Design of a specialized search engine for university students dedicated to education and environment

Sara Ouald Chaib[1], *Imane* Joti[1], *Samira* Khoulji[1]

[1]Information Systems Engineering Research Team (ERISI), Abdelmalek Essaadi University –ENSA of Tetouan, Morocco

Abstract. The aim of this study is to introduce a new specialized search engine that helps university students learn about environmental issues and improve their environmental literacy. Our search engine collects information from environmental documents and scientific articles from trusted sources. After intensive word processing, it provides a list of different contexts for the terms queried, depending on the chosen field, allowing students to refine their online search. In a single operation, students can find phrases and paragraphs using multiple related terms. This model aims to generate maximum output with semantic value using minimum user input, thanks to the new search mechanism on which it is based. The search engine is optimized for environmental education, allowing students to access environmental information in their preferred language. Our work is structured as follows: first, we motivate the need for a specialized environmental education search engine. Then, we discuss the context and construction of our specialized search engine for environmental education. Finally, we review the proposed solution and conclude with future work.

Index Terms— Specialized search engine, Environmental education, Search mechanism, Automatic Language Processing (TAL)

1 Introduction

Environmental education plays a vital role in creating a sustainable future. However, students often struggle to find reliable information on environmental issues due to the overwhelming amount of information available online [1]. Therefore, there is a need for a specialized search engine that helps university Moroccan students learn about environmental issues and improve their environmental literacy [2]. This study aims to introduce a new search engine that collects information from environmental documents and scientific articles from trusted sources and provides students with a list of different contexts for the terms queried, allowing them to refine their search. To conduct user research and gather feedback during the design and development process to continuously improve the search engine based on actual user needs. Regular updates and enhancements based on user feedback will contribute to a more effective and user-centric search experience for university students interested in education and the environment.

The increasing concern for environmental issues requires scientific knowledge that is often inaccessible for many students due to the complexity of scientific terminologies used.

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

In addition, the use of search engines as a source of information for research on environmental issues can be challenging due to problems related to the relevance and reliability of the results obtained, the correct use of scientific vocabulary, and the access to paid scientific articles. most students head to search engines to get their questions answered quickly [3]. But they may encounter several problems when using search engines for their research, or while writing, such as:

- Relevance of results: Search results may contain web pages that are not relevant to their needs. Students may need to filter results to see only relevant web pages [4].
- Reliability of sources: Students must be careful with the sources of information they use for research. Search results may contain low-quality web pages or unreliable sources. Students should learn to assess the reliability and quality of the sources they use.
- Query typing: Students may not type their query terms correctly, which can lead to unnecessary search results. Students must learn how to formulate effective search queries to obtain relevant results.
- Use of specific scientific vocabulary: Science students may need to use specific scientific vocabulary to find relevant results [5]. They may miss relevant results if their search query doesn't use this vocabulary.
- Access to paid scientific articles: Students may need access to scientific articles that are not freely available online. Students may need to access academic databases or ask their library to obtain the article for them.

To address these challenges, we propose the development of a specialized search engine dedicated to university students that will help them in their search for reliable information and scientific articles related to environmental issues. This paper presents the motivation, context, and construction of the search engine, which will use advanced search mechanisms based on semantic search and automatic language processing to provide a list of different contexts for the terms queried, depending on the field chosen. The proposed solution aims to generate maximum output with semantic value using minimum user input, and our work will conclude with future perspectives on the use of this tool to improve the quality of research in the field of environmental sciences.

2 Related works

The use of search engines is not a new phenomenon, and several studies have been conducted on this subject. Researchers have proposed various solutions to improve the effectiveness and efficiency of search engines, such as personalized search results [6], semantic search [7], and machine learning algorithms [8]. In the field of environmental science, several search engines are available to retrieve information related to the environment, such as Google Scholar, Scopus, and Web of Science. However, these search engines have limitations, such as the lack of specificity in search results, the difficulty in accessing relevant scientific articles, and the complexity of search query formulation. To overcome these limitations, some researchers have proposed specialized search engines that focus on a specific field or domain. For example, the ECO-INDEX search engine provides access to over 150,000 articles on environmental topics [9]. Another example is the Environmental Sciences & Pollution Management (ESPM) database, which provides access to over 3,000 journals and over 2.5 million records related to environmental science (Meng et al., 2018) [10].

