Page-Level Main Content Extraction from Heterogeneous Webpages

Julián Alarte[a], Josep Silva[a]

[a] *Valencian Research Institute for Artificial Intelligence (VRAIN)*
Universitat Politècnica de València

Abstract

The main content of a webpage is often surrounded by other boilerplate elements related to the template, such as menus, advertisements, copyright notices, and comments. For crawlers and indexers, isolating the main content from the template and other noisy information is an essential task, because processing and storing noisy information produce a waste of resources such as bandwidth, storage space, and computing time. Besides, the detection and extraction of the main content is useful in different areas, such as data mining, web summarization, and content adaptation to low resolutions. This work introduces a new technique for main content extraction. In contrast to most techniques, this technique not only extracts text, but also other types of content, such as images, and animations. It is a Document Object Model-based page-level technique, thus it only needs to load one single webpage to extract the main content. As a consequence, it is efficient enough as to be used online (in real-time). We have empirically evaluated the technique using a suite of real heterogeneous benchmarks producing very good results compared with other well-known content extraction techniques.

Artículo aceptado para su publicación en la revista ACM Transactions on Knowledge Discovery from Data.

Email addresses: jualal@doctor.upv.es (Julián Alarte), jsilva@dsic.upv.es (Josep Silva)