Aggregated Search Engine

SONG YUFENG
Assoc. Prof. Ts. Dr. Abdul Hadi Abd Rahman

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

ABSTRAK

This project uses different technical stacks to implement an aggregated search engine. The system uses Vue 3 and Ant Design Vue to implement a user-friendly interface that is more responsive and interactive.

On the back end, the system uses the Spring Boot framework for backend functionality, and the engine can get results from various data sources. The system uses MySQL as the primary database for storing and managing data.
The system also uses the Elastic Stack framework. Elasticsearch implements the search function, Logstash is used for data stream processing, and Kibana is used for data visualization to enable powerful search capabilities.

The system uses data crawling techniques to get data from other search engines and can combine many data sources for offline and real-time data collection using Jsoup and the HttpClient framework.

The system also uses many design patterns, including the Facade Pattern, Adapter Pattern, and Registry Pattern.
There are four methods to synchronize data, but it mainly uses the Logstash framework.

For project testing, JMeter is used for stress testing the search engine.
The result is a high-performance, scalable, and user-friendly aggregated search engine that meets diverse user needs.

PENGENALAN

The aggregate search platform mainly aims to solve the complex search problem in an efficient way.As the amount of data increases across websites and systems, users need a solution to quickly find the right information that they need.The project uses the latest technologies, including Elasticsearch for search functions, Spring Boot to support the backend, and Vue 3 to provide a user-friendly interface.By combining multiple data sources into a unified search interface, the platform can improve the user experience.

METODOLOGI KAJIAN

The development of the Aggregated Search Engine follows a full-stack development approach, integrating Spring Boot for backend services, Vue 3 for frontend UI, and Elasticsearch for search capabilities. The system leverages Logstash for data synchronization and Jsoup for web scraping, allowing it to aggregate search results from multiple data sources efficiently.

The project was developed iteratively, beginning with requirement analysis to identify challenges in existing search engines and define core functionalities. The system design phase focused on structuring the database, designing RESTful APIs, and implementing a user-friendly interface. The implementation phase involved backend service development, Elasticsearch integration, and frontend development using Vue 3

To ensure system reliability, functional and performance testing was conducted using JMeter, measuring query response times and scalability under different loads. Additionally, user feedback was collected to refine search algorithms and improve the user experience. The final system delivers a high-performance, scalable, and user-friendly search platform, capable of handling large query volumes while maintaining efficiency.

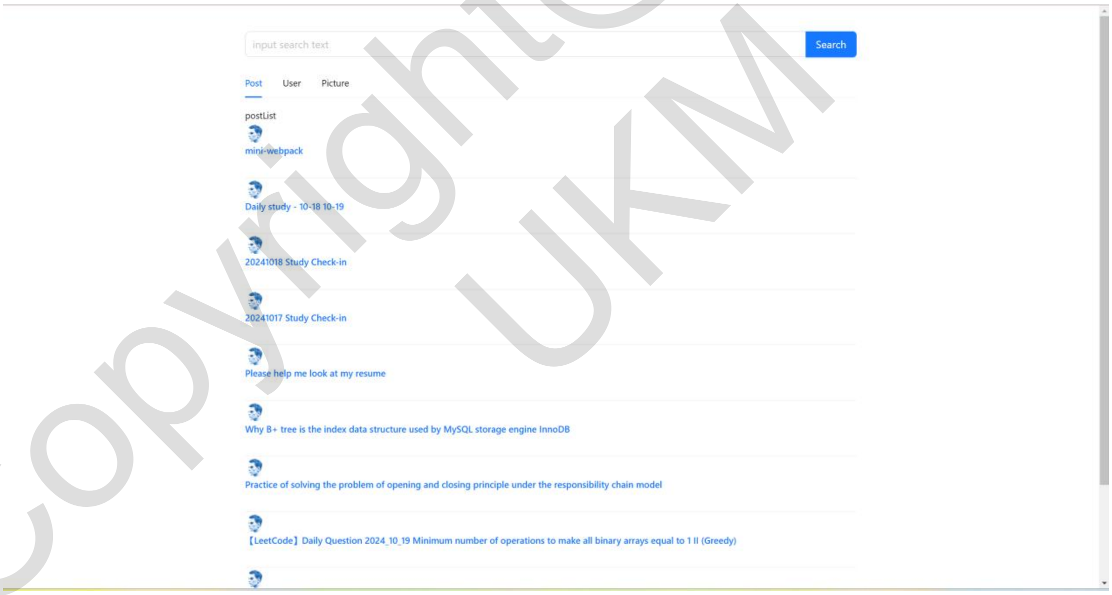
KEPUTUSAN DAN PERBINCANGAN

In summary, the development of the Aggregated Search Engine addresses the growing need for an efficient and unified search platform that can retrieve, aggregate, and optimize results from multiple data sources. The fragmentation of information across different websites and databases highlights the importance of a system that provides comprehensive, accurate, and categorized search results in a user-friendly manner. The proposed solution integrates Spring Boot for backend processing, Vue 3 for an interactive frontend, Elasticsearch for search optimization, and Logstash for real-time data synchronization. These technologies work together to ensure high-performance querying, keyword highlighting, and structured result categorization, ultimately enhancing the user experience.