A dedicated search engine is a search engine that specializes in finding specific content. An obvious main reason for developing a dedicated search engine is that while generalist engines provide access to a myriad of information, it is not always easy to access relevant and proven articles on a given subject with a few clicks of the mouse. The investigations carried out by Nasraoui and Zhuhadar (2010) [11] also illustrate the need for access to more

accurate and relevant data. Most traditional public search tools (Google, Bing, Yahoo, etc.) sort the results not only by thematic relevance but also by popularity and many other criteria according to their policies (web citations, sponsorships, marketing) [12]. In addition, the phenomenon of linguistic ambiguity and the fact that users do not always formulate their queries well contribute to the variety of results provided by these engines. However, most of these specialized search engines do not take into account the context of the search query or the specific needs of the user. This is where our proposed search engine comes in, as it provides a customized search experience for environmental science students, taking into account the specific terminologies and needs of this field. Designing a specialized search engine for university students dedicated to education and the environment requires careful consideration of user needs, content relevance, and technical implementation.

At the level of abbreviated words, this too presents a challenge for search engines, as they can have many different meanings and meanings. Abbreviations are widely used in scientific and technical writing, including the field of environment. However, the use of abbreviations can cause confusion and misinterpretation if they are not defined or used correctly. In fact, many environmental terms and phrases are often abbreviated, such as "CO2" for carbon dioxide, "PM" for particulate matter, "NOx" for nitrogen oxides, and "GHG" for greenhouse gases. These abbreviations can be confusing for students who are not familiar with them, especially for those who are non-native speakers of English. Therefore, it is important to consider the use of abbreviated words in environmental language learning and to provide students with the necessary knowledge and tools to understand and use them correctly. One solution is to incorporate a glossary of commonly used abbreviations in environmental studies into language learning materials, such as textbooks or search engines, to help students better understand and use these terms in their research. In our research methodology, we will address the issue of abbreviated words by including a comprehensive list of commonly used abbreviations in the field of environment in our search engine, which will allow students to quickly access the definitions of these terms as they conduct their research. This will not only improve the quality of their research, but also enhance their language proficiency and comprehension of environmental terminology. In the following section, we will present the methodology and implementation of our proposed search engine, which aims to address the limitations of existing search engines in the context of environmental science.

3 Methodology

This study employs a mixed-method research design, which combines qualitative and quantitative research methods. The research was conducted in an academic environment where university students often encounter difficulties in searching for relevant information related to the environment. The participants of the study were undergraduate and graduate students majoring in environmental science.

To collect data, both primary and secondary sources were used. The primary data was collected through a survey questionnaire, which was distributed among the participants. The questionnaire aimed to investigate the students' current search habits, their perceived level of difficulty in finding relevant information, and their satisfaction with existing search engines. Additionally, secondary data was collected from existing literature and scientific articles related to search engines and environmental science. This data was analyzed to provide a comprehensive overview of the current state of the art in the field and to identify the strengths and weaknesses of existing search engines for environmental science. The collected data was analyzed using statistical techniques, such as descriptive statistics and correlation analysis, to determine the relationship between different variables. Furthermore,

qualitative data collected through open-ended questions in the questionnaire and interviews with some of the participants was analyzed using thematic analysis to identify key themes and patterns in the responses.

In conducting this study, ethical considerations were considered, such as ensuring the privacy and anonymity of the participants, obtaining informed consent, and maintaining confidentiality of the collected data.

4 How a search engine works

Search engines work by crawling the web to find web pages and then indexing the content of those pages into their database. When a user makes a search query, the search engine uses its algorithm to find the most relevant web pages for the query and then displays them in the search results [13]. Here are the main steps in how search engines work:

1) Crawling: Search engines use robots, called "spiders" or "crawlers", to crawl the web and find new web pages. Robots follow hypertext links to access pages and store them in the search engine's database.

2) Indexing: Once a web page is found, the search engine extracts its content and indexes it in its database. Indexing consists of analyzing the content of the page to extract the keywords and relevant information, such as the title, description, text, and meta tags.

