Chapter 25
InfoScout: A Software Agent for Automated Web Information Retrieval

Saw Ming Sheng & Wan Hussain Wan Ishak*

School of Computing, Universiti Utara Malaysia, Sintok, Kedah, Malaysia

**hussain@uum.edu.my*

ABSTRACT
The exponential growth of web-based information has significantly increased the complexity of locating relevant, accurate, and timely content, particularly for students and novice researchers. Traditional manual search approaches are often time-consuming, cognitively demanding, and prone to information overload, leading to reduced productivity and inefficient knowledge acquisition. This project addresses these challenges through the development of a Web Crawler for Automated Information Retrieval, designed as a software agent that systematically extracts and categorises web links based on user-defined URLs and keywords. The primary objective of the study is to design, develop, and evaluate an automated web crawler that improves efficiency and accuracy in information retrieval while reducing user effort. The proposed system introduces a lightweight yet practical novelty by integrating rule-based keyword matching, structured output presentation, and persistent database storage within a user-friendly desktop application, targeting non-expert users without requiring advanced technical skills. The system was developed using Java, JavaFX, and Selenium, and evaluated through User Acceptance Testing involving 30 undergraduate students. Findings indicate high functional reliability, accurate keyword matching, effective link categorisation, and strong user acceptance, even among first-time web crawler users. From a societal perspective, the system benefits students, educators, and small organisations by reducing time spent on manual information searching, mitigating information overload, and supporting informed decision-making. By enabling faster access to relevant online resources, the proposed web crawler contributes to improved digital literacy, academic productivity, and more efficient use of web information in educational and professional contexts.

Keywords: Software agent, Web crawler, Automated information retrieval, Keyword-based classification, Information overload reduction

1. INTRODUCTION

The rapid expansion of web-based information has fundamentally transformed how knowledge is accessed, produced, and consumed. While search engines provide broad access to online resources, users—particularly students and novice researchers—often struggle to efficiently locate relevant, accurate, and timely information from an overwhelming volume of web content. This phenomenon, commonly referred to as *information overload*, increases cognitive effort, prolongs search time, and can negatively affect decision-making and learning outcomes (Dhanapal, 2008; Shaw et al., 2002).

Information Retrieval (IR) refers to the process of obtaining relevant information resources from large, often unstructured, data collections such as the World Wide Web. Traditional IR approaches largely depend on manual querying through search engines, where users must iteratively refine keywords, evaluate results, and filter irrelevant content. Although effective for experienced users, this process is inefficient and error-prone for non-expert users who may lack advanced search skills or domain knowledge (Hazem et al., 2015; Tu & Hsiang, 2000).

To address these limitations, *software agents* have been widely explored as an intelligent mechanism for automating information retrieval tasks. A software agent is an autonomous computational entity capable of perceiving its environment, making decisions, and performing actions to achieve predefined goals on behalf of users. In the context of IR, agent-based systems can autonomously search, extract, filter, classify, and manage information with minimal user intervention (Cabri et al., 2000; Mendes et al., 2011). Web crawlers, as a specific form of IR agent, systematically traverse web pages to collect and organise information according to defined rules or criteria.

Despite extensive research on agent-based IR architectures, many existing systems focus on complex models, large-scale infrastructures, or domain-specific applications, often requiring significant technical expertise to deploy and use (Mitra & Srivastava, 2020; Alves et al., 2018). This creates a gap between advanced IR technologies and the practical needs of everyday users, such as students and small organisations, who require lightweight, accessible, and task-oriented retrieval tools.

Motivated by this gap, this study aims to develop an Information Retrieval software agent that assists users in automated web information searching. Specifically, the objective is to design and implement a web crawler–based software agent that can automatically retrieve, categorise, and store relevant web links based on user-defined URLs and keywords, while reducing user effort and mitigating information overload. By integrating rule-based keyword matching, structured result presentation, and persistent database storage within a user-friendly desktop application, the proposed system seeks to provide a practical and usable IR solution aligned with real-world user needs.

2. LITERATURE REVIEW

Early research on agent-based information retrieval established foundational architectures and conceptual models for intelligent searching in distributed environments. Cabri et al. (2000) examined the role of agent mobility and coordination in IR systems, highlighting how autonomous agents can improve efficiency by distributing retrieval tasks across heterogeneous networks. Similarly, Tu and Hsiang (2000) proposed an architecture that integrates category knowledge into IR agents, enabling more structured and context-aware retrieval processes.

Shaw et al. (2002) extended this work by introducing a comprehensive agent-based architecture for intelligent information retrieval in distributed and heterogeneous environments. Their study demonstrated that agent cooperation and modular design could significantly enhance retrieval accuracy and system scalability. In parallel, Siraj and Ishak (2001) presented general-purpose search and classification algorithms for distributed information access, emphasizing automated classification mechanisms as a means to improve retrieval effectiveness in academic environments.

