Extracting Data behind Web Forms

Stephen W. Liddle*, David W. Embley*, Del T. Scott, and Sai Ho Yau

Brigham Young University, Provo, UT 84602, USA
{liddle,scottd}@byu.edu, {embley,yaus}@cs.byu.edu

Abstract. A significant and ever-increasing amount of data is accessible only by filling out HTML forms to query an underlying Web data source. While this is most welcome from a user perspective (queries are relatively easy and precise) and from a data management perspective (static pages need not be maintained and databases can be accessed directly), automated agents must face the challenge of obtaining the data behind forms. In principle an agent can obtain all the data behind a form by multiple submissions of the form filled out in all possible ways, but efficiency concerns lead us to consider alternatives. We investigate these alternatives and show that we can estimate the amount of remaining data (if any) after a small number of submissions and that we can heuristically select a reasonably minimal number of submissions to maximize the coverage of the data. Experimental results show that these statistical predictions are appropriate and useful.

1 Introduction

To help consumers and providers manage the huge quantities of information on the World Wide Web, it is becoming increasingly common to use databases to generate Web pages dynamically. Unlike ordinary Web pages mapped to standard URLs, information in this "deep Web" [1] or "hidden Web" [6,15] is not accessible through regular HTTP GET requests by merely specifying a URL and receiving a referenced page in response. Most commonly, data in the hidden Web is stored in a database and is accessible by issuing queries guided by HTML forms.

Automated extraction of data behind form interfaces is desirable when we wish to have automated agents search for desired information, when we wish to wrap a site for higher level queries, and when we wish to extract and integrate information from different sites. How can we automatically access this information? We began to answer this question in an earlier paper where we outlined our approach to automated form filling [10]. As our contribution in this paper we statistically streamline the process to make the information gathering process efficient, and improve our analysis of returned results. In the broader context of our project [7,5], information gathered during form processing will later be handed to a downstream data extraction process.

* Supported in part by the National Science Foundation under grant IIS-0083127.

A. Olivé et al. (Eds.): ER 2002 Ws, LNCS 2784, pp. 402–413, 2003.

1.1 Related Work

Others have also studied the problem of automatically filling out Web forms. Most common are tools designed to make it easier for an end user to fill out a form. Commercial services exist, for example, to provide information from a limited portfolio of user-specified information such as name, address, contact information, and credit card information [3,4,12]. These services, such as the Microsoft Passport and Wallet system, encrypt a user's personal information and then automatically fill in Web forms when fields can be recognized. Since many forms share common attributes (especially in the domain of e-commerce transactions), these tools can reliably assist users in entering personal information into Web forms.

One of the earliest efforts at automated form filling was the ShopBot project [2], which uses domain-specific heuristics to fill out forms for the purpose of comparison shopping. The ShopBot project, however, did not propose a general-purpose mechanism for filling in forms for non-shopping domains.

The most closely related work to our own is the Hidden Web Exposer (HiWE) project at Stanford [14,15]. HiWE provides a way to extend crawlers beyond the publicly indexable Web by giving them the capability to fill out Web forms automatically. Because of the formidable challenges to a fully automatic process, HiWE assumes that crawls will be domain specific and human assisted (we also rely on human assistance at key points, but we do not use domain specific information in retrieving data from a particular site). Although HiWE must start with a user-provided description of the search task, HiWE learns from successfully extracted information and updates the task description database as it crawls. Besides an operational model of a hidden Web crawler, a significant contribution of this work is the label matching approach used to identify elements in a form based on layout position, rather than proximity within the underlying HTML code. These researchers also present the details of several ranking heuristics together with metrics and experimental results that help evaluate the quality of the proposed process. The independently-developed details of our approach are complementary to HiWE. For example, we consider the task of duplicate record elimination and we use a statistics-based sampling approach to efficiently decide when a particular source has been sufficiently extracted.