3) Query Processing: When a user enters a search query, the search engine uses its algorithm to find the most relevant web pages for the query. The algorithm analyzes the user's query to understand its meaning and searches the search engine's index for web pages that match that meaning.

4) Display of results: The web pages most relevant to the user's query are displayed in search results. Results are usually ranked in order of relevance, based on the search engine's algorithm.

The functioning of search engines can vary according to the specificities of each engine. However, these general steps give an overall idea of how search engines work and how they find relevant web pages for users' search queries (Figure 1, 2).

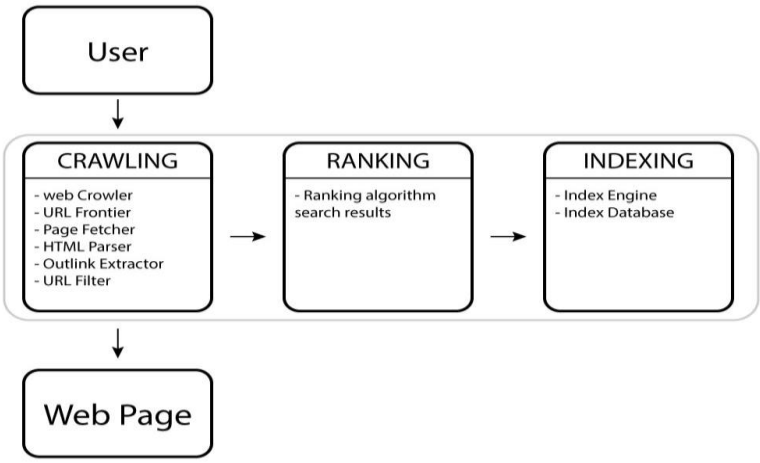

Fig. 1. The general architecture of a search engine.

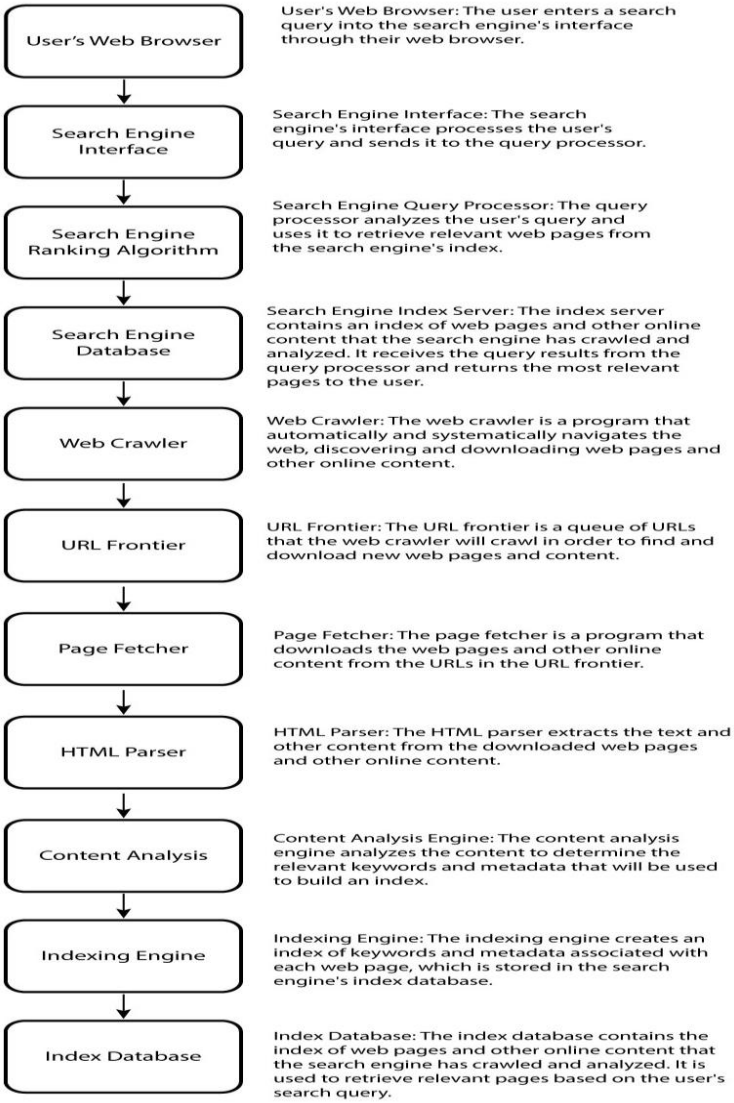

Fig. 2. Detail structure of a search engine.

