XPath Agent: An Efficient XPath Programming Agent Based on LLM for Web Crawler

Li Yu
lijingyu68@gmail.com

Wang Jixing
Science Dept.
Online City, CEUR 99099
jqq@ceur-ws.org

Hazul
Research Dept.
Science City, Sci 88088
yvo@science.rdept.net

Abstract

We introduce XPath Agent, a production-ready XPath programming agent tailored for web crawling tasks. A standout feature of XPath Agent is its capability to automatically program XPath queries from a set of sampled web pages. To illustrate its efficacy, we benchmark XPath Agent against a state-of-the-art XPath programming agent across a suite of web crawling tasks. Our findings reveal that XPath Agent excels in F1 score with minimal compromise on accuracy, while significantly reducing token usage and increase clock-time efficiency. The well designed 2 stage pipelines makes it readily integrable into existing web crawling workflows, thereby saving time and effort in manual XPath query development.

1 Introduction

Web scraping [1] automates data extraction from websites, vital for modern fields like Business Intelligence. It excels in gathering structured data from unstructured sources like HTML, especially when machine-readable formats are unavailable. Web scraping provides real-time data, such as pricing from retail sites, and can offer insights into illicit activities like darknet drug markets.

2 Related Work

...

In: W. Aigner, G. Schmiedl, K. Blumenstein, M. Zeppelzauer (eds.): Proceedings of the 9th Forum Media Technology 2016, St. Pölten, Austria, 24-11-2016, published at http://ceur-ws.org

3 Methodology

...

3.1 Information Extraction

...

3.2 XPath Program

...

4 Experiments

...

5 Conclusion

Third level headings must be flush left, initial caps and bold. One line space before the third level heading and 1/2 line space after the third level heading.

Fourth Level Heading

Fourth level headings must be flush left, initial caps and roman type. One line space before the fourth level heading and 1/2 line space after the fourth level heading.

5.1 Citations In Text

Citations within the text should indicate the author's last name and year[**?**]. Reference style[**?**] should follow the style that you are used to using, as long as the citation style is consistent.

5.1.1 Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page they appear on. Precede the footnote with a vertical rule of 2 inches (12 picas).

[1] This is a sample footnote

5.1.2 Figures

All artwork must be centered, neat, clean and legible. Do not use pencil or hand-drawn artwork. Figure number and caption always appear after the the figure. Place one line space before the figure, one line space before the figure caption and one line space after the figure caption. The figure caption is initial caps and each figure is numbered consecutively.

Make sure that the figure caption does not get separated from the figure. Leave extra white space at the bottom of the page to avoid splitting the figure and figure caption.

Figure 1 shows how to include a figure as encapsulated postscript. The source of the figure is in file `fig1.eps`.

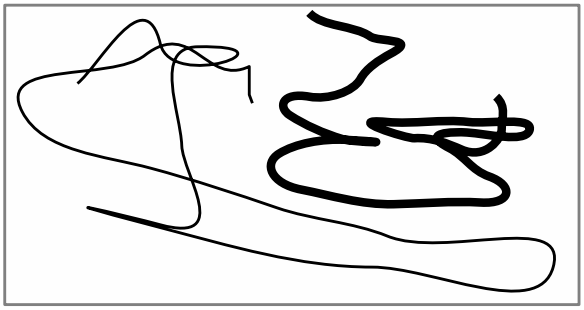

Figure 1: Sample EPS figure

Below is another figure using LaTeX commands.

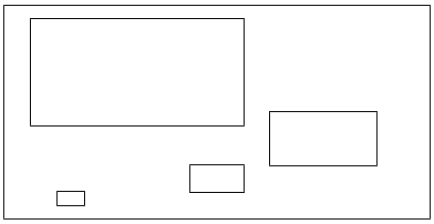

Figure 2: Sample Figure Caption

5.1.3 Tables

All tables must be centered, neat, clean and legible. Do not use pencil or hand-drawn tables. Table number and title always appear before the table.

One line space before the table title, one line space after the table title and one line space after the table. The table title must be initial caps and each table numbered consecutively.

5.1.4 Handling References

Use a first level heading for the references. References follow the acknowledgements.

Table 1: Sample Table

A	B	1
C	D	2
E	F	3

5.1.5 Acknowledgements

Use a third level heading for the acknowledgements. All acknowledgements go at the end of the paper.

References

[1] Moaiad Ahmad Khder. Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, 13(3), 2021.