The project successfully aligned with its objectives by implementing key search functionalities, system architecture design, and performance testing. Functional validation and JMeter stress testing confirmed that the search engine operates efficiently under high concurrency, ensuring fast response times and reliable performance.

While challenges such as data synchronization, scalability, and query optimization were encountered, they provided opportunities for continuous improvement. Through careful planning and development, the system was designed to be scalable, adaptable, and capable of future enhancements, such as AI-driven search ranking, real-time indexing, and improved search filtering mechanisms.

When user open the search website,the first page have 7 tags of category which include post,user,picture,video,news,IT(information technology),general.In post page,user can search the content from the post database and the search text will highlight.the user page search the content from the user database.



postList

mini-webpack

Daily study - 10-18 10-19

20241018 Study Check-in

20241017 Study Check-in

Please help me look at my resume

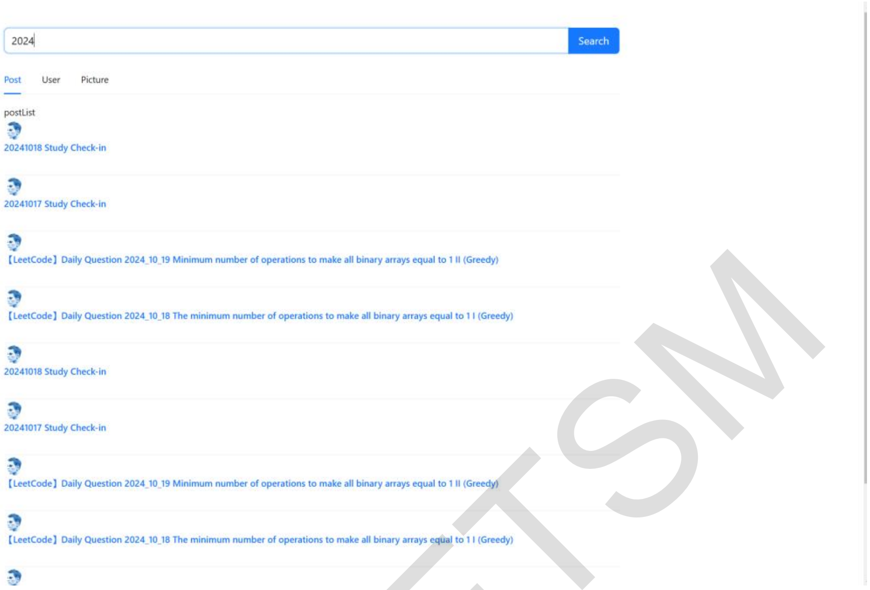
Why B+ tree is the index data structure used by MySQL storage engine InnoDB

Practice of solving the problem of opening and closing principle under the responsibility chain model

【LeetCode】 Daily Question 2024_10_19 Minimum number of operations to make all binary arrays equal to 1 II (Greedy)

Figure 1

Figure 1 shows this page features a post component and includes navigation options for users and pictures.



postList

20241018 Study Check-in

20241017 Study Check-in

【LeetCode】 Daily Question 2024_10_19 Minimum number of operations to make all binary arrays equal to 1 II (Greedy)

【LeetCode】 Daily Question 2024_10_18 The minimum number of operations to make all binary arrays equal to 1 I (Greedy)

20241018 Study Check-in

20241017 Study Check-in

【LeetCode】 Daily Question 2024_10_19 Minimum number of operations to make all binary arrays equal to 1 II (Greedy)

【LeetCode】 Daily Question 2024_10_18 The minimum number of operations to make all binary arrays equal to 1 I (Greedy)

Figure 2

Figure 2 shows search for the text '2024' and provide the results along with the data source from MySQL

