An Academic Information Collection System based on Large Language Model

Qiwei Lang
Software College, Jilin University, Changchun, China
langqw5520@mails.jlu.edu.cn

Shiqi Lyu
Software College, Jilin University, Changchun, China
lvsq5520@mails.jlu.edu.cn

Haoyi Wang
Software College, Jilin University, Changchun, China
wanghy5520@mails.jlu.edu.cn

Rui Zhang*
College of Computer Science and Technology, Jilin University, Changchun, China
*rui@jlu.edu.cn

2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI) | 979-8-3503-8694-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICECAI62591.2024.10675124

***Abstract*—To effectively assess a scholar's capabilities, it is essential to begin by gathering their basic information, forming the foundational step for subsequent analysis. This entails sourcing pertinent details such as the scholar's affiliation, title, research focus, and published papers from online sources. We curated a comprehensive dataset comprising the homepages of numerous scholars and develop a system to collect information. Our methodology involved harnessing search engines and webpage classifiers to locate scholars' homepages across the internet. Once identified, these pages underwent meticulous parsing to extract academic information. Recent advancements in Large Language Models (LLMs) have demonstrated remarkable proficiency across various domains, particularly in fundamental Natural Language Processing (NLP) tasks. By leveraging LLM capabilities, we conducted text extraction from webpage content, resulting in the successful compilation of academic data.**

Keywords-large language models; webpage classification; natural language processing

I. INTRODUCTION

Assessing the capabilities of scholars is a multifaceted endeavor that entails the collection, analysis, and interpretation of pertinent information. At its core lies the acquisition of fundamental scholar details, encompassing affiliations, academic titles, educational and professional backgrounds, as well as scholarly publications. These foundational data not only facilitate a deeper understanding of an individual scholar's contributions to academia but also serve as cornerstones for comparative analysis and the establishment of assessment frameworks. In today's era of pervasive internet connectivity, webpages have emerged as primary vessels of information dissemination. For scholars, their webpages serve as virtual portals, offering glimpses into their basic information and academic achievements. Thus, we have opted to focus on scholars' academic webpages as our data collection targets, procuring several such pages to construct a dataset of scholar webpages.

With the exponential growth in webpage volume, the Internet hosts an increasingly vast and intricate array of information [1]. This complexity renders the identification and localization of a scholar's webpage relatively challenging. Modern search engines provide effective tools for filtering Internet content, employing keywords to sift through webpages and ultimately return more relevant results. Following the initial filtration of webpage subsets using search engines, we employ our webpage classification method, PLM-GNN [2], to categorize these pages. Subsequently, based on the classification outcomes and predefined rules, we successfully identify the academic webpages of the targeted scholars.

Once these webpages are pinpointed, rigorous parsing is necessary to extract pertinent academic information. Large Language Models (LLMs) have demonstrated their exceptional ability across various applications [3,10,11] as LLMs have been trained on vast amount of textual data [4-6]. Owing to their zero-shot and few-shot abilities, LLMs achieve impressive performance on various foundational Natural Language Processing (NLP) tasks, thus serving as potent tools for task completion. Following the localization of academic webpages, we utilize webpage content extraction tools to extract the main contents of these pages and employed the formidable text parsing capabilities of LLMs to analyze the extracted text, ultimately gathering the scholarly information contained within the scholar webpages.

Based on the aforementioned efforts, we develop a system capable of locating scholars' academic webpages on the Internet based on their names and affiliations, then extracting the scholarly information contained within these pages. Our primary contributions are as follows:

1) We select scholars' academic webpages as targets for acquiring academic information, constructing a dataset of academic webpages to support both webpage classification and academic information collection. Additionally, we introduce the search engine and webpage classification methods for the identification and localization of corresponding academic webpages.

2) Leveraging the formidable general capabilities of LLMs, we parse the textual information contained within academic webpages to extract academic information.

3) We develop a system capable of locating scholars' webpages using their names and affiliations and utilizing large language models to extract the scholarly information pertaining to these scholars.