Subsequent studies focused on enhancing agent intelligence and autonomy. Dhanapal (2008) proposed an intelligent information retrieval agent capable of improving relevance through adaptive decision-making, demonstrating that autonomous agents can reduce manual search effort. Hazem et al. (2015) further explored intelligent mobile agents

for automated IR, showing that agent mobility enables faster information acquisition and reduced network load compared to centralized retrieval approaches.

Agent-based IR systems have also been applied in specialised domains. Mendes et al. (2011) introduced SASAgent, an agent-based architecture for searching and composing scientific models, illustrating the suitability of agents for complex knowledge-intensive tasks. Alves et al. (2018) developed an IR tool for biomedical patents, highlighting the importance of domain adaptation and structured result presentation for effective retrieval in specialised repositories. These studies demonstrate the versatility of agent-based IR but often target expert users or narrow application contexts.

More recent research has explored scalable and real-time agent-based IR solutions. Al-Akashi and Inkpen (2021) proposed a new intelligent data retrieval paradigm, focusing on improving retrieval efficiency through structured processing. This work was later extended by Al-Akashi and Inkpen (2022), who introduced a scalable real-time agent-based IR engine designed to handle high-volume data streams. While effective, such systems typically require substantial computational resources and advanced configuration.

Mitra and Srivastava (2020) examined agent-based web searching mechanisms and reaffirmed the role of autonomy and adaptability in improving retrieval quality. At the same time, Jamaluddin and Ishak (2011) demonstrated how virtual repository systems supported by automated retrieval mechanisms can enhance information sharing within academic departments, reinforcing the practical value of structured IR systems in educational contexts.

Most recently, Poniszewska-Maranda and Kopa (2025) explored autonomous agents for information retrieval using large language models (LLMs), signalling a shift toward AI-driven IR solutions. While promising, LLM-based systems introduce new challenges related to transparency, computational cost, and accessibility, particularly for small-scale or educational applications.

Overall, existing literature confirms the effectiveness of agent-based approaches for automated information retrieval. However, many solutions prioritise architectural sophistication, scalability, or AI-driven intelligence over usability and accessibility. There remains a need for lightweight, rule-based IR software agents that balance automation, accuracy, and ease of use for non-expert users. The present study addresses this gap by proposing a practical web crawler–based IR agent focused on usability, keyword-driven retrieval, and structured information management.

3. METHODOLOGY

This study adopts the Agile methodology due to its iterative and incremental nature, which enables continuous refinement of system functionality based on user feedback and testing outcomes. This approach is particularly suitable for developing an information retrieval system intended for students and non-expert users, as it allows usability and system effectiveness to be evaluated and enhanced throughout the development process rather than only at the final stage.

The development process began with the planning stage, where the core problem of inefficient manual web searching was analysed and system requirements were identified. Functional requirements included automated web crawling, keyword-based filtering, link categorisation, and persistent storage of retrieved information. Non-functional requirements such as ease of use, responsiveness, and minimal user intervention were also prioritised to ensure the system could effectively assist users in information searching tasks.

Following this, the system design was carried out to define the overall architecture and workflow of the software agent. The design focused on integrating the web crawler,

rule-based keyword matching mechanism, database storage, and user interface into a cohesive system. Emphasis was placed on modularity and clarity to ensure that retrieved information could be presented in a structured and easily interpretable manner for users.

The development phase involved implementing the designed components as a desktop application using Java, JavaFX, and Selenium. The web crawler was developed to autonomously traverse web pages starting from user-defined URLs, while the keyword matching mechanism filtered and categorised retrieved links according to relevance. Retrieved data were stored in a database to enable result management and reuse. Development was conducted iteratively, allowing enhancements and refinements to be introduced across multiple cycles.

Testing was performed continuously throughout development and formally evaluated through functional testing and User Acceptance Testing. Functional testing verified the correctness of crawling operations, keyword matching accuracy, link categorisation, and data storage. User Acceptance Testing involved 30 undergraduate students and focused on usability, system reliability, and user satisfaction. Feedback obtained from users was incorporated into subsequent iterations to improve system performance and user experience.

After successful testing, the system was deployed in a controlled academic environment for practical use. Deployment ensured that the application could operate reliably on standard user machines without complex configuration. The final review stage involved evaluating the system against the initial objectives, particularly its ability to assist users in automated information searching, reduce manual effort, and mitigate information overload. Insights gained from this evaluation were used to assess system effectiveness and identify potential areas for future enhancement.

4. FINDINGS

The evaluation results indicate that *InfoScout* effectively fulfils its objective of assisting users in automated web information retrieval. User Acceptance Testing demonstrates strong overall performance across key functional and usability dimensions. The highest mean score was recorded for input validation (4.70/5), suggesting that users found the system reliable in handling user-defined URLs and keywords, with minimal errors during interaction. This reflects the robustness of the system's front-end controls and validation mechanisms.