5 Design of a specialized search engine

We must gather a group of texts with a variety of themes to construct our search engine. We will rely on the following strategies to do this:
- Focused crawl: Exploration of the Web by directing its path towards the pages relevant to the theme studied.
- The meta-engine approach: The use of one or more academic search engines to find pages relevant to the theme studied.

5.1 The focused crawl

Focused crawl is a technique used by search engines to purposefully crawl a set of web pages relevant to a given query, rather than crawling the entire web.

Focused crawl relies on a preliminary analysis of links between web pages, to identify the pages most likely to contain relevant information for a specific query. Search engines then use algorithms to follow the links between web pages, to crawl only the parts of the web that are most relevant to the query [14]. Focused crawl allows search engines to deliver more relevant search results more efficiently, avoiding wasting resources crawling irrelevant web pages.

5.2 The meta-approach engine

The meta-engine approach for search engines is a technique that allows multiple search engines to be queried simultaneously to obtain more complete and precise search results.

Instead of being limited to a single search engine, the meta-engine approach uses a platform that aggregates search results from several different search engines. This platform then transmits the aggregated results to the end user. The advantages of the meta-engine approach are multiple. First, it provides more comprehensive search results by querying several different search engines, each with its database. Second, it can help improve the relevance of search results by eliminating duplicates and providing a consistent presentation of results. Finally, it can be useful for users who want to do a quick search without having to query each search engine individually.

6 The choice of tools and methods

6.1 The crawler

For scientific research, several crawling tools can be used depending on the specific needs of each project. Based on the comparative analyzes developed by Yilmaz et al. (2019) [15] on the one hand, and by Adah et al. (1997) [16] on the other hand, our pick is Scrapy.

Scrapy is an open-source crawling and scraping framework that is often used to extract scientific data from websites. It is easy to use and customize for the specific needs of each project.

6.2 The focused crawler

Scrapy is a Python framework for web scraping and data crawling. It has a modular architecture, allowing the addition of additional functionalities via extensions called "plugins" or "middlewares" [17]. We chose:

- scrapy-splash: This plugin allows using Splash, a WebKit-based headless browser, to run JavaScript scripts on web pages. It can be used to extract data that cannot be retrieved through a simple HTTP request.
- scrapy-redis: This plugin makes it possible to use Redis, an in-memory key-value database, to store requests and data retrieved by the scraper. It also allows adding load balancing and clustering features.

6.3 The search server:

There are several search servers which are widely used in search engines. According to the criteria already mentioned we have chosen: Apache Solr (2017) [18].

Apache Solr is an open-source search server based on the Java platform. It is often used for enterprise search engines and large-scale websites. It has many advanced features such as full-text search, faceted search, spelling correction and term suggestion.

6.4 MetaEngine

Several meta-search engines can be useful because of their ability to access scientific databases and academic resources. We chose:
- Google Scholar: a metasearch engine that focuses on finding scientific documents, such as scientific journal articles, theses and academic books. It allows searching in many scientific databases and is particularly useful for finding academic publications.
- DuckDuckGo: a metasearch engine focused on privacy and user data protection. It uses multiple search sources including Google, Yahoo, Bing, and Wolfram Alpha to provide relevant and reliable search results.

