Imperial College London

Department of Earth Science and Engineering

MSc in Applied Computational Science and Engineering

Independent Research Project
Project Plan

Current Content Discovery for Module Teaching

by

Guanyuming He

Email: guanyuming.he24@imperial.ac.uk
GitHub username: esemsc-gh124
Repository: https://github.com/ese-ada-lovelace-2024/irp-gh124

Supervisors:

Sean O’Grady
Rhodri Nelson

10th June 2025

Abstract

TBD. **Keywords:** content discovery, information retrieval, LLM, search engines, case method, Business school,

1 Introduction

1.1 Problem background

Since the emergence of the first Business schools in the late 19th century [42, 35], several distinct pedagogical teaching strategies have been applied. First institutionalized at Harvard Business School [17, 5] in the early 20th century, a method about teaching students with real world business cases (will be called *case method* in the rest of the document) has been found more effective and engaging [44, 3, 26] than traditional big lecture methods, and gained wide adoption across the world [43, 11].

However, the case method is constrained by the availability and collection of timely and relevant case material [38, 30]. In particular, Christensen has identified in his classical article that instructors would have to conduct "extensive preparation" [8] for case method, and some others emphasise the importance of up-to-date material [9, 27].

1.2 Past advancements in information retrieval

During the past two centuries, a number of key developments have profoundly expanded an individual's capacity to retrieve information about the world. In the early 19th century, the transmission of information was still traditional — carried by person on paper or simply remembered. The invention of telegraph in the 1840s by Morse, Cornell, and Henry [36, 25], was perhaps the first big advancement in history. This technological innovation was considerably improved by the invention of the telephone by Bell in 1876 [46, 13], enabling communication directly by human voice, instead of encoded Morse code.

Wired communication is critically limited by geographical features on where the wires were laid. Around the late 19th century, Marconi's experiments with wireless telegraphy [12] and the first successful transatlantic signal in 1901 [2] introduced electromagnetic wave-based wireless communication, eventually accumulating into the world's first voice broadcast by radio in 1906 [41].

The next many decades have seen people improving on the serious limitations of wireless communication: signal strength, interferences and attenuation, carrying capacity, and even deliberate sabotage during war times [16, 47, 1]. Theoretically, Hartley observed a logritham pattern of information capacity [18] and then Shannon expanded on it to first define *bits* and *entropy*, giving a formal mathematical theory of information [39] in 1948. Meanwhile, engineers were experimenting with alternative modulation techniques, notably frequency modulation (FM), and the concepts were formalized in the 1930s [19].

As the theoretical understanding of information progressed, people began to have the idea of *searching* for information based on content and by relevance, instead of by unique identifier [33]. The term *information retrieval* was coined by Mooers in 1950 [28]. Since then, information retrieval systems have quickly evolved, and Griffiths and King identified four phases of its evolvement: "(1) manual and mechanical devices; (2) offline computing; (3) online computing, vendor access; (4) distributed, networked, and mass computing." [15], with the last three substantially contributed to by the invention of the Internet [24], and consequently the emergence of search engines in the 1990s [37, 29]. For the next two decades, search engines greatly expanded in speed and coverage and has been significantly impacting the society's information for at least a decade [7, 21].

Recently, LLMs have had a huge impact on various areas [10]. One strong appeal of LLMs is that they could easily work with natural language input & output [49, 23]. However, because of the inherent limit of neural networks, some argue that they could not formally reason about what they output [34, 22, 32]. Indeed, hallucination [20, 31] and other forms of distortion of facts, is a big problem of LLMs. On the other hand, although search engines rely on determinstic algorithms that give precise reference to searched result, they could not compare with LLMs' ability to process search prompts and summarize results.

Therefore, it is a current research direction to integrate LLMs with search engines [48, 40, 45]. Specifically, Xiong et al. proposes to categorize them into "LLM4Search" and "Search4LLM", where "A4B" means using A to improve B [48]. Here is where my thesis will build upon.

1.3 Goals

In this project, I will attempt to improve the current state of information retrieval (mostly limited to business teach related information) further by designing and developing a software system. More precisely, these are the goals:

1. By combining LLMs and search engines, compensate each other's weakeness in information retrieval.
    (a) The system shall improve search engines on prompt engineering to the extent that a user would only need to give a rough and vague natural language description of the target information to the system.
    (b) The system shall improve LLMs on result credibility and interpretibility. If pure LLMs lack the two properties, then combination with deterministic and well-designed ranking and searching algorithms in search engines will compensate for that.
2. Enhance the automation of the current general public information retrieval tools, e.g., Google search, ChatGPT, to the extent that a user would only need to occasionally configure the system to run them.
3. Specially tailor the system to business teaching related information, aiming to gather information from authentic and reliable sources.
4. The system shall support up-to-date information retrieval. More precisely, the system shall support the user to specify a time range of information, and the user is allowed to set the end of the range to the current time.
5. The system could support user feedback and learning from it to provide information more close to one's need.