1.2 Overview

In the remainder of the paper we describe the details of our contribution. We have created a prototype tool that automatically retrieves the data behind a particular HTML form. We refer the reader to [10] and [9] for details on our tool and approach. After briefly summarizing our approach, we discuss our form-submission plan in Section 2. In particular, we discuss the use of statistical procedures as a means to intelligently cover the search space and to determine how much of the available data has likely been retrieved and how close we are to completion. Section 3 presents the results of experiments we conducted to verify our approach to automatically extracting data behind Web forms. We

summarize, report on our implementation status, and give plans for future work in Section 4

1.3 Summary of Our Approach

As described in [10], our approach to automated form filling starts with parsing the HTML page containing the target form. Next we construct an HTTP request that is equivalent to "filling in the form manually." This request is issued to the host Web server, and our tool analyzes the result page. If the result page contains links to other pages of results (e.g. the result displays, say, records 1–10 of 53), the tool follows the required links to retrieve all the result records. Often, it is the case that the default query (which consists of using the form's default values for all fields) returns all the records in the underlying database. But in most cases, we have to submit a variety of different queries in order to retrieve all the records. The same underlying records might be returned in response to different queries. So each time we add the results of a query to the accumulated extracted data, we first eliminate any duplicate records. The tool also deals intelligently with error pages and unexpected results.

2 Form Submission Plan

Our goal is to retrieve all the data within the scope of a particular Web form. One way to do this is to fill in the form in all possible ways. Ignoring the issue of fields with unbounded domains (i.e., text boxes), there are still two problems with this strategy. First, the process may be time consuming. Second, we may have retrieved all (or at least a significant percentage) of the data before submitting all the queries. Many forms have a default query that obtains all data available from the site. If the default query does not yield all data, it is still likely that we can extract a sufficient percentage of the data without exhaustively trying all possible queries.

Our strategy involves several phases: (1) issue the default query, (2) sample the site to determine whether the default query response is likely to be comprehensive, and (3) exhaustively query until we reach a limiting threshold. In the exhaustive phase, we can often save considerable effort by using limiting thresholds. For example, we can estimate the size of the database behind a form, and then continue issuing queries until we have reached a certain percentage of completeness. The user can specify several thresholds:

- *Percentage of data retrieved*: What percentage of the estimated data has actually been retrieved so far? Typical values for this threshold might be 80%, 90%, 95%, or 99%. This threshold controls the quality of the crawl.
- *Number of queries issued*: How many total queries have been issued to this site? It may be prudent to limit the burden placed on individual sites by terminating a crawl after a particular number of queries. This threshold controls the burden placed on crawled sites.

- *Number of bytes retrieved*: How many bytes of unique data have been retrieved so far?
- *Amount of time spent*: How much total time has been spent crawling this site? This threshold and the previous one control the resources required on the crawler side to support the crawl.
- *Number of consecutive empty queries*: How many consecutive queries have returned no new data? The probability of encountering new data goes down significantly as the number of consecutive empty queries goes up.

Each of these thresholds constitutes a *sequential stopping rule* that can terminate the crawl before trying all possible queries (by "all" we mean all combinations of choices for fields with bounded domains — we exclude fields with unbounded domains in this study).

2.1 Sampling Phase

Earlier we described the method for determining and issuing the default query. We now discuss the sampling procedure used to determine whether the default query is likely to have returned all the data behind a particular form.

There are several parameters of interest to us. First, we characterize each form field as a "factor" in our search space. Let $f_1, f_2, ...f_n$ be the n factors corresponding to fields with bounded domains, and let $|f_i|$ represent the number of choices for the i^{th} factor. Then the total number of possible combinations N for this form is:

$$N = \prod_{i=1}^{n} |f_i|.$$

We are also interested in the cardinality c of the largest factor: $c = \max(|f_1|, |f_2|, ..., |f_n|)$.

Next, we define C to be the size of a sampling batch. We want each sampling batch to be large enough to cover the margins of our sample space — that is, we want to have fair coverage over all the factors. So we let $C = \max(c, \lceil \log_2 N \rceil)$. This accounts for the case where there are many factors of small cardinality. For example, if there were 16 factors each of cardinality 2, then $N = 2^{16} = 65536$, but $c = 2$. We want c to be representative of the size of our search space, so we require that it be at least $\log_2 N$, which is a statistically reasonable number to use when sampling populations of known size (consider, for example, the 2^k factorial experiments method [16,11]).