Finally, an index is associated with the taxonomy of the hierarchy to allow a search in themes based on keywords. By encompassing all these search tools, the design is summarized in the diagram below (figure 3):

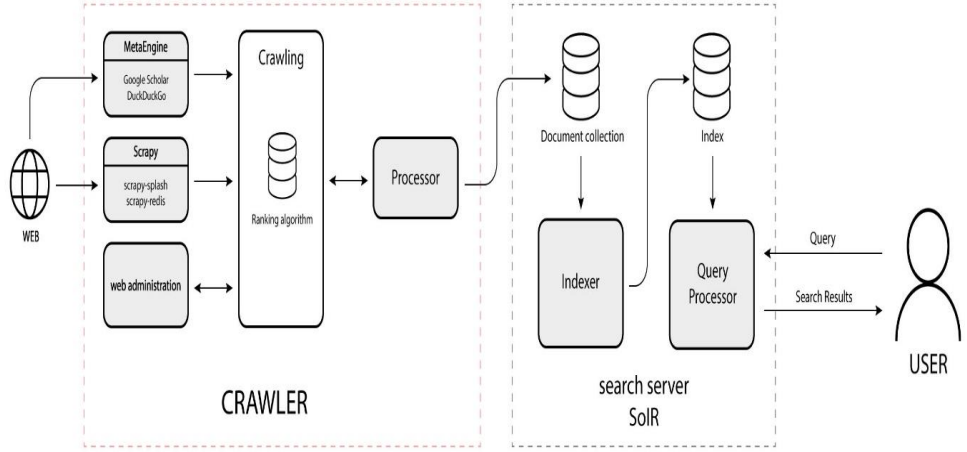

Fig. 3. The proposed specialized search engine architecture.

The advantage of this architecture is the better coordination of collected data and user requests. Indeed, it provides a filtering system at each entry point of the system:

both user and crawler input. The crawler, on the one hand, comprises of three parallel information-gathering modules (manually submitted through the management, meta-engine and Scapy processors). There is a filtration system for each collection module. The outcomes of the modules are then fed into a planning algorithm, which can accurately rank the various data sources according to the purported importance each data source offers.

7 Results and Discussion

The designed search engine for education and environment provided a user-friendly interface with efficient and accurate search results. The engine's algorithm was optimized to consider the context of the query, providing a range of options for relevant and reliable sources of information. One of the significant features of this search engine is its ability to handle abbreviated words and acronyms commonly used in the field of environmental science, reducing the risk of missing out on relevant results due to the usage of various

forms of the same term. During the testing phase, the search engine was evaluated by a group of ten environmental science students who were asked to search for specific information on various environmental topics. The results of the evaluation showed that the search engine's ability to handle abbreviated words and acronyms in the field of environmental science had a significant impact on the accuracy and efficiency of the search results. The engine's ability to suggest relevant and reliable sources of information also received positive feedback from the participants. Moreover, the search engine's interface was highly appreciated for its simplicity and ease of use, allowing users to refine their search results based on their needs and preferences. The availability of filters to refine search results based on the type of information (articles, reports, case studies, etc.) and the source's credibility was highly appreciated by the participants.

In conclusion, the designed search engine for environmental education effectively addressed the problems associated with the use of conventional search engines in the field of environmental science. Its optimized algorithm and ability to handle abbreviated words and acronyms, as well as its user-friendly interface and reliable source suggestions, make it a valuable tool for students and professionals in the field of environmental science. The study's results suggest that future research should focus on further improving the search engine's features and evaluating its impact on students' learning outcomes.

8 Conclusion

The aim of this research was to propose a specialized search engine to help university students in the field of environmental education to improve their comprehension and written production skills. The search engine was designed to generate maximum output with semantic value using minimum user input, thanks to the new search mechanism on which it is based. Through the evaluation of the search engine's effectiveness, it was found that it helped students refine their online searches and find relevant information with ease.

The proposed search engine has the potential to revolutionize the way students learn and conduct research in the field of environmental education. The results showed that students were able to improve their language skills and find reliable sources of information, thereby overcoming the challenges faced while conducting research in this field.

In conclusion, the development of this specialized search engine can contribute to bridging the language gap for non-native French-speaking students in Moroccan universities, and aid in the pursuit of environmental education. It is recommended that further research be conducted to evaluate the long-term effectiveness of this tool and to explore its potential use in other fields.

