases), leaving only a slight minority of 49% to the rest of commerce sites

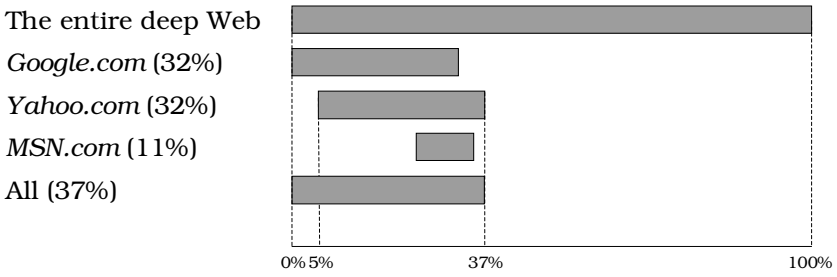


Figure 3: Coverage of search engines.

(broadly defined). In comparison, the subject distribution of the surface Web, as [3] characterized, showed that commerce sites dominated with an 83% share. Thus, the trend of "deepening" emerges not only across all areas, but also relatively more significantly in the non-commerce ones.

(Q5) ***How do search engines cover the deep Web?*** Since some deep Web sources also provide "browse" directories with URL links to reach the "hidden" content, how effective is it to "crawl-and-index" the deep Web as search engines do for the surface Web? We thus investigated how popular search engines "index" data on the deep Web. In particular, we chose the three largest search engines Google (*google.com*), Yahoo (*yahoo.com*) and MSN (*msn.com*).

We randomly selected 20 Web databases from the 190 in our sampling result. For each database, first, we manually sampled 5 objects (result pages) as test data, by querying the source with some random words. We then, for each object collected, queried every search engine to test whether the page was indexed, by formulating queries specifically matching the object page. (For instance, we used distinctive phrases that occurred in the object page as keywords and limited the search to only the source site.)

Figure 3 reports our finding: Google and Yahoo both indexed 32% of the deep Web objects, and MSN had the smallest coverage of 11%. However, there was significant overlap in what they covered– Together, the combined coverage of the three largest search engines increased only to 37%, indicating that they were indexing almost the same objects. In particular, as Figure 3 illustrates, Yahoo and Google overlapped on 27% objects of their 32% coverage– *i.e.*, a $\frac{27}{32} = 84\%$ overlap. Moreover, MSN's coverage was entirely a subset of Yahoo, and thus a 100% overlap.

The coverage results reveal some interesting phenomena: On one hand, contrast to the common perception, the deep Web is probably not inherently "hidden" or "invisible"– The major engines were able to each index one-third (32%) of the data. On the other hand, however, the coverage seems bounded by an intrinsic limit– Combined, these major engines covered only marginally more than they did individually, due to their significant overlap. This phenom-

	Number of Web Databases	*Coverage*
completeplanet.com	70,000	15.6%
lii.org	14,000	3.1%
turbo10.com	2,300	0.5%
invisible-web.net	1,000	0.2%

Table 2: Coverage of deep-Web directories.

enon clearly contrasts with the surface Web– where, as [3] reports, the overlap between engines is low, and combining them (or "metasearch") can greatly improve coverage. Here, for the deep Web, the fact that 63% objects were not indexed by any engines indicates certain inherent barriers for crawling and indexing data: 1) Most Web databases remain "invisible," providing no link-based access, and are thus not indexable by current crawling techniques. 2) Even when crawlable, Web databases are rather dynamic, and thus crawling cannot keep up with their updates promptly.

(Q6) ***What is the coverage of deep-Web directories?*** Besides traditional search engines, several deep-Web portal services have emerged online, providing deep-Web directories which classify Web databases in some taxonomies. To measure their coverage, we surveyed four popular deep-Web directories, as Table 2 summarizes. For each directory service, we recorded the number of databases it claimed to have indexed (on their Web sites). As a result, *completeplanet.com* was the largest such directory, with over 70,000 databases[2]. As Table 2 reports, compared to our estimate, it covered only 15.6% of the total 450,000 Web databases. However, other directories covered even less, in the mere range of $0.2\% - 3.1\%$. We believe this extremely low coverage suggests that, with their apparently manual classification of Web databases, such directory-based indexing services can hardly scale for the deep Web.

