Adaptive Record Extraction From Web Pages[*]

Justin Park
University of Calgary
2500 University DR NW
Calgary, AB, Canada
parkj@cpsc.ucalgary.ca

Denilson Barbosa
University of Calgary
2500 University DR NW
Calgary, AB, Canada
denilson@ucalgary.ca

ABSTRACT

We describe an adaptive method for extracting records from web pages. Our algorithm combines a weighted tree matching metric with clustering for obtaining data extraction patterns. We compare our method experimentally to the state-of-the-art, and show that our approach is very competitive for rigidly-structured records (such as product descriptions) and far superior for loosely-structured records. (such as entries on blogs).

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Textual Databases; H.3.3 [**Information Search and Retrieval**]: Clustering

General Terms

Algorithms, Experimentation.

1. INTRODUCTION

A substantial fraction of the web consists of the so-called *deep web*: pages that are dynamically generated using predefined templates populated with data from databases. Because these databases are maintained by organizations with vested interests in their accuracy and usefulness, the information in deep web pages tends to be of very high-quality. However, deep web sites are intended for human consumption, much like other web sites, and do not provide access to their data to computer applications. Recently, there has been considerable interest in building automatic tools capable of extracting data from disparate web sites and representing them in a form amenable to processing by other applications [2, 3, 4, 6, 8]. In particular, there has been considerable work based on using the tree-edit distance [7] metric as a basis for finding data extraction patterns.

Most of the previous tools are tailored to specific web sites. The MDR [4] and Depta [6] approaches aim at extracting product listings or data displayed in tabular form; Zhao et al. [8] focus on extracting entries from result pages from search engines; the News Extractor by Reis et al. [5] only extracts news articles. A major shortcoming of these works is that they use tree-edit distance strictly. Doing so works well on strictly structured records, such as product descriptions, but not so well on loosely structured records, such as blog entries.

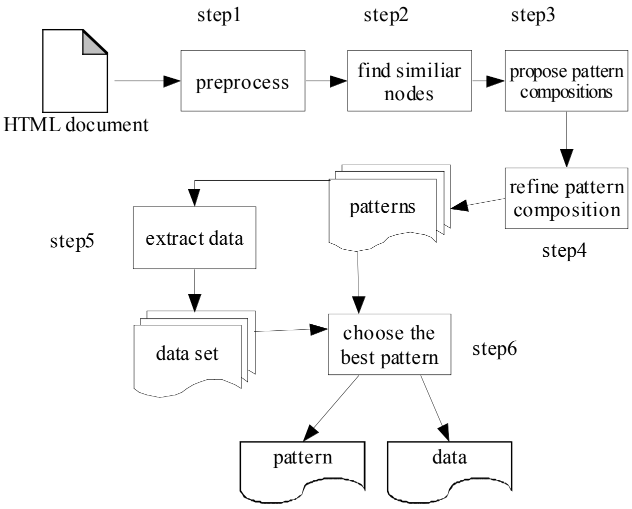

Figure 1: Overview of our method.

In this paper, we describe a general purpose web data extractor, which we call WDE, for simplicity, that performs well in both rigidly and loosely structured records in an HTML document. We validate our tool on several kinds of web pages, including product listings, search engine results pages, sports scoreboards, forums, and blogs.

2. OUR METHOD

Figure 1 gives an overview of our data extractor. Given the URL of a web page, WDE automatically discovers all repeating patterns found in the page, as follows. Initially, we use a standard tool for *tidying* the HTML page[1]; our preprocessing step also merges all non-structural HTML tags (e.g., those specifying fonts, colors, etc.). The result of this step is a tree representation of the web page.

The next step is identifying clusters of tree nodes with similar structure. As in previous work, we compute the similarity between two nodes in the tree by a tree matching algorithm that approximates the true tree-edit distance [7] value between those nodes. (This is due to the high cost in finding the actual tree-edit distances between all pairs of

[*]This work was funded in part by grants from the Natural Sciences and Engineering Council of Canada and the Alberta Ingenuity Fund.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

[1] `http://people.apache.org/~andyc/neko/doc/html`.

	WDE			MDR		
Findgift.com	10	0	0	10	0	0
Ebay.com	50	0	0	50	0	0
Pricerunner.com	25	0	0	25	0	0
Backcountry.com	31	1	0	32	0	0
Download.com	9	1	0	9	1	0
Shopping.yahoo.com	15	0	0	11	0	4
Radioshack.com	0	17	0	17	3	0
Nextag.com	15	0	0	15	0	0
Indio.ca	10	0	0	0	10	0
Dealtime.com	14	1	0	14	1	0
del.icio.us	10	0	0	0	0	0
Barnsandnoble.com	10	0	0	0	0	10
Youtube.com	20	0	0	20	0	0
Imdb.com	8	4	0	0	0	12
allrecipes.com	20	0	0	20	0	0
foodtv.ca	10	0	0	10	0	0
weblog.xanga.com	10	0	0	0	10	0
nhl.com	14	0	0	0	13	1
calgarypubliclibrary.com	20	0	0	0	20	0
mls.ca	0	10	0	0	0	10
Average Recall	89.9%			80.1%		
Average Precision	100%			86.3%		

Table 1: Accuracy on product listings.

nodes in the tree.) Unlike in previous work [6], we assign different weights to nodes depending on their height (internal nodes get higher weights).