Our decision rule for determining whether the default query returned all the data available from a particular site is based on a sample of C queries. If all C queries return no additional data (i.e. after duplicates have been eliminated), then we assume the default query did indeed retrieve all available data.

However, we need to be careful about how we choose the C queries. Suppose the form of interest contains two bounded fields with choices of 7 and 4 possibilities respectively. Then we have $N = 7 \times 4$ and $C = \max(\max(4, 7), \lceil \log_2 28 \rceil) = \max(7, 5) = 7$. If we simply choose C random queries, we might end up with

a sample set like the one shown in Table 1. This table shows a two-way layout [16,8,11,13] that helps us use a two-factor method [16,11,13] to choose a query sample. Notice that a_6, a_7, and b_3 were not considered in any of the sample queries, while b_1 is oversampled.

One solution is to keep track of how many times we have sampled each factor, and spread the samples evenly, as Table 2 shows. This approach for constructing a sampling search pattern yields "maximal coverage."

In Table 2 we use a regular pattern to cover both factors as broadly as possible. The sample consists of the sequence (a_1, b_1), (a_2, b_2), (a_3, b_3), (a_4, b_4), (a_5, b_1), (a_6, b_2), (a_7, b_3). If we were to continue, the next query we would choose would be (a_1, b_4). The algorithm chooses a next sample that is as far away from all previous samples as possible. Since we have categorical, not quantitative, data each of the a_i choices for factor A is equally distant from all others. Thus, our distance function simply measures the number of coordinates that are different. For example, the distance between (a_1, b_1) and (a_2, b_2) is 2, while the distance between (a_1, b_1) and (a_7, b_1) is 1.

To ensure that a regular pattern does not bias our results, we introduced a stochastic element by randomly choosing next samples from the list of all those that are equally furthest from the set selected so far. This yields a layout like the one in Table 3. Note that our technique is general for n dimensions, $n \geq 2$.

Table 1. Two-Way Layout with Random Sampling

		Factor A						
		a_1	a_2	a_3	a_4	a_5	a_6	a_7
Factor B	b_1	x	x		x			
	b_2		x		x			
	b_3							
	b_4			x		x		

Table 2. Regular Sampling with Maximal Coverage

		Factor A						
		a_1	a_2	a_3	a_4	a_5	a_6	a_7
Factor B	b_1	x				x		
	b_2		x				x	
	b_3			x				x
	b_4				x			

Table 3. Random Sampling with Maximal Coverage

		Factor A						
		a_1	a_2	a_3	a_4	a_5	a_6	a_7
Factor B	b_1				x			x
	b_2	x					x	
	b_3					x		
	b_4		x	x				

After issuing C queries stochastically in a maximally covering fashion, if we have received no new data we judge the default query to be sufficient and halt,

claiming that we have successfully retrieved all data. In practice, this rule has been highly successful (as judged by human operators manually verifying the decision). Although this rule never reported that all data had been retrieved when it had not, we did encounter sites where our rule was too strict. (The copy detection system reported new bytes in subsequent queries, but there were not really any new records; this can be the result of personalized advertising information or other dynamic variations in Web pages).

Our decision rule for determining whether a query yields "new" data is based on a size heuristic. First we strip a returned page of its HTML tags, leaving on the special sentence boundary tags (<s.>) described in [10]. Then we use the copy detection system to remove duplicate sentences, and count the number of bytes U remaining. If U is at least 1000, we assume we have received new data. If U is less than 1000, we use order analysis to prune small results. We first find the minimum size M returned by the previous C queries. If U is greater than M then we assume we have received new data. We have also tested a percentage threshold, comparing U to the total number of bytes returned in the page (excluding HTML tags). Both heuristics perform reasonably well. Since we are using a statistical approach, some noise in the system is acceptable, and indeed some even cancels itself out.