2 Methods

2.1 Architecture

Based on the goals, the system is designed to be a chain of tools invoking each other.

What directs the system is the user's configuration, which is expected to include

1. A time range of information to retrieve.

2. A description of the types of information to retrieve.
3. How the retrieved information is sent to the user.
4. A frequency of running the tool.

Based on the configured running frequency, the tool will

1. Use LLMs to translate the configured description and generate a list of search engine prompts.
2. The prompts are fed to search engines, with the configured time constraint.
3. The search results are gathered and processed. How the information is processed and filtered may rely on other AI systems like a rule-based expert system.
4. The processed results may be summarized by LLMs.
5. The results are sent to the users via the configured ways.
6. Optionally, the user gives feedback to the results. The tool can thus improve on the search engine prompts, the information processing, and LLM parameters, based on the feedback. The improvement algorithm may use something like recommendation algorithms. Then, the tool can iterate on the user's configuration automatically.

Figure 1 demonstrates my architecture design and the workflow of the system.

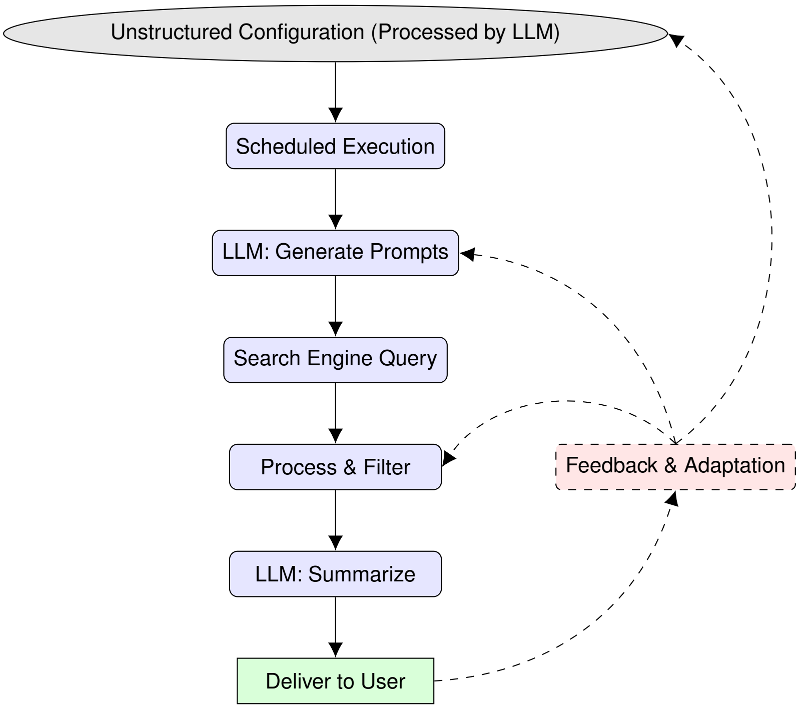

Figure 1: Architecture and operational flow of the software system

What directs the system is the user's configuration, which is expected to include a time range of information to retrieve, a natural language description of the types of information to retrieve, how the retrieved information is sent to the user, and a frequency of running the tool.

2.2 Implementation plan

In this section, I give the plan of my implementation. It is only meaningful for the project plan, as it will be replaced by the specifics in the final report.

2.2.1 Configuration format

LLMs have demonstrated unprecedented ability to process unstructured data [6, 4]. Thus, it is desirable for the system to accept unstructured configuration from the user. The user may give a natural language description of the wanted information, give some examples, e.g., URLs, or simply feed the LLM the teaching slides and videos from previous classes.

Nevertheless, there are a few configurations that must be formatted.

1. A range of time for the information retrieved.
2. (Optional) A list of sources (e.g. websites) to always include or exclude.
3. The destination of the collected information (e.g. email addresses of teaching staff).

Of course, LLMs can still try to infer them from the given unstructured data, but they must be presented to the user for her to decide and correct, because these configurations are strict and a little error will have great effects.

Another important output from unstructured configuration is the search prompts or insturctions generated by the LLMs. As this output direct controls the information retrieved. LLMs for this task will be specifically tuned to ensure

- Diversity of results.
- A specific distribution of results. E.g., a more authentic website will contribute relatively more information.

2.2.2 Scheduled Execution

Based on the configured running frequency, the system will run every once a while automatically. The scheduling component must handle different frequencies (daily, weekly, monthly) and manage execution timing to avoid overwhelming external APIs or generating excessive costs. This involves implementing a robust scheduling system that can handle failures, retries, and load balancing across multiple users.

