IMPROVING RECALL IN IMAGE-BASED DEEP WEB CONTENT EXTRACTION: A STATISTICAL STUDY OF WEID AND WEIDJ

Ily Amalina Ahmad Sabri[1,*], Mustafa Man[1], Xu Jianhui[2], Astari Retnowardhani[3]

[1]*Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, Kuala Nerus, 21030, MALAYSIA*
[2] *School of Medical Information Engineering, Shenyang Medical College, Liaoning, Shenyang, 110034, CHINA.*
[3]*Information Systems Management Department, BINUS Graduate Program - Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia 11480*

ABSTRACT

Main content extraction from Deep Web pages is challenging due to dynamic layouts and the increasing use of image-based content. This paper presents two extraction techniques, WEID and WEIDJ, designed to identify main content regions from web page images. The methods model page layouts as structured representations and apply quantitative criteria to separate informative regions from non-content elements. A statistical evaluation framework is used to assess extraction performance based on precision, recall, and F1-score. Experiments were conducted on a dataset comprising 54,321 images collected from 10 websites, covering diverse and complex layout structures. Results demonstrate that WEIDJ achieves a precision of 99% and improves recall by 85% over WEID, which attains a precision of 60% and a recall of 78%, reflecting a more balanced but less complete extraction profile. The overall average F_1-score across all experiments reaches 86%, confirming the strong effectiveness of the proposed framework. Beyond extraction performance, the proposed techniques support downstream AI and machine learning pipelines by enabling reliable content isolation for training data preparation, annotation workflows, and feature generation. The findings underscore the critical role of rigorous statistical evaluation in advancing image-based data extraction within AI-driven web mining research.

***K*eywords** Deep Web · content extraction · WEID · WEIDJ · precision and recall ·

References

[1] Cai, D., Yu, S., Wen, J. R., & Ma, W. Y. , VIPS: A vision-based page segmentation algorithm. Microsoft Technical Report, MSR-TR-2003-79, 2003.
[2] Chang, C. H., Kayed, M., Girgis, M. R., & Shaalan, K. F., A survey of web information extraction systems. IEEE Transactions on Knowledge and Data Engineering, 18(10), 1411–1428, 2006.
[3] Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. Knowledge-Based Systems, 70, 301–323, 2014.
[4] Laender, A. H., Ribeiro-Neto, B. A., da Silva, A. S., & Teixeira, J. S., A brief survey of web data extraction tools. ACM SIGMOD Record, 31(2), 84–93, 2002.
[5] Reis, D. C., Golgher, P. B., Silva, A. S., & Laender, A. F., Automatic web news extraction using tree structure algorithms. Proceedings of the 13th International Conference on World Wide Web (WWW), 502–511, 2004.
[6] Sun, F., Song, D., & Liao, L. DOM based content extraction via text density. Proceedings of the 34th International ACM SIGIR Conference, 245–254, 2011.

[*]*Corresponding Author's E-mail: ilylina@umt.edu.my*