Another way to test for uniqueness of data would be to call on our downstream record extraction process to actually extract and structure the records from this page. Then we could perform database-level record comparisons to determine when we have found unique records. However, this makes the Web-form retrieval less general by tying it to a specific domain ontology. In the future we will investigate more fully the ramifications of this alternate approach.

2.2 Exhaustive Phase

If the C sample queries yield new data, we proceed by sampling additional batches of C queries at a time, until we reach one of the user-specified thresholds or we exhaust all the possible combinations. Sampling proceeds according to the maximal covering algorithm discussed earlier — each new query is guaranteed to be as far away from previous queries as possible, thus maximizing our coverage of the factors.

However, before proceeding, we first estimate and report the maximum possible space needed for storing the results and the maximum remaining time needed to finish the process. We also give the user the choice of specifying the various thresholds for completeness of retrieved data, maximum number of queries to be issued, maximum storage space to use, and maximum time to take. Additionally, the user can decide to stop and use only the information already obtained. (All this information can also be specified ahead of time so the process can run unattended.)

We estimate the maximum space requirement S by multiplying the total number of queries N by the average of the space needed for data retrieved from

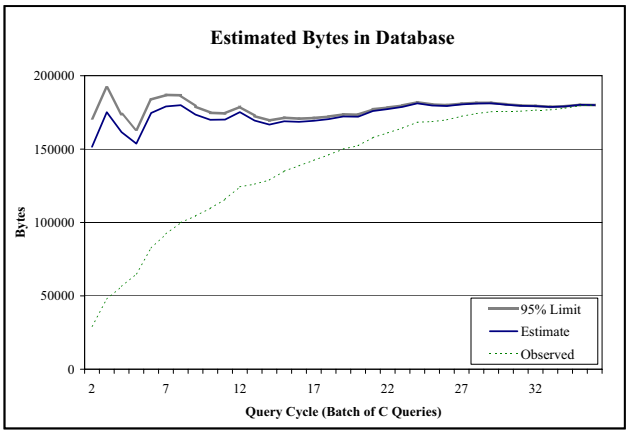

Fig. 1. Completeness Measures for Automobile Ads Web Site (`www.slc-classifieds.com`, Form 1)

m sample queries:

$$S = \left(\frac{N}{m}\right) \sum_{i=1}^{m} b_i$$

where b_i is the size in bytes of the i^{th} sample query, and typically $m = C$. We estimate the remaining time required T similarly:

$$T = \left(\frac{N}{m}\right) \sum_{i=1}^{m} t_i - \sum_{i=1}^{m} t_i = \left(\frac{N-m}{m}\right) \sum_{i=1}^{m} t_i$$

where t_i is the total duration of the i^{th} sample query. Note that we subtract the time already spent in the initial sampling phase. Also observe that since we follow "next" links, what we are calling a "query" could actually involve a fair number of HTTP GET requests. For larger sites, this can take a significant amount of time.

After establishing the user's preferences, we proceed to process additional batches of C query samples. At the end of each sample batch, we test the thresholds to see if it would be productive to continue processing. Note that each batch provides maximal coverage of the various factors, using unique combinations that have not yet been tried.

Figure 1 illustrates how we measure our progress with respect to the percentage of information retrieved so far. The "Observed" line indicates how many actual unique bytes we have seen after each query cycle. The "Estimate" line shows our estimate of the number of unique bytes we would encounter if we were to exhaustively crawl this Web site (trying all possible combinations of queries). The "95% Limit" line uses the standard deviation of our prediction estimate for the probability of finding additional data to determine the level at which we can claim 95% confidence about our estimate. That is, based on what we have seen so far, we are 95% confident that the real number is less than the "95% Limit"

