- Text cue: If most of the children of a DOM node are text nodes or virtual text node, we prefer not to divide it.
- Size cue: We predefine a relative size threshold (the node size compared with the size of the whole page or sub-page) for different tags (the threshold varies with the DOM nodes having different HTML tags). If the relative size of the node is small than the threshold, we prefer not to divide the node.

Based on these cues, we can produce heuristic rules to judge if a node should be divided. If a node should not be divided, a block is extracted and we will set the DoC value for this block. We list the heuristic rules in Table 1 by their priority.

Table 1. Heuristic rules in block extraction phase

Rule 1	If the DOM node is not a text node and it has no valid children, then this node cannot be divided and will be cut.
Rule 2	If the DOM node has only one valid child and the child is not a text node, then divide this node.
Rule 3	If the DOM node is the root node of the sub-DOM tree (corresponding to the block), and there is only one sub DOM tree corresponding to this block, divide this node.
Rule 4	If all of the child nodes of the DOM node are text nodes or virtual text nodes, do not divide the node. ● If the font size and font weight of all these child nodes are same, set the DoC of the extracted block to 10. ● Otherwise, set the DoC of this extracted block to 9.
Rule 5	If one of the child nodes of the DOM node is line-break node, then divide this DOM node.
Rule 6	If one of the child nodes of the DOM node has HTML tag <HR>, then divide this DOM node
Rule 7	If the sum of all the child nodes' size is greater than this DOM node's size, then divide this node.
Rule 8	If the background color of this node is different from one of its children's, divide this node and at the same time, the child node with different background color will not be divided in this round. ● Set the DoC value (6-8) for the child node based on the html tag of the child node and the size of the child node.
Rule 9	If the node has at least one text node child or at least one virtual text node child, and the node's relative size is smaller than a threshold, then the node cannot be divided

	● Set the DoC value (from 5-8) based on the html tag of the node
Rule 10	If the child of the node with maximum size are small than a threshold (relative size), do not divide this node. ● Set the DoC based on the html tag and size of this node.
Rule 11	If previous sibling node has not been divided, do not divide this node
Rule 12	Divide this node.
Rule 13	Do not divide this node ● Set the DoC value based on the html tag and size of this node.

For different DOM nodes with different HTML tags, we will apply different rules. We listed them in Table 2.

Table 2. Different rules for different DOM nodes

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
Inline Text Node	√	√	√	√	√	√	√		√	√		√	
<TABLE>	√	√	√					√		√			√
<TR>	√	√	√				√	√		√			√
<TD>	√	√	√	√					√	√	√		√
<P>	√	√	√	√	√	√	√		√	√		√	
Other Tags	√	√	√	√		√	√		√	√		√	

Let us consider an example shown in Figure 1. At the first round of block extraction, VB1, VB2_1, VB2_2, VB2_3, VB3 and VB4 will be extracted and put into the pool. Below we explain how the VB2_1, VB2_2 and VB2_3 are extracted in details. Figure 7(b) is a table, which is a part of the whole web page. Its DOM tree structure is shown on the left side. In the block extraction process, when the <TABLE> node is met, it has only one valid child <TR>. We trace into the <TR> node according to the rule 2. The <TR> node has five <TD> children and only three of them are valid. The first child's background color is different from its parent's background color. According to the rule 8, the <TR> node is split and the first <TD> node is not divided further in this round and is put into the pool as a block. The second and fourth child of <TR> node is not valid and will be cut. According to the rule 11, the third and fifth children of <TR> will