Wrapper Induction based on Minimum Description Length using a Suffix Tree

Dae-Ki Kang and Kiwook Sohn
National Security Research Institute
161 Gajeong-Dong, Yuseong-Gu, Daejeon, South Korea
E-mail: dkkang@etri.re.kr kiwook@etri.re.kr

Abstract: Automated generation of wrappers has been one of important topics in Web mining because wrappers bridge between HTML hypertext pages on the Web and business applications that need useful information from the HTML pages in a structured form. In this paper, we propose an efficient and accurate wrapper induction technique to elicits concise patterns that occur frequently in the HTML pages using minimum description length (MDL) principle and a suffix-tree sequence storage mechanism. To induce accurate and concise wrapper patterns from Web pages, our algorithm, MDL-Wrapper, uses MDL principle as a trade-off criterion between the number of occurrence of important patterns and the length of the patterns. The estimation of the occurrence is efficiently calculated by and obtained from suffix tree storage mechanism. Experiments on wrappers for price information and news information from popular Web pages as unlabeled examples show that MDL-Wrapper is efficient and effective for wrapper induction tasks.

1. Introduction

With the rapid growth of World Wide Web, automated generation of wrappers becomes one of important topics in Web data mining, because wrappers transform HTML hypertext pages on the Web into a structured form for business applications that need information from the HTML pages. Since the useful information for the application is generally stored and utilized as a set of structured records and the records are available in the HTML pages without the explicit protocol of the structure, "detecting frequently occurred patterns" inside the HTML pages is a crucial problem for automated wrapper generation. However, most wrapper induction methods simply rely on ad-hoc methods or complicated dictionary search to extract the necessary information.

Considering this background, we introduce a novel wrapper induction technique that elicits concise patterns that occur frequently in the HTML pages efficiently. Our algorithm, MDL-Wrapper, considers minimum description length (MDL) as a trade-off criterion between the number of occurrence of patterns and their lengths. Given the HTML document, MDL-Wrapper removes needless HTML tags and tokenizes the rest in the document, which is analogous to noise removal process in image processing. After the removal of needless tags, MDL-Wrapper inserts the tokenized sequence into a generalized suffix tree.

Experiments on popular Web pages as unlabeled examples show that our MDL-Wrapper is efficient and effective for wrapper induction tasks.

The rest of the paper is organized as follows: section 2 briefly introduces suffix tree and MDL; section 3 presents MDL-Wrapper; and section 4 describes preliminary experimental results on online stores and news sites.

2. Preliminaries

Before we describe our algorithm, it is helpful to cover fundamental concepts and notions adapted in our research.

2.1 Suffix Tree

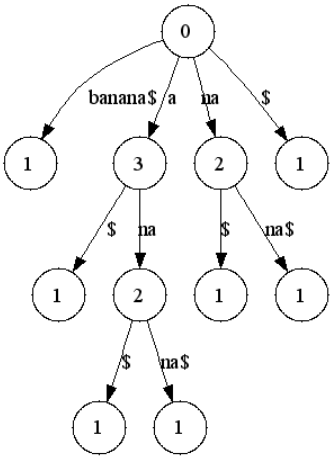

Fig 1 A suffix tree of a string "banana$". A number in each node represents a number of occurrence of patterns.

Suffix tree is a data structure for indexing a string so that we can easily find a pattern in the string. Figure 1 shows an example suffix tree for a string "banana$", where '$' denotes the end of the string. A number in each node represents a number of occurrence of patterns. For example, in the string "banana$", 'a' occurs three times and 'na' occurs twice. Ukkonen[1] devised a linear time algorithm for constructing the suffix tree. When the length of a string is *n*, then it takes a linear *O(n)* time to build a suffix tree for the string. Once a suffix tree is generated, then it takes *O(m)* time to find a pattern with length *m*. Also, with edge-label compression, it only needs *O(n)* space for a suffix tree. In practice, to store multiple strings, a generalized suffix tree is used.

