Incremental Discovery of Sequential Pattern from Semi-structured Document Using Grammatical Inference

Ramesh Thakur[1], Suresh Jain[2], and Narendra S. Chaudhari[3]

[1] IIPS, DAVV, Indore, India
[2] KCB Technical Academy, Indore, India
[3] Indian Institute Of Technology, Indore

Abstract. On the World Wide Web a large numbers of information is available in the form of semi-structured format. Knowledge discovery in semi-structured document has been recognized as promising task. Since semi structured document is typically hidden within HTML formatting intended for human viewing the details of which vary widely from site to site and frequent changes made to their formatting so we can't construct a global schema, discovery of interesting rules form it is complex and tedious process. Most of the existing system uses hand-coded wrappers to extract information, which is monotonous and time consuming. An intelligent and automated method is needed for their processing. Learning grammatical information from given sample of semi-structured documents has attracted lots of attention in the past decades. To understand "what say the data" is necessary to know the structure of data to discover the syntactic-semantic knowledge of its language.

The problem of learning the correct grammar for the unknown language form finite example of the language is known as grammatical inference problem. In automated grammar learning, the task is to infer grammar rules from given information about the target language. If example belongs to the target language it is called positive example otherwise it is called negative example. In this paper we propose a grammar inference methodology to automate the construction of grammar rules and facilitate the process of information extraction. We are using hybrid technique of association analysis and sequential algorithm to generate context free grammar rules from semi-structured document (HTML document).

Our algorithm that infers a sequential pattern from a sequence[1] of discrete HTML tags. The basic insight is that sub-string is selected on the basis of high support factor[2] by taking entire sentences into account. Which appears more frequently in string can be replaced by a grammatical rule that generate the sub-string, and this process is repeated many times, producing a single length rules of the sequence. The result is strictly a context-free grammar rules, which provide a compact summary of corpora that aids understanding of its properties.

Keywords: Knowledge discovery, sequence mining and grammar inference.

---

[1] Sequence: A sequence is an ordered list of alphabet symbols. We denote a sequence by <s1s2….sn) where si is a symbol. A sequence <a1, a2, …. An) is a subsequence of another sequence < b1, b2, ……bm> if there exist integers i1<i2<….in such that a1=bi1, a2=bi2, ……an=bin. For example, the sequence bob is subsequence of bbobbb.

[2] Support factor: The support Factor (SFβ) for sub-sequences in corpora C.

$$SF\beta = \sum_{i=1}^{N} \text{count of } \beta \text{ in sentence} \times \text{length of } \beta \; / \text{ length of sentence}$$

Where N= number of sentences in Corpora C and β is a candidate sub-sequence for replacement.

R. Ramanujam and S. Ramaswamy (Eds.): ICDCIT 2012, LNCS 7154, p. 269, 2012.