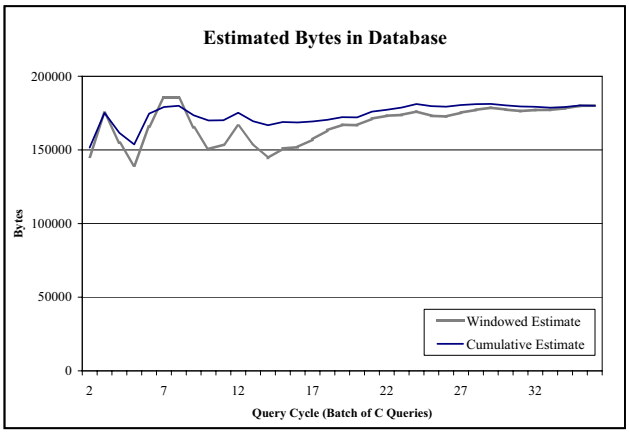

Fig. 2. Windowed vs. Cumulative Estimates

number. In this example, we cross the 80% completeness threshold in cycle 14 (out of 36 total). After 22 cycles we estimate that we are 90% complete. We reach 95% in cycle 25, and 99% in cycle 31. Only after all 36 cycles have been exhaustively attempted would we determine that we are 100% complete. In this example, $N = 2124$ and $C = 59$. Thus, if 80% completeness is sufficient for the user, we prune 1297 out of 2124 queries (61%) from the list.

The formula for estimating the database size D_i after i queries is shown in Equation 1:

$$D_i = O_i \left(1 + \frac{N-i}{i} p_i\right) \tag{1}$$

where O_i is the number of unique bytes observed after i queries, and p_i is the estimate of the probability of finding new data in query $i+1$. The only random variable in Equation 1 is p_i, which is defined as the number of queries that returned new data divided by i. Thus, p_i is a cumulative probability estimate reflecting the ratio of successful queries to total queries. We compute D_i by predicting that in the remaining $N-i$ queries the proportion that return new data will be approximately p_i. Further, we estimate that on average successful queries will return approximately $\frac{O_i}{i}$ new bytes. We can compute $\hat{D}_i$, that is, D_i with 95% confidence by including in p_i a measure of the standard deviation of p_i over the previous two query cycles, σ_i, as shown in Equation 2:

$$\hat{D}_i = O_i \left(1 + \frac{N-i}{i}(p_i + 1.645\sigma_i)\right) \tag{2}$$

Rather than using a cumulative estimate of database size, we could instead use a window to ignore older data points. For example, Figure 2 compares a cumulative estimate of D_i with a windowed estimate, where the probability p_i is only computed using the last two query cycles. One might think that a windowed probability estimate should be more accurate because it reflects more recent experience. But in practice the more stable cumulative estimate appears

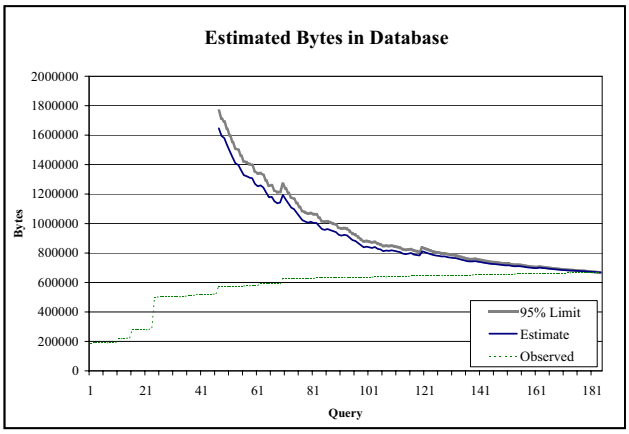

Fig. 3. A Contrasting Automobile Search (`www.slc-classifieds.com`, Form 2)

to give better performance among the sites we have examined. Both converge relatively quickly and show nicely declining variance over time.

3 Experimental Results

A considerable amount of our work was done by simulation, crawling a site once and then playing with different sampling scenarios to understand the ramifications of various approaches. However, we also tried our tool on 15 different Web sites from several different application domains to evaluate its performance empirically. In all cases, we manually verified the decision reported by the system regarding whether the default query retrieved all data. In six of the cases, the default query did indeed return all the data.

