Merging Prediction by Partial Matching with Structural Contexts Model*

Joaquín Adiego[1], Pablo de la Fuente[1] and Gonzalo Navarro[2]

[1]Dpto. de Informática, Universidad de Valladolid, Valladolid, España.

{jadiego, pfuente}@infor.uva.es

[2]Dpto. de Ciencias de Computación, Universidad de Chile, Santiago, Chile.

gnavarro@dcc.uchile.cl

Our goal in this work is to explore the possibility of considering the text structure in the context of compressed structured documents. Structure has semantic meaning but classical compressors do not profit from structure. We aim at taking advantage of the structure. An initial approach is XMLPPM, where the context given by the path in the structure tree is used to model the text in the subtree. That is, different models are used to code tags names, attribute names, attribute values, textual content, and so on. XMLPPM is based on the intuition that the text under similar structural elements (i.e., XML tags) should follow a similar distribution. Our idea is then to use separate models to compress the text that lies inside different tags. For example, in an email archive, a different model would be used for each of the fields `From:`, `Subject:`, `Date:`, `Body:`, etc.

We propose a compression technique for structured documents, called *SCMPPM*, which combines the Prediction by Partial Matching technique with Structural Contexts Model idea which takes advantage of the context information usually implicit in the structure of the text. The idea is to use a separate PPM model to compress the text that lies inside each different structure type (e.g., different XML tag). The intuition behind the idea is that the distribution of all the texts that belong to a given structure type should be similar, and different from that of other structure types, allowing that PPM model can make better predictions. We test our idea using the classic PPM character-based variant (PPMD+) and we assumed that each tag has its own PPM model and that a default PPM model is used for the text that is not under any tag. Our experimental results show significant gains over the methods that are insensitive to the structure and over the current methods that consider the structure. We have shown that the method actually improves compression ratios by more than 35% with respect to the SCM basic technique (that uses word-based Huffman coders). We have compared our prototype against state-of-the-art compression systems, showing that our prototype obtains the best compression for all collections improving over XMill by 77%, over MG System by 38% and over XMLPPM by 25%.

*This work was partially supported by CYTED VII.19 RIBIDI project (all authors), Millenium Nucleus Center for Web Research, Grant P01-029-F, Mideplan, Chile. (third author) and the TIC2003-09268 project from MCyT, España (first and second authors)