# Extracting Structured Data from Web Pages (Poster)

Arvind Arasu
Stanford University, CA
arvinda@cs.stanford.edu

Hector Garcia-Molina
Stanford University, CA
hector@cs.stanford.edu

Many web sites contain a large collection of "structured" web pages. These pages encode data from an underlying structured source, and are typically generated dynamically. An example of such a collection is the set of book pages in Amazon. There are two important characteristics of such a collection: first, all the pages in the collection contain structured data conforming to a common *schema*; second, the pages are generated using a common *template*.

Our goal is to *automatically* extract structured data from a collection of pages described above, without any human input like manually generated rules or training sets. Extracting structured data gives us greater querying power over the data and is useful in information integration systems. Most of the existing work on extracting structured data assumes significant human input, for example, in form of training examples of the data to be extracted. To the best of our knowledge, ROADRUNNER project is the only other work that tries to automatically extract structured data. However, ROADRUNNER makes several simplifying assumptions. These assumptions and their implications are discussed in our paper [2].

Structured data denotes data conforming to a schema or type. We borrow the definition of complex types from [1]. Any value conforming to a schema is an instance of the schema. For example, the schema $S = \langle \mathcal{B}, \{\mathcal{B}\}, \mathcal{B} \rangle$ represents a tuple of 3 attributes. The first and third attributes are "atomic"; the second attribute is a set of atomic values. The value $x = \langle t, \{f_1, f_2\}, c \rangle$ denotes an instance of schema $S$. A template is a *pattern* that describes how instances of a schema are encoded. An example template for schema $S$ above is $T = \langle A * B\{\}_E C * D \rangle$ where each letter denotes a string. Template $T$ encodes the first attribute of $S$ between strings $A$ and $B$, the second between $B$ and $C$, and so on. Further, the set of elements of the second attribute are separated from each other by string $E$. Encoding value $x$ above using $T$ results in page $AtBf_1Ef_2CcD$.

The extraction problem is formalized as follows: *given a set of $n$ pages created from unknown template and values, extract the values encoded in the pages*. This statement of the problem is not complete since there could be more than one solution for a given set of pages. However, for a set of real pages a human rarely has any ambiguity in picking the "right" template and values. Our goal is to solve the extraction problem for real pages, *i.e.*, produce the values that a human would consider semantically correct.

Our approach consists of two stages. In the first stage, the unknown template used to create the pages is deduced. In the second stage, the deduced template is used to extract the values. We focus on the first stage since it is more challenging. Each word in an input page either occurs as part of the template or as part of the encoded data used to create the page. The task of deducing the template is hard because, syntactically, there is no difference between a word that is part of the template and a word that is part of encoded data. Our main contribution is the observation that, for real pages, the words that are part of the template have a high correlation of occurrence in input pages with other words in the template. The words that are part of encoded data, on the other hand, do not have high occurrence correlation with other words. This observation can be used to design an algorithm for deducing the template from input pages. The full version of the paper contains formal definition of high occurrence correlation and our algorithm.

We evaluated our approach by considering 9 real collections of pages. For each collection, we measured the number of atomic attributes that were correctly extracted by our algorithm. For the above set of collections our algorithm was able to extract more than 90% of the attributes correctly, thereby validating our approach.

## References

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison Wesley, Reading, Massachussetts, 1995.

[2] A. Arasu and H. Garcia-Molina. Extracting structured data from web pages. Technical report, Stanford University, Database Group, 2002. Available at `http://dbpubs.stanford.edu/pub/2002-40`.

Proceedings of the 19th International Conference on Data Engineering (ICDE'03)
1063-6382/03 $ 17.00 © 2003 IEEE