In our testing, processing a single HTTP request took anywhere from 2 to 25 seconds on average, depending on the Web site. A single query, which includes following "next" links, averaged between 5 seconds and 14 minutes. Some pages had no "next" links, but others had as many as 140 such links! Thus, the sampling phase could take anywhere from a few dozen seconds to several hours. Storage requirements were modest by modern standards, requiring anywhere from several megabytes to hundreds of megabytes per tested site.

We encountered two typical data patterns. First is the relatively sparse behavior exhibited by the site crawled in Figure 1. The second is fairly dense, as typified by Figure 3. Both of these data sets come from the same Web site, but through different forms to different sub-portions of the site. In this second example, there are only 8 query cycles of 23 queries each (8 choices for one bounded field, 23 for another), for a total of $N = 184$. However, it is not until the end of cycle 6 (query 138) that we estimate we have retrieved 80% of the available data. And even then, there is still an estimated 42% probability of obtaining additional data with another query. Thus, we can only save a small portion of

the 184 queries (25% or fewer). Even in such cases, however, the 15-25% savings can be significant.

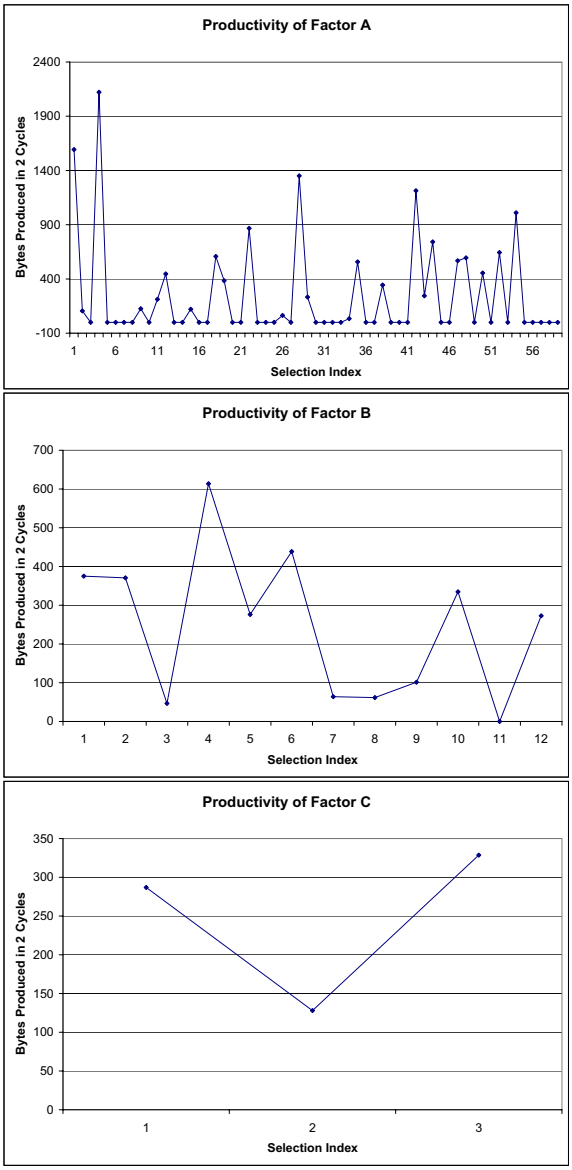

Fig. 4. Relative Productivity of Various Factors

4 Conclusion

In this paper we have described our domain-independent approach for automatically retrieving the data behind a given Web form. We have prototyped a synergistic tool that brings the user into the process when an automatic decision is hard to make. We use a two-phase approach to gathering data: first we sample the responses from the Web site of interest, and then, if necessary, we methodically try all possible queries until either we believe we have arrived at a fixpoint of retrieved data, or we have reached some other stopping threshold, or we have exhausted all possible queries.

We have created a prototype (mostly in Java, but also using JavaScript, PHP, and Perl) to test our ideas, and the initial results are encouraging. We have been successful with a number of Web sites, and we are continuing to study ways to improve our tool.

