대형언어모델 기반 다중 에이전트 시스템을 사용한 자동 웹 크롤러 생성

곽지호*, 김민근*, 정유철**

Automated Web Crawler Generation Using LLM-Based Multi-Agent Systems

Jiho Gwak*, Mingeun Kim*, and Yuchul Jung**

요 약

데이터를 효율적으로 수집하기 위한 대표적인 도구로 크롤러가 있으며, 기존 크롤러의 구현을 위해서는 HTML 구조를 사람이 분석하는 과정이 필요하므로 이에 대한 추가 비용이 발생한다는 한계가 있다. 본 논문에서는 이러한 문제를 해결하기 위해 대형언어모델을 활용한 멀티 에이전트를 기반으로 하여 자동 크롤러 작성 기법을 제안하여, 사람의 개입 없이도 효율적인 데이터 수집의 가능성을 확인하였다.

Abstract

Crawlers are typical tools for efficiently collecting data. Existing crawlers have the limitation that humans must analyze the HTML structure for implementation, which incurs additional costs. This paper proposes an automated crawler generation technology based on a multi-agent system utilizing LLM to solve this problem and confirms the possibility of efficiently collecting data without human intervention.

Key words

multi agent system, automatic crawler, data collection, large language model

Ⅰ. 서 론

컴퓨터 시스템이 발전함에 따라 데이터는 시장 분석, 동향 분석 등에서 핵심 자원으로서 중요한 역할을 한다. 이러한 데이터를 효율적으로 수집하기 위한 도구로 웹 크롤러가 있으며, 이는 웹 페이지에서 데이터를 자동으로 수집하여 유용한 정보를 제공한다. 그러나 웹의 데이터는 대부분 비정형 데이터로 존재[1]하여, 정형화된 정보를 얻기 위해 웹 페이지의 구조를 탐색하고 HTML 형식의 코드를 분석하는 등 사람의 개입이 필요하다는 한계가 있다. 대형 언어 모델 에이전트는 다양한 분야에서 활용되고 있으며[2], 최근에는 이를 통해 웹 콘텐츠를 분석하는 도구[3], 계층 구조를 이용해 점진적 이해

* 국립금오공과대학교 컴퓨터공학과. wlgh6709@kumoh.ac.kr, bbbababb@kumoh.ac.kr,,
** 국립금오공과대학교 인공지능공학과, jyc@kumoh.ac.kr 교신저자

※ The work is supported by the Korea Basic Science Institute (Project No.A438200)

를 돕는 AutoCrawler 연구[4], 동적 및 정적 크롤러 병합 기술에 관한 연구[5]도 진행되었다. 본 연구에서는 정형 데이터를 효과적으로 수집하기 위해 대형언어모델 멀티 에이전트 시스템을 적용해 동적 및 정적을 결합한 웹 크롤러 개발 기법을 제안한다. 이 기법은 웹 페이지 구조 인식, 크롤러 코드 작성, 디버깅 작업을 사람의 개입 없이 에이전트가 반복적으로 수행하여, 적은 비용으로 정형화된 데이터를 추출하고, 수집할 수 있는 것을 보였다.

II. 에이전트 기반 크롤러 코드 생성

본 논문에서는 실험을 위해 GPT-4o-mini API를 에이전트로 활용하였다. 에이전트의 상태는 State function을 통해 관리하며, 이는 대화 기록, 프로세스 상태, 요청 횟수 등의 정보가 포함한다. 에이전트는 크롤링 코드를 구현하는 development 에이전트와 해당 코드를 실행하는 executor 에이전트로 구성된다. development 에이전트는 동적 에이전트와 정적 에이전트로 나뉘며, 각각 크롤링 방식에 따라 구현된다. 그림 1은 전체 프로세스의 구성을 나타낸 것이다.