II. DATASET CONSTRUCTION

We select scholars' academic homepages as the target for gathering academic information, as these pages serve as the most comprehensive repositories reflecting their current status and past experiences. Whether for the construction of classifiers in identifying academic homepages or for the construction of samples in extracting academic information, a dataset containing academic homepages serves as the foundation. We curate our own dataset, named *Academic Homepages of Scholars* (AHS), which comprises two distinct components: AHS-scholars (AHS-s) and AHS-others (AHS-o). To facilitate academic information extraction, we crawled the academic homepages of faculty members from 22 universities, thereby constituting AHS-s. Subsequently, to further support webpage classification, specifically the identification of academic homepages, we utilized the webpages from AHS-s as positive samples. Additionally, we incrementally crawled webpages of NBA players and movie information to constitute AHS-o as negative samples. The detailed information of the AHS dataset is presented in Table 1.

Table 1. Information of The Ahs Dataset

Dataset	Split	#Pages	Language
AHS	AHS-s	8937	Chinese, English
	AHS-o	2307	

III. METHODOLOGY

In this section, we will elucidate the two most crucial components of the methodology employed in the entire academic information retrieval system: firstly, the approach to locating a scholar's academic homepage; and secondly, the process of parsing textual content from the homepage using LLM.

A. Identification of the Academic Webpage

Before parsing the homepage, it is imperative to first locate and identify it. Given the vast number of webpages across the Internet, discerning the pertinent academic homepage can be a formidable task. Recognizing the utility of search engines in conducting keyword-based searches across the Internet, we initially incorporate search engine functionality for preliminary web page screening. Following the retrieval results by the search engine, we extract the links contained within the top N retrieved pages and subsequently request the corresponding webpage source codes.

After obtaining the initial retrieval results, we further categorize these web pages to facilitate the screening of academic homepages. We employ the PLM-GNN [2] for classification. PLM-GNN models webpages based on textual information and the graph information contained in the HTML DOM tree, and constructs classifiers based on both text and graph features. Specifically, let W denote the set comprising all web pages retrieved from the search engine. Given a webpage $w \in W$, PLM-GNN firstly parsing it into an HTML DOM tree $\mathcal{T} = (\mathcal{N}, \mathcal{C}, \mathcal{R})$, where $\mathcal{N}$ denotes the nodes set, $\mathcal{C}$ denotes the contents which contained in the HTML document, and $\mathcal{R}$ denotes the relation between the nodes. Following the parsing of the DOM tree, PLM-GNN extracts the textual content T from the HTML and utilizes a pretrained language model (PLM) as a text encoder to obtain text embeddings x_T. Simultaneously, PLM-GNN constructs a graph $\mathcal{G}$ based on $\mathcal{T}$ and employs a graph neural network (GNN) as a graph encoder to obtain the graph representation x_G. After regularization and concatenation of features, the final webpage feature x_H is obtained, which is then fed into a multi-layer perceptron for classification.

B. Academic Information Extraction via LLMs

LLMs excel at understanding textual information and completing downstream NLP tasks. Due to their remarkable zero-shot capability, simple prompts suffice to engage them in tasks effectively. For a webpage $w \in W$, we initially employ a main content extractor $\varphi(\cdot)$ to retrieve the main content from the webpage, yielding text $T' = \varphi(T)$. Subsequently, we utilize the LLM to comprehend and extract information from T'. Formally, denoting the LLM as $M(x, c; \theta)$, where x represents the input text, c denotes the prompt, and θ signifies the model parameters. Taking T as input, we obtain the generated extraction result T_g, i.e.,

$$T_g = M(\varphi(T), c; \theta) \tag{1}$$

Since different prompts may affect the performance, we conducted experiments and ultimately standardized the prompts into a unified format. The prompt template designed for the task of collecting academic homepage information consists of three main parts: role setting, task definition, and guideline instructions. Specifically, considering the role-playing ability of LLMs [7] and the enhancement of task performance when the model is set to a role highly relevant to the task, we include the role-setting part in the prompt. Additionally, owing to the model's strong ability to follow instructions, acquired through instruction-following fine-tuning during training, we include the task definition and guideline instructions in the prompt to facilitate the model in accomplishing the task more effectively.

IV. EXPERIMENTS

In this chapter, our primary aim is to address the following two inquiries through experimentation: 1) Can the webpage identification method outlined in Section 3.1 effectively categorize academic homepages? 2) Is LLM proficient in extracting information effectively from academic homepages?

A. Can the Identification Method effectively Categorize Academic Homepages?

Dataset. We conduct experiments on the entire AHS dataset using the PLM-GNN mentioned in Section 3.1 for webpage classification tasks to assess its effectiveness in distinguishing academic homepages from other webpages. To ensure label balance, we sample 2000 samples on two partitions and split them into training and testing sets in an 8:2 ratio.