The display accuracy score (4.43/5) indicates that retrieved and categorised information was presented in a clear, structured, and understandable format. Users reported that the organisation of search results improved their ability to quickly identify relevant web resources, thereby reducing cognitive effort during information exploration. The crawling effectiveness score (4.13/5) further confirms that the web crawler was able to retrieve relevant links based on keyword matching with satisfactory accuracy, even for first-time users with limited technical background.

In terms of system dependability, reliability and stability achieved a score of 4.00/5, demonstrating consistent system behaviour during crawling and data storage operations. The overall satisfaction score (4.00/5) reflects positive user perception of the system's usefulness and ease of use. Collectively, these findings indicate that InfoScout successfully reduces the time and effort required for manual web searching, enhances efficiency in research-related tasks, and provides a practical automated information retrieval solution for students and novice users.

5. CONCLUSION

This study proposes InfoScout, a software agent for automated web information retrieval. By integrating web crawling, rule-based keyword matching, structured result presentation, and persistent storage within a user-friendly desktop application, the system addresses common challenges associated with manual web searching, including information overload and inefficient retrieval processes.

The findings demonstrate that InfoScout achieves its primary objective of assisting users in information searching by significantly reducing user effort and improving retrieval efficiency. High evaluation scores across usability, accuracy, and system reliability confirm strong user acceptance, particularly among non-expert users. From a practical perspective, the system contributes to improved academic productivity by enabling faster access to relevant online resources and supporting more informed decision-making.

REFERENCES

Al-Akashi, F., & Inkpen, D. (2021). A new approach of intelligent data retrieval paradigm. *Artificial Intelligence Advances, 3*(2), 1–12.

Al-Akashi, F., & Inkpen, D. (2022). A scalable real-time agent-based information retrieval engine. *International Journal of Software Innovation, 10*(1). https://doi.org/10.4018/IJSI.292022

Alves, T., Rodrigues, R., Costa, H., & Rocha, M. (2018). Development of an information retrieval tool for biomedical patents. *Computer Methods and Programs in Biomedicine, 159*, 125–134. https://doi.org/10.1016/j.cmpb.2018.03.012

Cabri, G., Leonardi, L., & Zambonelli, F. (2000). Agents for information retrieval: Issues of mobility and coordination. *Journal of Systems Architecture, 46*(15), 1419–1433. https://doi.org/10.1016/S1383-7621(00)00033-3

Dhanapal, R. (2008). An intelligent information retrieval agent. *Knowledge-Based Systems, 21*(6), 466–470. https://doi.org/10.1016/j.knosys.2008.03.002

Hazem, M., Alaa, R., Ahmed, A., Sameh, G., & Mitkas, N. (2015). A new automated information retrieval system using intelligent mobile agents. In *Recent advances in artificial intelligence, knowledge engineering and databases* (pp. 339–351).

Jamaluddin, Z., & Ishak, W. H. W. (2011). A virtual repository approach to departmental information sharing. *American Journal of Economics and Business Administration, 3*(1), 18–23. https://doi.org/10.3844/ajebasp.2011.18.23

Mendes, L. F., Silva, L., Matos, E., Braga, R., & Campos, F. (2011). SASAgent: An agent-based architecture for search, retrieval and composition of scientific models. *Computers in Biology and Medicine, 41*(7), 449–462. https://doi.org/10.1016/j.compbiomed.2011.04.007

Mitra, U., & Srivastava, G. (2020). A study on agent-based web searching and information retrieval. In S. Choudhury et al. (Eds.), *Intelligent communication, control and devices* (Advances in Intelligent Systems and Computing, Vol. 989, pp. 575–584). Springer. https://doi.org/10.1007/978-981-13-8618-3_59

Poniszewska-Maranda, A., & Kopa, M. (2025). Autonomous agents in software development for information retrieval using LLM models. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering* (pp. 1240–1241). https://doi.org/10.1145/3696630.3731432

Shaw, N. G., Mian, A., & Yadav, S. B. (2002). A comprehensive agent-based architecture for intelligent information retrieval in a distributed heterogeneous environment. *Decision Support Systems, 32*(4), 401–415. https://doi.org/10.1016/S0167-9236(01)00128-2

Siraj, F., & Ishak, W. H. W. (2001). General purpose search and classification algorithms for distributed information access. In *Prosiding Seminar Penyelidikan UUM ke-6* (pp. 258–271). Universiti Utara Malaysia.

Tu, H.-C., & Hsiang, J. (2000). An architecture and category knowledge for intelligent information retrieval agents. *Decision Support Systems, 28*(3), 255–268. https://doi.org/10.1016/S0167-9236(99)00089-5