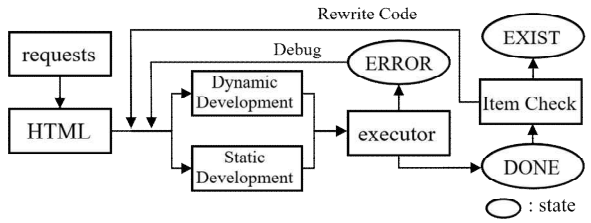


그림 1. 제안 기법 프로세스
Fig. 1. Process of the proposed technique

사용자가 수집을 원하는 사이트의 URL을 입력하면, requests 요청을 통해 HTM 형식의 사이트 정보를 응답받고, 이를 각 Development 에이전트에 제공한다. 동적 에이전트는 Selenium을 기반으로 하여 동적으로 데이터가 있는 페이지를 탐색하며, 정적 에이전트는 BeautifulSoup를 활용하여 정적 페이지 분석을 통해 구조화된 데이터를 추출하는 역할을 한다. 에이전트는 실행될 수 있는 python code를 생성하여 해당 코드는 executor 에이전트를 통해 실행된다. 실행 과정에서 오류가 발생하면 로그와 코드를 다시 제시하여 디버깅을 진행하고, 실행이 정상적으로 완료되면 작업을 마친다. 이후 수집된 Item을 확인한 후 Item이 부족하거나 잘못되면 추가 개발을 요구하고, 충분한 경우 프로세스를 종료하게 된다.

III. 평 가

평가는 nserc(https://www.nserc-crsng.gc.ca/ase-oro/index_eng.asp), nsf(https://www.nsf.gov/awardsearch/), 금오공과대학교 게시판(https://www.kumoh.ac.kr/ko/sub06_01_01_01.do), 네이버 지식인(https://kin.naver.com/qna/list.naver)네 개 사이트에 대해 항목(Item)을 올바르게 수집했는지에 대해 진행되었으며, 수작업으로 작성한 데이터를 정답으로 두고 두 결과를 비교하였다. 평가 항목은 수집된 Item의 수와 정보의 정확성 두 가지로 구성하였으며, 각각 1, 3회 반복 작업에 따른 결과 비교를 진행하였다.

표 1. 평가 결과
Table 1. Evaluation results

사이트	제안 기법의 수집 Item 수		정확히 수집한 Item 수		실제 Item 수
반복	1	3	1	3	
nserc	7	16	6	15	16
nsf	12	24	12	24	28
본교 게시판	7	7	7	7	7
지식인	1	1	1	1	4

위 결과와 같이, 제안한 기법이 높은 수집률과 정확도를 보였으며, 반복 과정을 수행한 후 수집률과 정확도가 더욱 향상됨을 확인할 수 있었다. 사이트의 형태가 어느 정도 구조화된 경우는 정보 대부분을 수집할 수 있는 코드가 작성되었으나, 주요 정보의 판단이 어려운 지식인 같은 경우는 코드 작성에 한계를 확인할 수 있었다. 또한, 본 기법은 1회 평균 0.0009$에서 3회 평균 0.0013$의 매우 적은 비용으로 크롤러를 구축할 수 있음을 확인하였다.

Ⅳ. 결 론

본 논문에서는 멀티 에이전트 기반으로 한 크롤러 코드 작성 기법을 제안하였다. 제안된 기법은 에이전트를 기반으로 하여 사람의 개입 없이 높은 수집률을 가지는 크롤러를 효율적으로 구축할 수 있음을 보였다. 그러나 사이트 구조가 복잡한 경우 데이터 수집을 위한 코드 작성에 한계가 존재하며, 이러한 문제는 추후 연구에서 다룰 예정이다.

참 고 문 헌

[1] https://en.wikipedia.org/wiki/Unstructured_data [Accessed: Oct. 21, 2024]

[2] A. Dorri, S. S. Kanhere, and R. Jurdak, "Multi-Agent Systems: A Survey," IEEE Access, vol. 6, pp. 28573–28593, 2018. doi: 10.1109/ACCESS.2018.2831228

[3] https://dev.to/jadouse5/building-an-openai-swarm-web-scraping-and-content-analysis-streamlit-web-app-with-multi-agent-systems-bl0 [Accessed: Oct 21, 2024]

[4] W. Huang, Z. Gu, C. Peng, Z. Li, J. Liang, Y. Xiao, L. Wen, and Z. Chen, "AutoScraper: A Progressive Understanding Web Agent for Web Scraper Generation," arXiv preprint, 2024. Available: http://arxiv.org/abs/2404.12753

[5] Q. Liu, R. Yahyapour, H. Liu, and Y. Hu, "A novel combining method of dynamic and static web crawler with parallel computing," Multimedia Tools and Applications, vol. 83, no. 21, pp. 60343–60364, 2024. doi: 10.1007/s11042-023-17925-y

[6] https://platform.openai.com/docs/models/gpt-4o-mini [Accessed: Oct. 22, 2024]