One improvement that seems particularly promising is to perform an analysis of the productivity of various factors in order to emphasize those that yield more data earlier in the search process. For example, Figure 4 shows that choices 1, 4, and 28 of Factor A have been significantly more productive than average in the first two query cycles. Similarly, choices 4 and 6 of Factor B, and choices 1 and 3 of Factor C are highly productive. A straightforward two-way interaction analysis indicates which pairs of choices for various factors are most productive. There are challenging issues with determining how best to partition the search space, but we are studying how a directed search algorithm might perform.

Our research group is generally working in the broader context of ontology-based data extraction and information integration. If we combine the hidden Web retrieval problem with the tool of domain-specific ontologies, we could automatically fill in text boxes with values from the ontologies. While this makes the retrieval process task-specific, it also increases the likelihood of being able to extract just the relevant subset of data at a particular Web site. (Our assumption in this paper has been that the user wants to retrieve all or most of the data at a given site.) Also, it is likely that we could avoid a fair amount of the manual intervention in our current process if we use ontologies.

Users need and want better access to the hidden Web. We believe this will be an increasingly important and fertile area to explore. This paper represents a step in that direction, but a great deal remains to be done. We look forward to continuing this promising line of research.

References

1. M.K. Bergman. *The Deep Web: Surfacing Hidden Value*. BrightPlanet.com, July 2000. Downloadable from `http://www.brightplanet.com/deep_content/deepwebwhitepaper.pdf`, checked August 10, 2001.
2. R.B. Doorenbos, O. Etzioni, and D.S. Weld. A scalable comparison-shopping agent for the World-Wide Web. In *Proceedings of the First International Confence on Autonomous Agents*, pages 39–48, Marina del Rey, California, February 1997.

3. Patil systems home page. `http://www.patils.com`. Describes LiveFORM and ebCARD services. Checked August 10, 2001.
4. eCode.com home page. `http://www.eCode.com`. Checked August 10, 2001.
5. D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith. Conceptual-model-based data extraction from multiple-record Web pages. *Data and Knowledge Engineering*, 31:227–251, 1999.
6. D. Florescu, A.Y. Levy, and A.O. Mendelzon. Database techniques for the World-Wide Web: A survey. *SIGMOD Record*, 27(3):59–74, 1998.
7. Home Page for BYU Data Extraction Group, 2000.
URL: `http://www.deg.byu.edu`.
8. T. Leonard. *A Course In Categorical Data Analysis*. Chapman & Hall/CRC, New York, 2000.
9. S.W. Liddle, D.W. Embley, D.T. Scott, and S.H. Yau. Extracting data behind Web forms. Technical report, Brigham Young University, June 2002. Available at `http://www.deg.byu.edu/papers/`.
10. S.W. Liddle, S.H. Yao, and D.W. Embley. On the automatic extraction of data from the hidden web. In *Proceedings of the International Workshop on Data Semantics in Web Information Systems (DASWIS-2001)*, pages 106–119, Yokohama, Japan, November 2001.
11. R.A. McLean and V.L. Anderson. *Applied Factorial and Fractional Designs*. Marcel Dekker, Inc., New York, 1984.
12. Microsoft Passport and Wallet services. `http://memberservices.passport.com`. Checked August 10, 2001.
13. R.L. Plackett. *The Analysis of Categorical Data,* 2^{nd} *Edition*. Charles Griffin & Company Ltd., London, 1981.
14. S. Raghavan and H. Garcia-Molina. Crawling the hidden Web. Technical Report 2000-36, Computer Science Department, Stanford University, December 2000. Available at `http://dbpubs.stanford.edu/pub/2000-36`.
15. S. Raghavan and H. Garcia-Molina. Crawling the hidden Web. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB 2001)*, Rome, Italy, September 2001.
16. A.C. Tamhane and D.D. Dunlop. *Statistics and Data Analysis: From Elementary to Intermediate*. Prentice-Hall, New Jersey, 2000.