2.2.3 Multi-Source Search Integration

One problem of search engine APIs is that they have ungenerous search limits [14], allowing about 1000 queries per month, which is not enough for frequent testing and prototyping.

As a result, I plan to use some other self-developed system in parallel with search engines, which do only a small amount of tasks that truly require whole Internet searching. The majority of the searching will be handled by scraping a limited sets of websites, most likely.

2.2.4 Content Processing and Filtering

Search results will be gathered and processed through a multi-stage filtering system. This involves content deduplication, relevance scoring, source credibility assessment, and quality filtering. The processing pipeline will need to handle various content formats including articles, reports, videos, and social media posts.

The filtering system will combine rule-based approaches (such as keyword matching and source whitelisting) with machine learning techniques for relevance scoring and content classification. Natural language processing techniques will be employed to extract key information and assess content quality. The system will also need to handle different languages and cultural contexts in business information.

2.2.5 LLM-Based Summarization

Processed results will be summarized using LLMs with business-focused prompting strategies. The summarization component must balance comprehensiveness with conciseness, ensuring that key business insights are preserved while making information digestible for educators. This involves developing prompting strategies that can identify the most relevant information for case method teaching and present it in a structured format.

The summarization process may involve extractive techniques (selecting key sentences from original content) and abstractive techniques (generating new summaries). Different summarization styles may be needed for different types of content and user preferences.

2.2.6 Delivery and User Interface

Results shall be sent to users via configured methods, which may include email summaries, web dashboards, or integration with existing educational platforms. The delivery component shall handle different user preferences and ensure information is presented in a format suitable for educational use.

2.2.7 Feedback and Adaptation System

The system shall collect user feedback on result quality and relevance, using this information to improve future searches. This involves designing feedback mechanisms that capture both explicit user ratings and implicit behavioral signals. Possible algorithms include recommendation algorithms and other machine learning models.

2.3 Technical Considerations

The system will be a high-level program, which calls for simple programming languages like Python. On the other hand, the system may have specific performance requirements, especially for LLMs and other machine learning algorithms. The likely result is then a combination of different programming languages and technical frameworks.

2.4 Evaluation and Validation

It may be desirable to develop an external evaluation framework for the system, if I have enough time in the end.

References

[1] E. F. W. Alexanderson. Transatlantic radio communication. *Proceedings of the American Institute of Electrical Engineers*, 38(10):1077–1093, 1919.

[2] W. J. G. Beynon. Marconi, radio waves, and the ionosphere. *Radio Science*, 10(7):657–664, 1975.

[3] Kevin M. Bonney. Case study teaching method improves student performance and perceptions of learning gains. *Journal of Microbiology &amp; Biology Education*, 16(1):21–28, 2015.

[4] William Brach, Kristián Koštál, and Michal Ries. The effectiveness of large language models in transforming unstructured text to standardized formats, 2025.

[5] Todd Bridgman, Stephen Cummings, and Colm McLaughlin. The case method as invented tradition: revisiting harvard's history to reorient management education. In *Academy of Management Proceedings*, volume 2015, page 11637. Academy of Management Briarcliff Manor, NY 10510, 2015.

[6] Andry Castro, João Pinto, Luís Reino, Pavel Pipek, and César Capinha. Large language models overcome the challenges of unstructured text data in ecology. *Ecological Informatics*, 82:102742, 2024.

[7] Junghoo Cho and Sourashis Roy. Impact of search engines on page popularity. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, page 20–29, New York, NY, USA, 2004. Association for Computing Machinery.

[8] C Roland Christensen. Teaching and the case method harvard business school. `https://www.hbs.edu/teaching/case-method/Pages/default.aspx`, 1987.

[9] Tom Clark. Case method in the digital age: how might new technologies shape experiential learning and real-life story telling? LSE Impact of Social Sciences, 2016.

[10] Qinxu Ding, Ding Ding, Yue Wang, Chong Guan, and Bosheng Ding. Unraveling the landscape of large language models: a systematic review and future perspectives. *Journal of Electronic Business & Digital Economics*, 3(1):3–19, 2023.

[11] Rebekka Eckhaus. Supporting the adoption of business case studies in esp instruction through technology. *Asian ESP Journal*, 14:280–281, 2018.

[12] Gabriele Falciasecca. Marconi's early experiments in wireless telegraphy, 1895. *IEEE Antennas and Propagation Magazine*, 52(6):220–221, 2010.

[13] J.E. Flood. Alexander graham bell and the invention of the telephone. *Electronics and Power*, 22:159–162, 1976.

[14] Google Developers. Usage limits, 2025. Accessed: 2025-06-09.