Setup. We configure the task as a binary classification task and train models on the partitioned training data. We choose RoBERTa and Longformer as text encoders, utilizing *sum*$(\cdot)$ and *max*$(\cdot)$ as readout functions, respectively. Additionally, a two-layer MLP is employed as the classification head.

Evaluation Metrics. Given that the experimental task is a classification task, we employ accuracy, precision, and F1 score as evaluation metrics to gauge the model's classification performance.

Table 2. Model Performance on Webpage Classification Task

Model	A.	R.	P.	F1
RoBERTa+sum	1.000	0.999	0.999	0.999
Longformer+sum	0.999	1.000	0.998	0.999
RoBERTa+max	0.989	0.989	0.991	0.997
Longformer+max	0.848	0.620	0.921	0.649

Result and Analysis. Table 2 illustrates the results of PLM-GNN on webpage classification tasks, with evaluation metrics denoted by A., R., and P. for accuracy, recall, and precision, respectively. It is evident that both models perform well when utilizing the summation function as the readout function. However, there is a noticeable decrease in performance when employing the maximum function as the readout function, particularly evident with the Longformer model. Additionally, the effectiveness of the $max(\cdot)$ as the readout function is inferior to that of the $sum(\cdot)$. We speculate that the reason for this phenomenon may be attributed to overfitting caused by the excessive model capacity of Longformer. Moreover, the presence of abundant noise text irrelevant to the main content in webpages may interfere with text representation, underscoring the necessity of eliminating irrelevant noise from webpages in practical applications.

The webpage classification results still belong to a subset of scholar homepages. We further experimented on how to precisely filter the classified set to obtain the scholars' main homepages, which will be discussed in Section 5.1 on constructing the mentioned localization module.

B. Is LLM Proficient in Extracting Information effectively from Academic Homepages?

Dataset. We conduct experiments on the AHS-s split using the methods mentioned in Section 3.2. Since the AHS dataset is bilingual, we need to test the model's performance on different languages separately. We partition AHS-s into Chinese and English segments and randomly sampled 100 pages from scholar homepages in each language as the test set.

Setup. We conduct tests on the sampled test sets for information extraction tasks. Considering that we are dealing with two languages, Chinese and English, we choose the open-source and well-capable ChatGLM2-6B [9] as the base model. We utilized the prompts and concatenated webpage texts designed in Section 3.2 as inputs and performed inference on four NVIDIA V100 GPUs. Due to the excessively long textual information in individual webpages, exceeding the context window length supported by the model, we segment the input at the 8k context supported by ChatGLM2-6B.

Evaluation Metrics. Since LLM essentially accomplishes various tasks through generation, we evaluate the model's performance using evaluation metrics for natural language generation tasks. We select BLEU and ROUGE to assess the overlap between generated content and reference. Additionally, since the output of LLM is subject to randomness, and some generated responses may not adhere to the JSON format requirement, we consider responses that do not meet the format requirement as erroneous generations. We evaluat the stability of LLM's generation by calculating the error rate in generating responses and complying with the JSON format requirement as enforced in the prompts.

Table 3. Model Performance on Ahs-s

Segment	B-1	B-2	B-4	R-1	R-2	R-L	Err
Zh-CN	92.54	92.53	92.40	96.38	94.45	96.02	5%
EN	90.63	90.17	91.13	93.41	93.56	94.19	8%

Results and Analysis. Table 3 presents the results of LLM on the English and Chinese segments of AHS-s, where BLEU and ROUGE are denoted as “B” and “R” respectively, and Err represents the ratio of erroneously generated samples by the model. It can be observed that the model performs relatively well on the Chinese segment but slightly less so on the English segment. Additionally, the model is more prone to generating responses that do not adhere to the JSON format on English data. The task of extracting structured information from webpages itself is relatively complex, as structured information is hierarchical, requiring the model to first comprehend the input text before extracting and generating nested hierarchical textual information. Moreover, complex tasks tend to induce hallucinations in the model to some extent. Analysis of the generated content reveals that the model fabricates some information while extracting paper details from the webpage, leading to a decrease in model performance. Furthermore, since we interacted with the model using uniform Chinese prompts, the model's ability to understand English content in zero-shot scenarios is somewhat diminished.