In the third step, we enumerate all possible *candidate* records in the page. Candidate records can have one or more nodes, but cannot have more than one node from the same cluster. Also, we assume that records do not overlap. The fourth step computes the similarity of all pairs of records identified in the previous step. This is done as follows. Each record is converted into a tree whose root is a *dummy* node containing all nodes in that record as children. The similarity between two records is computed similarly as in step 2. Finally, we cluster the records based on their similarity.

From the clusters obtained after step 4, we generate data extraction patterns that navigate the HTML tree and extract the actual data that form the records. At the time of writing, we report to the user all patterns extracted from clusters with high pairwise similarity among records. Our future work consists of finding automatic ways of evaluating these patterns (step 6).

3. EXPERIMENTAL EVALUATION

We compared our tool to the MDR [4] method, which is the state-of-the-art in web data extraction based on tree edit distance. We used 30 web sites in our comparison; 20 of these sites contain product listings, usually organized in tabular format; the other sites contain user comments, similar to discussion groups or blogs. We use the standard notions of precision and recall [1] to evaluate our tool. Recall is defined as the percentage of the intended data records that are retrieved by the tool; precision is defined as the percentage of the returned data records that are correct. We determine correctness (i.e., precision) manually, on a best-effort basis.

On average, our tool returned less than 5 patterns per site. We chose the pattern that best identified the records for the comparison below after a quick visual inspection. (Almost always, the chosen pattern was the one producing the largest number of records.) Tables 1 and 2 show the results of our experiments. The tables show three values for each web site and each method: the number of *correct* records retrieved, the number of *missed* records, and the number of *false positives*, in this order. The tables also show the average precision and recall for each method on all web sites in each group.

	WDE			MDR		
forums.gentoo.org	25	0	0	25	0	0
forum.java.sun.com	9	2	0	0	11	0
Youtube.com	10	0	0	0	10	0
operawatch.com	5	0	0	0	5	0
shoutwire.com	7	3	0	0	10	0
engadget.com	7	0	0	0	7	0
gizmodo.com	6	0	0	6	0	19
thinkprogress.org	42	19	0	0	61	0
discussion.forum.nokia.com	15	0	0	15	0	0
messages.yahoo.com	18	0	0	18	0	1
Average Recall	85.7%			38.1%		
Average Precision	100%			76.2%		

Table 2: Accuracy on discussion forums.

On product listings (Table 1), WDE showed very high precision, although it sometimes missed a few records. The reason for this is that in the refining step, outlier records that are different than the more common ones are eliminated. On the other hand, this refining step also eliminated all false positives, thus resulting 100% in precision. While WDE is very competitive with MDR with strictly structured records, it showed drastically superior performance on our second test, with loosely structured records (Table 2).

4. CONCLUSION

We proposed a novel method for extracting records from web pages based on and adaptive, weighted tree matching algorithm and the clustering of similar records. Unlike previous methods, our method performs well for both strictly and loosely structured records, as confirmed by our comprehensive experimental evalutaion. Our future work consists of studying criteria for ranking patterns, and experimenting with other kinds of records.

5. REFERENCES

[1] R. Baeza-Yates, and B. Ribeiro-Neto. Modern Information Retrieval. Addison-Welsey, 1999.

[2] V. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In *VLDB* 2001: p. 109–118.

[3] A. Laender, A. da Silva, B. Ribeiro-Neto, and J. Teixeira. A Brief Survey of Web Data Extraction Tools. *SIGMOD Record* 31(2): p. 84–93.

[4] B. Liu , R. Grossman , Y. Zhai. Mining Data Records in Web Pages. In *KDD* 2003: pg. 601–606.

[5] D. Reis, P. Golgher, A. Silva, and A. Laender, Automatic Web News Extraction Using Tree Edit Distance. In *WWW* 2004:, pp. 502–511.

[6] Y. Zhai, and B. Liu. Web Data Extraction Based on Partial Tree Alignment. In *WWW* 2005: p. 76–85.

[7] K. Zhang, and D. Shasha. Tree Pattern Matching. In *Pattern Matching Algorithms*; Oxford University Press, 1997.

[8] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu. Fully Automatic Wrapper Generation for Search Engines. In *WWW* 2005: p. 66–75.