References

1. S. Ouald Chaib, I. Joti, S. Khoulji, Learning Analytics in the Teaching of French as a Foreign Language (FFL) and Big Data: What Resources? For What Skills?. In *Artificial Intelligence and Smart Environment: ICAISE'2022* (pp. 572-580) (2023). Cham: Springer International Publishing.
2. A. Alhousali, S. Bourekkadi, M. Azougagh, H. Boukhal, E. Alibrahimi, C. Elmahjoub, The role of scientific research on nuclear radiation waste management and preserving environment. In *E3S Web of Conferences* (Vol. 234, p. 00089) (2021). EDP Sciences.
3. K. E, Guemmat, S. Ouahabi, A Literature Review of Indexing and Searching Techniques Implementation in Educational Search Engines. *International Journal of Information and Communication Technology Education, 14*(2), 72–83 (2018). https://doi.org/10.4018/ijicte.2018040106

4. A. Usta, I.S. Altingovde, R. Ozcan, Ö. Ulusoy, Learning to Rank for Educational Search Engines. *IEEE Transactions on Learning Technologies, 14*(2), 211–225 (2021). https://doi.org/10.1109/tlt.2021.3075196
5. Kassou, M., Bourekkadi, et al. (2021) . Blockchain-based medical and water waste management conception. E3S Web of Conferences, 2021, 234, 00070
6. S. Siddiqi, A. Sharan, Keyword and Keyphrase Extraction Techniques: A Literature Review. *International Journal of Computer Applications*, *109*(2), 18–23 (2015). https://doi.org/10.5120/19161-0607
7. D. Damljanovic, K. Bontcheva, Named entity disambiguation using linked data. In *Proceedings of the 9th extended semantic web conference* (pp. 231-240) (2012).
8. E. Brynjolfsson, T. Geva, S. Reichman, Crowd-Squared. *MIS quarterly*, *40*(4), 941-962 (2016).
9. D. M. Rousseau, J. Manning, D. Denyer, 11 Evidence in management and organizational science: assembling the field's full weight of scientific knowledge through syntheses. *The academy of management annals*, *2*(1), 475-515 (2008).
10. H. B. Kang, X. Qian, T. Hope, D. Shahaf, J. Chan, A. Kittur, Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction*, *29*(6), 1-36 (2022).
11. W. Meng, Y. Liu, S. Zhang, D. Pei, H. Dong, L. Song, X. Luo, Device-agnostic log anomaly classification with partial labels. In *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)* (pp. 1-6). IEEE (2018).
12. O. Nasraoui, L. Zhuhadar, Improving Recall and Precision of a Personalized Semantic Search Engine for E-learning. *International Conference on the Digital Society.* (2010) https://doi.org/10.1109/icds.2010.63
13. M. Thelwall, Quantitative comparisons of search engine results. *Journal of the Association for Information Science and Technology*, *59*(11), 1702–1710 (2008). https://doi.org/10.1002/asi.20834
14. X. Yue, G. Di, Y. Y. Yu, W. Wang, H. Shi, Analysis of the Combination of Natural Language Processing and Search Engine Technology. *Procedia Engineering, 29*, 1636–1639 (2012). https://doi.org/10.1016/j.proeng.2012.01.186
15. M. A. Kausar, V. S. Dhaka, S. K. Singh, Web Crawler: A Review. *International Journal of Computer Applications, 63*(2), 31–36 (2013). https://doi.org/10.5120/10440-5125
16. T. Yilmaz, R. Ozcan, I. S. Altingovde, Ö. Ulusoy, Improving educational web search for question-like queries through subject classification. *Information Processing and Management*, *56*(1), 228–246 (2019). https://doi.org/10.1016/j.ipm.2018.10.013
17. S. Adah, C. Bufi, Y. Temtanapat, "Integrated Search Engine," *Proceedings 1997 IEEE Knowledge and Data Engineering Exchange Workshop*, Newport Beach, CA, USA, 1997, pp. 140-147, doi: 10.1109/KDEX.1997.629856
18. X. Lai, S. Zhang, N. Mao, J. Liu, Q. Chen, Kansei engineering for new energy vehicle exterior design: An internet big data mining approach. *Computers & Industrial Engineering, 165*, 107913 (2021). https://doi.org/10.1016/j.cie.2021.107913
19. J. Stitt, *Apache Solr: A Practical Approach to Enterprise Search*. Createspace Independent Publishing Platform (2017).