We additionally explored the effect of the model after deleting the different prompt components, as shown in Table 4.

Table 4. Ablation Study of Prompt Components

Prompt	B-1	B-2	B-4	R-1	R-2	R-L	Err
w.o. Role	53.19	50.52	47.29	56.11	50.67	52.30	10%
w.o. Task	63.20	61.43	58.95	63.90	57.45	60.67	10%
w.o Gui.	53.41	48.61	42.40	35.44	27.31	30.96	11%

In the table, “Role” denotes the role setting, “Task” denotes the task definition, and “Gui.” denotes the guideline instruction. It can be seen that the performance of the model decreases significantly when any part of the cue is removed, while the error rate also increases. Among them, removing the guideline instruction has the most significant decrease, removing the role setting has the second largest decrease, and removing the task definition has the smallest decrease. The reason for this phenomenon is that, firstly, the instruction is the most important part of the cue, and LLM uses the instruction to follow the fine-tuning during training. Second, as mentioned in Section 3.2, LLM has a strong role-playing capability, and adding role setting to the prompt can indeed significantly improve the model performance from the performance point of view.

V. THE ACADEMIC INFORMATION COLLECTION SYSTEM

In the preceding section, we delineated two pivotal methods employed in our academic information collection system: utilizing search engines and webpage classifiers to identify academic homepages, and employing LLMs for academic information extraction. In this section, we will introduce the system we have developed for collecting academic information, which is capable of locating and extracting academic information from specified scholars' homepages. Figure 1 shows the overall framework of the academic information system. This system primarily comprises two major components: a localization module, which searches and locates the homepages of specified scholars on the Internet based on input; an extraction module, which extracts academic information from the homepages retrieved by the localization module, ultimately yielding the academic information of the scholars.

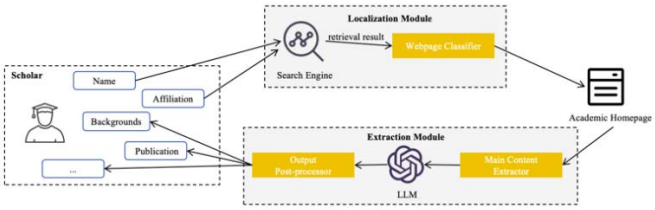

Figure 1: Overview of the Academic Information Collection System

A. Localization Module

In the academic information collection pipeline and the practical system, the initial step involves locating scholars' academic homepages on the Internet. We employ the method mentioned in Section 3.1 to accomplish this task. Considering the possibility of name ambiguity among scholars in different affiliation, we aim to mitigate the issue of the ambiguity. We did the empirical investigation and discovered that combining a scholar's name with their affiliated institution serves as a robust identifier, effectively circumventing ambiguity issues. Utilizing this combination as search engine keywords facilitates efficient webpage filtering.

Based on the aforementioned analysis, our entire system takes the scholar's name and affiliation as input. Initially, we invoke a search engine for retrieval, setting $N = 5$ to retrieve the top 5 pages of search results. Subsequently, we obtain and locally store the source code of all pages contained in the results via their respective page links. To classify the retrieved pages, we employ the PLM-GNN method for classification, leveraging the AHS dataset for training. Furthermore, we use RoBERTa [8] as the text encoder, and the summation function is utilized as the GNN's readout function.

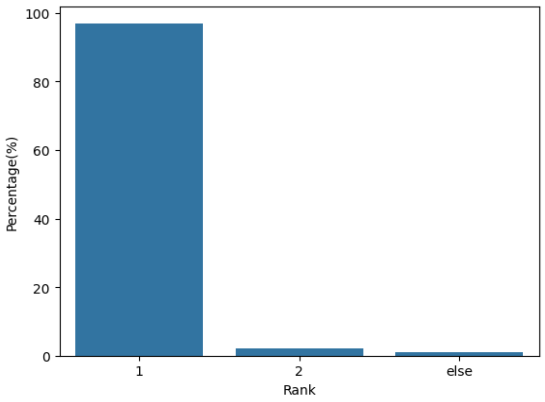

Figure 2: Scholar Homepage Ranking

After classification, we obtain a collection of pages and need to determine which one corresponds to the scholar. To identify the final academic homepages, we conduct experiments on 100 scholars randomly select from AHS-s. Combining the retrieval ranking from search engines with the results of the classifier inference, we perform manual evaluation. The outcomes are depicted in Figure 2, where it was observed that the webpage most likely to be ranked first is the academic homepage we aimed to locate. Thus, through the process before, we successfully locate the scholar's homepage based on his name and affiliation.