2.2 Minimum Description Length

With minimum description length (MDL) principle [2], MDLWrapper controls the extent of generalization during the wrapper induction. MDL is a criterion that trades off the accuracy and the size of a theory. That is, MDL principle chooses the theory (or model) from a set of data that minimizes (1) the sum of the length of the theory and (2) the length of the data that are encoded according to the theory. In our setting, the theory is an inferred pattern of

the wrapper, and the data is the tokenized sequence converted from an HTML document and stored in a generalized suffix tree.

3. Minimum Description Length guided Wrapper (MDL-Wrapper)

3.1 Problem Definition

We briefly define the problem of wrapper induction from HTML documents as follows: Let *D(i)* be a document *i* that consistes of words from finite alphabet Σ, and let *TOK :* $\Sigma^* \rightarrow \Sigma_{TOK}^*$ be a tokenizing function that removes needless HTML tags and tokenized HTML tags and words into tokens $\in \Sigma_{TOK}$, then the goal is to find a token sub-sequence that maximizes the MDL fomula for evaluating conciseness and accuracy.

3.2 String Generation from a Web Page

Since HTML web pages are generated for visual presentation purpose, the pages mostly have HTML tags used for decoration of their contents. Although, the tags are for visual decoration, a few of them sometime imply the systematic structure of the contents. For example, <*table*>, <*th*>, and <*td*> tags are for tables, but those tags strongly imply that they are used to present relational data, i.e. multiple records. With these considerations, we remove most of these HTML tags which are not suitable for extracting wrapper patterns [3]. This removal (or abstraction) stage also makes the wrapper robust to small changes of HTML documents, because most HTML tags for visual presentation are removed and disregarded in wrapper induction.

After removing the HTML tags for visual presentation, we generate one token sequence from one HTML document. In the sequence, each token denotes a HTML tag or text data. After the generation of a token sequence, the problem of wrapper induction is reduced to the problem of finding a token sub-sequence that, when converted back to the original tags and text, covers relational data in the HTML page. Figure 2 shows some rules to the tokenizing process.

<TABLE>	→	T
</TABLE>	→	t
<TR>	→	R
</TR>	→	r
<TD>	→	D
</TD>	→	d
<A>	→	A
</A>	→	a
hyperlink (HREF=…)	→	U
other (general text)	→	X

Fig 2 Tokenizing rules.

Note that, from the rules in figure 2, we treat all general texts equally (as 'X'), but maintain the HTML tags for table structure because they provide significant clues for relational data.

3.3 Algorithm

The major steps for MDLWrapper is as folllows:

MDLWrapper(D)

1. For each document $D_i \in \boldsymbol{D}$, remove needless HTML tags and normalize relative URL's.
2. Generate token sequence from the HTML documents and insert the token sequences into a generalized suffix tree.
3. Sort all the nodes in the suffix tree in descending order, and for each suffix node, obtain the pattern *p* corresponding to the node.
4. Choose the node and its pattern that maximizes the MDL score.

Considering that we are interested in the patterns that are occurred frequently and we want the pattern to be meaningful (i.e. long enough), minimum description length for wrapper induction can be formulated as follows:

$$MDL(tok) = \#(tok) \bullet \sum_i w(tok_i) + \alpha \bullet l(tok)$$

where $w(tok_i)$ is a user-specified weight for a token which reflects the user's domain knowledge, *#(tok)* is the number of occurrence of a token sub-sequence, l(tok) is the length of a token sequence, and α is a user-supplied parameter.

4. Experiments

We have performed experiments of our MDLWrapper on price information and news information from popular Web pages. The preliminary experimental results show that MDL-Wrapper is efficient and effective for wrapper induction tasks.

For example, for news sites, we have found that very simple patterns like 'AUXa', 'XAUa', and 'AUaX' occurs frequently and are effective as wrapper patterns for news articles, although we have found more diverse patterns for online stores's price information.

References

[1] E. Ukkonen, "On-line construction of suffix-trees," *Algorithmica*, 1995, 14, 249-260.

[2] J. Rissanen, "Modeling by shortest data description," *Automatica*, 1978, 14, 465-471.

[3] D.-K. Kang, and J. Choi, "MetaNews: An Information Agent for Gathering News Articles on the Web," *International Symposium on Methodologies for Intelligent Systems*, Maebashi City, Japan, 2003.