Conclusion

For further discussion, we summarize the findings of this survey for the deep Web in Table 3. We thus have the following conclusions:

First, while important for information search, the deep Web remains largely unexplored– At this point, it is neither well supported nor well understood. The poor coverage of both its data (by search engines) and databases (by directory services) suggests that access to the deep Web is not adequately supported. This survey seeks to understand better the deep Web. In some aspects, the deep Web does resemble the surface Web: It is large, fast growing, and diverse.

[2]However, we noticed that *completeplanet.com* also indexed "site search," which we have excluded; thus, its coverage could be overestimated.

Aspect	*Findings*
scale	The deep Web is of a large scale of 307,000 sites, 450,000 databases, and 1,258,000 interfaces. It has been rapidly expanding, with 3-7 times increase between 2000-2004.
diversity	The deep Web is diversely distributed across all subject areas– Although e-commerce is a main driving force, the trend of "deepening" emerges not only across all areas, but also relatively more significantly in the non-commerce ones.
structural complexity	Data sources on the deep Web are mostly structured, with a 3.4 ratio outnumbering unstructured sources, unlike the surface Web.
depth	Web databases tend to locate shallowly in their sites; the vast majority of 94% can be found at the top-3 levels.
search engine coverage	The deep Web is not entirely "hidden" from crawling– Major search engines cover about one-third of the data. However, there seems to be an intrinsic limit of coverage– Search engines combined cover roughly the same data, unlike the surface Web.
directory coverage	While some deep-Web directory services have started to index databases on the Web, their coverage is small, ranging from 0.2% to 15.6%.

Table 3: Summary of findings in our survey.

However, they differ in other aspects: The deep Web is more diversely distributed, is mostly structured, and suffers an inherent limitation of crawling.

Second, to support effective access to the deep Web, while the *crawl-and-index* techniques widely used in popular search engines today have been quite successful for the surface Web, such an access model may not be appropriate for the deep Web. To begin with, crawling will likely face the limit of coverage, which seems intrinsic because of the hidden and dynamic nature of Web databases. Further, indexing the crawled data will likely face the barrier of structural heterogeneity across the wide range of deep Web data: The current keyword-based indexing (which all search engines do), while serving the surface-Web pages well, will miss the schematic "structure" available in most Web databases– Imagine one can only search for flight tickets by keywords, but not destinations, dates, and prices!

Third, as traditional access techniques may not be appropriate for the deep Web, it is crucial to develop more effective techniques. To look ahead, we speculate that the deep Web will likely be better served with a *database-centered, discover-and-forward* access model. A search engine will automatically discover *databases* from the Web, by crawling and indexing their query interfaces (and not their data pages). Upon user querying, the search engine will forward users to the right databases for the actual search of data. Querying at the databases will use their data-specific interfaces and thus fully leverage their structures– *i.e.*, we can query flights with the desired attributes! Several recent research projects, *e.g.*, MetaQuerier [7] and WISE-Integrator [8], are exploring this exciting direction.

References

[1] Thanaa M. Ghanem and Walid G. Aref. Databases deepen the web. *IEEE Computer*, 73(1):116–117, 2004.

[2] BrightPlanet.com. The deep web: Surfacing hidden value. Accessible at `http://brightplanet.com`, July 2000.

[3] Steve Lawrence and C. Lee Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, 1999.

[4] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th International World Wide Web Conference*, pages 669–678, 2004.

[5] Ed O'Neill, Brian Lavoie, and Rick Bennett. Web characterization. Accessible at "`http://wcp.oclc.org`".

[6] GNU. wget. Accessible at "`http://www.gnu.org/software/wget/wget.html`".

[7] Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. Toward large scale integration: Building a metaquerier over databases on the web. In *CIDR 2005 Conference*, 2005.

[8] Hai He, Weiyi Meng, Clement Yu, and Zonghuan Wu. Wise-integrator: An automatic integrator of web search interfaces for e-commerce. In *Proceedings of the 29th VLDB Conference*, 2003.