B. Extraction Module

After locating the academic homepages, we proceed to parse them using the method described in Section 3.2. The extraction module primarily comprises three components: the main content extractor, the text parser, and the output post-processor.

For the pages obtained from the localization module, we initially utilize the main content extraction tool *newspaper3k* to extract the webpage. The rationale is that for LLMs, their context window is limited. Webpages often contain noisy text content, which not only wastes valuable context window but also interferes with extraction performance as noise. *Newspaper3k* is a text extraction library designed for news webpages, effectively filtering out irrelevant text such as navigation bars and extracting the main content. During the deployment of the main content extractor, we observed that when processing pages with many tables, the tool may overlook some content within the tables. Consequently, we separately extract the content from tables and input it into the text parser.

Here we continue to utilize ChatGLM2-6B as the base model of text parser, which demonstrates proficient understanding in both languages. We configure the model's prompts according to the format mentioned in Section 3.2. We pre-defined some academic information entities in the prompt of LLM, such as names, institutions, titles, and publications, in order to ensure precise extraction. Moreover, to maintain uniformity in the generated responses, we instruct the model to output in JSON format. It's important to note that due to constraints on context length and inference time, we did not include examples of text extraction in the prompts for in-contextual learning (ICL). The primary reasons are: 1) the main content of webpages nearly fills

the input context window, leaving no space for the inclusion of ICL mechanisms, and 2) increasing the input length significantly raises inference time and memory consumption, which is highly undesirable for single-sample inference.

The construction of the output post-processor in the extraction module is relatively straightforward. Since the output of LLMs exhibits some randomness, the output of the text parser may include descriptive content. We apply regular expressions to filter out unnecessary descriptive text from the output, thus obtaining the final extracted academic information.

VI. CONCLUSION AND FUTURE WORK

In support of assessing the capability of scholars, we devise a comprehensive pipeline for collecting scholars' academic information and develop a system to execute this information collection. Firstly, we construct our own academic homepage dataset, AHS, to facilitate the classification of academic homepages and extraction of academic information. Secondly, we design methodologies for locating academic homepages on the Internet and extracting the information contained therein. We introduce search engines and webpage classifiers to identify academic homepages, and leveraged the powerful capabilities of LLMs to parse the extracted content.

The entire academic information collection system comprises two main modules: the localization module and the extraction module. In the localization module, we utilize scholars' names and affiliations as identifiers, employing them as keywords to search through search engines. We employ the webpage classification method PLM-GNN along with manually defined rules to ultimately pinpoint the academic homepages. In the extraction module, we extract the main content of webpages and employ ChatGLM2-6B to parse the textual content. In the future, we aim to further optimize and iterate upon the entire system to develop a system capable of efficiently and accurately collecting scholars' academic information.

REFERENCES

[1] L. Deri, M. Martinelli, D. Sartiano, and L. Sideri, “Large Scale Web-Content Classification:,” in Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, Portugal: SCITEPRESS - Science and and Technology Publications, 2015, pp. 545–554.

[2] Q. Lang, J. Zhou, H. Wang, S. Lyu, and R. Zhang, “PLM-GNN: A Webpage Classification Method based on Joint Pre-trained Language Model and Graph Neural Network.” arXiv, May 09, 2023. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2305.05378

[3] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23).

[4] J.Hoffmann et al., “Training compute-optimal large language models,” in Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 2176, 30016–30030.

[5] S. Min et al., “Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?,” in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11048–11064.

[6] P. Liang et al., “Holistic Evaluation of Language Models.” arXiv, Oct. 01, 2023. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2211.09110

[7] Z. M. Wang et al., “RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models.” arXiv, Oct. 01, 2023. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2310.00746

[8] Y. Liu et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv, Jul. 26, 2019. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/1907.11692

[9] A. Zeng et al., “GLM-130B: An Open Bilingual Pre-trained Model.” arXiv, Oct. 25, 2023. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2210.02414

[10] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, “Is ChatGPT a General-Purpose Natural Language Processing Task Solver?,” in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1339–1384.

[11] B. Rozière et al., “Code Llama: Open Foundation Models for Code.” arXiv, Jan. 31, 2024. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2308.12950