Title: Web Information Extraction: An Algorithm for Detecting Main Content in News Pages

Student: Hamza Salem
Supervisor: Radek Burget

Extended abstract:
Web information extraction is a fundamental component of data retrieval and content analysis within the expansive World Wide Web. This abstract presents an algorithm aimed at the identification and extraction of primary content from news pages, seamlessly combining two pivotal concepts discussed in "Handbook of Mathematical Models for Languages and Computation" by Prof. Meduna, specifically, the application of regular expressions(Chapter 5) and tree graphs(Chapter 3).

To initiate the process, the algorithm employs robust regular expression techniques for pattern matching within HTML documents. It prioritizes the search for content enclosed within <h1> tags, a common practice as headlines in web articles often reside within these tags.

For the identification of the article's main body, the algorithm capitalizes on tree structures. By constructing an HTML tree that mirrors the hierarchical structure of the web page, it incorporates the essence of tree graphs, deviating from mere reliance on HTML tag semantics and adopting an innovative approach.

The core concept involves singling out the node within the HTML tree that boasts the highest number of direct child nodes. This chosen node, in unison with its children, encapsulates the most extensive textual content. By prioritizing nodes with significant textual information and a dense branching structure, the algorithm effectively and efficiently extracts the principal body of the article.

This research contributes to the realm of web information extraction by fusing the principles of regular expressions and tree graphs. The algorithm's proficiency in pinpointing the most pertinent content underpins more efficient data retrieval and analysis, serving as a linchpin for a wide spectrum of applications, including content summarization, sentiment analysis, and information retrieval in the contemporary digital landscape.