[15] J.-M. Griffiths and D.W. King. Us information retrieval system evolution and evaluation (1945-1975). *IEEE Annals of the History of Computing*, 24(3):35–55, 2002.

[16] Brian N. Hall. The british army and wireless communication, 1896–1918. *War in History*, 19(3):290–321, 2012.

[17] John S Hammond. *Learning by the case method*. Harvard Business School Boston, MA, 1980.

[18] R. V. L. Hartley. Transmission of information. *Bell System Technical Journal*, 7(3):535–563, 1928.

[19] Raymond A. Heising. Modulation methods. *Proceedings of the IRE*, 50(5):896–901, 1962.

[20] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025.

[21] Olagoke Olawale Israel and Olatunji Olusoji Samson. The impact of search engines in the world today. *International Journal of Management, IT and Engineering*, 8(3):10–22, 2018.

[22] Andrei Kucharavy. Fundamental limitations of generative llms. In *Large Language Models in Cybersecurity: Threats, Exposure and Mitigation*, pages 55–64. Springer Nature Switzerland Cham, 2024.

[23] Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260, 2024.

[24] Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. A brief history of the internet. *SIGCOMM Comput. Commun. Rev.*, 39(5):22–31, October 2009.

[25] Kenneth B Lifshitz. *Makers of the Telegraph: Samuel Morse, Ezra Cornell and Joseph Henry*. McFarland, 2017.

[26] Alberto Lusoli. Teaching business as business: The role of the case method in the constitution of management as a science-based profession. *Journal of Management History*, 26(2):277–290, 2020.

[27] Donovan A McFarlane. Guidelines for using case studies in the teaching-learning process. *College Quarterly*, 18(1):n1, 2015.

[28] Calvin Mooers. The theory of digital handing of non-numerical information and its implications to machine economics. In *Proceedings of the meeting of the Association for Computing Machinery at Rutgers University*, 1950.

[29] Rhoda Okunev. *History of the Internet, Search Engines, and More*, pages 9–16. Apress, Berkeley, CA, 2023.

[30] Sandeep Puri. Effective learning through the case method. *Innovations in Education and Teaching International*, 59(2):161–171, 2022.

[31] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models, 2023.

[32] Walid S Saba. Stochastic llms do not understand language: towards symbolic, explainable and ontologically based llms. In *International conference on conceptual modeling*, pages 3–19. Springer, 2023.

[33] Mark Sanderson and W. Bruce Croft. The history of information retrieval research. *Proceedings of the IEEE*, 100(Special Centennial Issue):1444–1451, 2012.

[34] Erin Sanu, T Keerthi Amudaa, Prasiddha Bhat, Guduru Dinesh, Apoorva Uday Kumar Chate, and Ramakanth Kumar P. Limitations of large language models. In *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)*, pages 1–6, 2024.

[35] Steven A Sass. *The pragmatic imagination: A history of the Wharton School, 1881-1981*. University of Pennsylvania Press, 2016.

[36] Mischa Schwartz and David Hochfelder. Two controversies in the early history of the telegraph. *IEEE Communications Magazine*, 48(2):28–32, 2010.

[37] Tom Seymour, Dean Frantsvog, Satheesh Kumar, et al. History of search engines. *International Journal of Management & Information Systems (IJMIS)*, 15(4):47–58, 2011.

[38] Binod Shah. Case method of teaching in management education. *Research Gate*, 2019.

[39] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

[40] Xiang Shi, Jiawei Liu, Yinpeng Liu, Qikai Cheng, and Wei Lu. Know where to go: Make llm a relevant, responsible, and trustworthy searchers. *Decision Support Systems*, 188:114354, 2025.

[41] Elliot N. Sivowitch. A technological survey of broadcasting's "pre-history," 1876–1920. *Journal of Broadcasting*, 15(1):1–20, 1970.

[42] John-Christopher Spender. The business school in america: a century goes by. *The future of business schools: Scenarios and strategies for*, pages 9–18, 2020.

[43] Carlos JO Trejo-Pech and Susan White. The use of case studies in undergraduate business administration. *Revista de administração de empresas*, 57:342–356, 2017.

[44] Klára Vítečková, Tobias Cramer, Matthias Pilz, Janine Tögel, Sascha Albers, Steven van den Oord, and Tomasz Rachwał. Case studies in business education: an investigation of a learner-friendly approach. *Journal of International Education in Business*, 18(2):149–176, 2025.

[45] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation, 2023.

[46] Thomas A. Watson. How bell invented the telephone. *Proceedings of the American Institute of Electrical Engineers*, 34(8):1503–1513, 1915.

[47] JONATHAN REED WINKLER. Telecommunications in world war i. *Proceedings of the American Philosophical Society*, 159(2):162–168, 2015.

[48] Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing*, 17(6):4558–4577, 2024.

[49] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6), April 2024.