Automatic Hidden-Web Table Interpretation by Sibling Page Comparison

Cui Tao* and David W. Embley* 
Brigham Young University, Provo, Utah 84602, U.S.A.

Abstract. The longstanding problem of automatic table interpretation still eludes us. Its solution would not only be an aid to table processing applications such as large volume table conversion, but would also be an aid in solving related problems such as information extraction and semi-structured data management. In this paper, we offer a conceptual modeling solution for the common special case in which so-called sibling pages are available. The sibling pages we consider are pages on the hidden web, commonly generated from underlying databases. We compare them to identify and connect nonvarying components (category labels) and varying components (data values). We tested our solution using more than 2,000 tables in source pages from three different domains—car advertisements, molecular biology, and geopolitical information. Experimental results show that the system can successfully identify sibling tables, generate structure patterns, interpret tables using the generated patterns, and automatically adjust the structure patterns, if necessary, as it processes a sequence of hidden-web pages. For these activities, the system was able to achieve an overall F-measure of 94.5%.

1 Introduction

The World Wide Web serves as a powerful resource for every community. Much of this online information, indeed, the vast majority, is stored in databases on the so-called hidden web. Hidden-web information is usually only accessible to users through search forms and is typically presented to them in tables. Automatically understanding hidden-web pages is a challenging task. In this paper, we introduce a domain independent, web-site independent, unsupervised way to automatically interpret tables from hidden-web pages.

Tables present information in a simplified and compact way in rows and columns. Data in one row/column usually belongs to the same category or provides values for the same concept. The labels of a row/column describe this category or concept.

Although a table with a simple row and column structure is common, tables can be much more complex. Figure 1 shows an example. Tables may be nested

* Supported in part by the National Science Foundation under Grant #0414644.
1 There are more than 500 billion hidden-web pages. The surface web, which is indexed by common search engines only constitutes less than 1% of the World Wide Web. The hidden web is several orders of magnitude larger than the surface web [10].