An Algorithmic Framework for Precise Main Content Extraction from News Websites

Алгоритмический фреймворк для точного извлечения основного содержимого с новостных веб-сайтов

Хамза Салем

Университет Иннополис

This thesis presents the design, implementation, and evaluation of a novel, open-source algorithm for Main Content Extraction (MCE) from web pages. The proposed algorithm operates on the Document Object Model (DOM) tree of an HTML document and employs a multi-criteria heuristic approach to identify the primary content node. It combines three key metrics: the node with the highest number of direct text-containing children, the node with the most text content that lacks text-bearing children, and the node closest to the middle depth of the DOM tree. This methodology is intentionally language-agnostic, relying on structural features rather than linguistic cues, making it particularly effective for multilingual content and languages with complex tokenization.

The algorithm's performance was rigorously evaluated against two established content extraction tools, Readability and Boilerpipe, using metrics including precision, recall, F1-score, and accuracy. Results demonstrate that the proposed MCE algorithm significantly outperforms these existing solutions, achieving near-perfect scores (e.g., Precision: 99.96%, Recall: 99.69%, F1-Score: 99.80%, Accuracy: 99.65% on the primary dataset) and consistently maintaining superior performance on a secondary benchmark (Webz.io dataset). The work contributes not only a highly accurate and efficient extraction tool but also a standardized benchmark dataset to foster future research. The practical implications are substantial, offering a cost-effective method to enhance Large Language Models (LLMs) and improve global information accessibility by accurately extracting content across diverse languages and web structures.

В данной диссертации представлены разработка, реализация и оценка нового алгоритма с открытым исходным кодом для извлечения основного контента (MCE) с веб-страниц. Предлагаемый алгоритм работает на основе дерева объектной модели документа (DOM) HTML-документа и использует многокритериальный эвристический подход для определения основного узла контента. Он объединяет три ключевые метрики: узел с наибольшим количеством дочерних элементов, содержащих текст, узел с наибольшим количеством текстового контента, но без дочерних элементов, содержащих текст, и узел, ближайший к середине дерева DOM. Эта методология

намеренно не зависит от языка и опирается на структурные особенности, а не на лингвистические подсказки, что делает ее особенно эффективной для многоязычного контента и языков со сложной токенизацией.

Производительность алгоритма была тщательно оценена с использованием двух известных инструментов извлечения контента, Readability и Boilerpipe, с использованием таких метрик, как точность, полнота, F1-оценка и достоверность. Результаты показывают, Предлагаемый алгоритм MCE значительно превосходит существующие решения, достигая практически идеальных результатов (например, точность: 99,96%, полнота: 99,69%, оценка F1: 99,80%, достоверность: 99,65% на первичном наборе данных) и стабильно сохраняя превосходную производительность на вторичном эталонном наборе данных (набор данных Webz.io). Работа представляет собой не только высокоточный и эффективный инструмент извлечения, но и стандартизированный эталонный набор данных для будущих исследований. Практические выводы весьма существенны: это экономически эффективный метод улучшения больших языковых моделей (LLM) и повышения глобальной доступности информации за счет точного извлечения контента на различных языках и в различных веб-структурах.