A Rule-based Transformation Engine for Web Page Re-authoring

Péter Cserkúti and Hassan Charaf

This paper introduces SmartWeb, a system for web page re-authoring. It reveals its main algorithms and highly extensible architecture. The purpose of web page re-authoring (a.k.a. web content adaptation) is to transform the original web content in a way that it can adequately be displayed on a client device, let it be a mobile, a PDA or a desktop computer. The adaptation process should take the capabilities of the rendering client into consideration such as screen resolution, network bandwidth, type and speed of processor, amount of memory or software configuration and produce the best version of the original content for that. This results a renderer independent web browsing capability.

SmartWeb is a proxy-based solution. It is always situated between the client and the web server that contains the requested web page. It catches all the queries of the client and before returning the requested content it fulfills an adaptation process on the original page.

The essence of SmartWeb is the adaptation process itself. Its input is the original html page and its output is a modified version of it. SmartWeb handles html pages as trees that we call transformation tree. The adaptation process is realized by the help of graph transformation. One can define special graph transformation rules that the engine can execute against the transformation tree. A rule is made up by two parts: a left hand side (LHS) pattern and a right hand side (RHS) tree with witch the LHS pattern must be replaced. The paper introduces the method how SmartWeb facilitates defining LHS patterns representing common block types in web pages, and exhibits its technique for describing RHS trees for them. A major point of the graph transformation process is matching the specified patterns in the transformation tree. There are existing algorithms for pattern matching such as Ullmann's algorithm or the VF2 algorithm. However because of the complexity of our pattern definition method they can not be used here or must be extended. This paper will also introduce an algorithm for matching these patterns in the transformation tree and an other algorithm for transforming them according to the RHS tree. There are two things that make these algorithms powerful and SmartWeb easy and convenient to configure: first when defining an LHS pattern one can freely refer to an other previously matched pattern as a node of the graph and second when defining the RHS tree referring to subtrees of the LHS pattern is allowed.

SmartWeb is an extensible framework for web content adaptation. Its extensibility has got two levels. First, there are pipelines defined in the system and all the algorithms implemented are attached to an adequate one. New algorithms that analyze or transform the transformation tree can easily be implemented and attached to the system. Second, the set of transformation rules can be extended. New patterns and rules can be defined in XML. The paper also introduces the architecture of SmartWeb.