Comparison

Both algorithms mentioned in our two papers basically try to do the same thing. They both try to find the data-rich section or table on a webpage. This is important for the wide research field of webdata extraction to get suitable information for newsclients or mobile devices as pdas. Whereas the machine learning approach of table detection is a more special way to solve the problem of webdata extraction, the DSE-algorithm is a more universal way to find a data-rich section. This is an advantage for the DSE-algorithm because this algorithm can be easily adapted in order to solve similar problems. Therefore the machine learning approach is suited especially for this problem and it would be very difficult if not impossible to adapt to other problems.

The ground truth database from which the machine learning method learns to get the right tables can be seen as an advantage, because it makes this approach more resistant. For example when there is a webpage that has not been built by a CMS and there are no similar pages the DSE-algorithm fails. If the papers can be trusted the machine learning method achieves a accuracy of max. 95% while the DSE hits up to 100% of success. Unfortunately there is no way to control these performance statistics so that we can't come to a scientific based conclusion on this issue. The complexity of these algorithms also make it very difficult for us to evaluate realistic performance data.

The machine learning method is a concept with many separate parts that can be easily extended. So it is probably no problem to find new significant features and to integrate them into the algorithm. A good example for this extendability is the usage of two different learning algorithms.

The papers gave us an insight view of how the algorithms work and how successful they are with given inputs, but it's kind of hard to decide which algorithm is the better one. They are both superior for different problems. While the DSE-algorithm seems to work better for pages which are part of a big web-site, the machine-learn-approche is superior for finding data-rich sections on stand-alone web-sites.