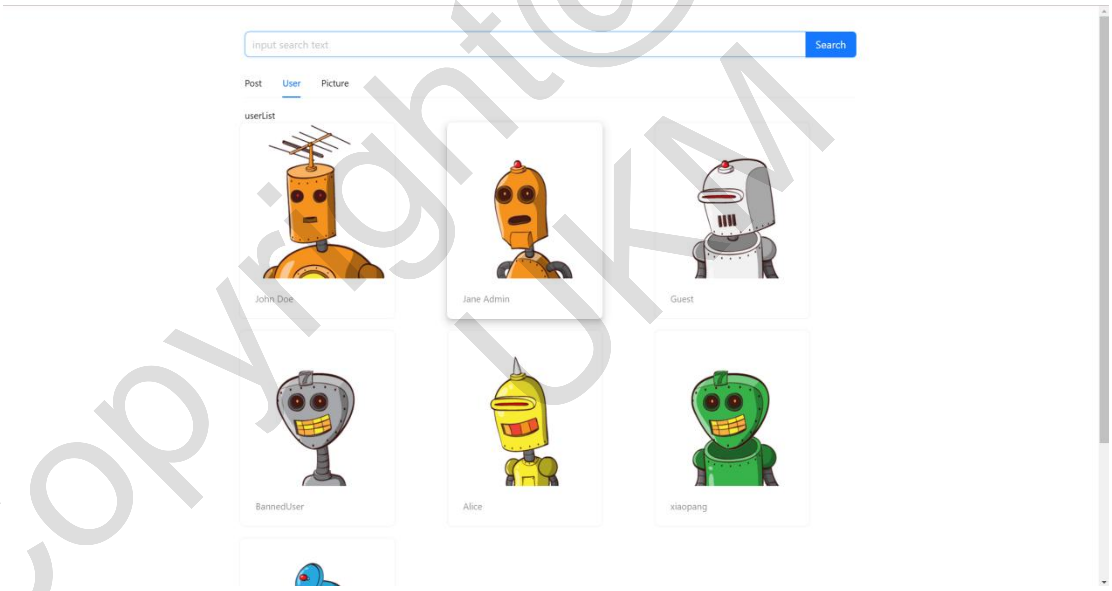


Figure 3

Figure 3 shows the user page and pictures are generated from the RobotHash website.

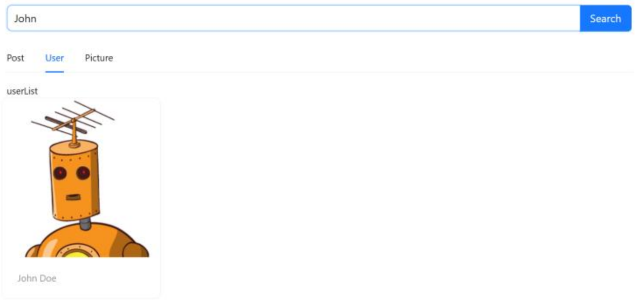


Figure 4

Figure 4 shows search the text ‘JOHN’ and also provide the results

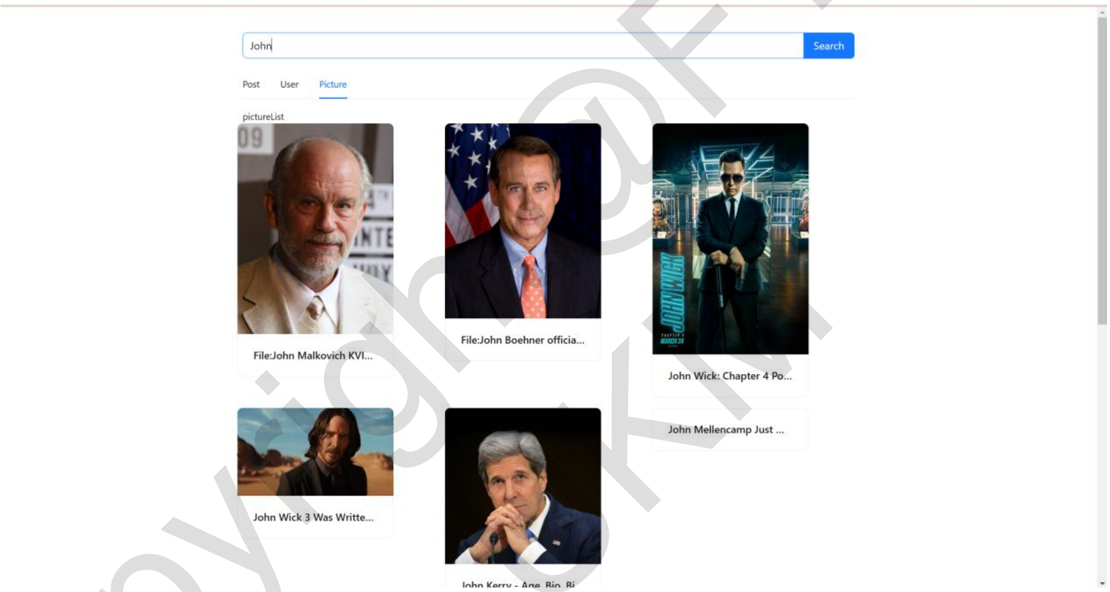

John
Search
Post
User
Picture
pictureList
File:John Malkovich KVI...
File:John Boehner officia...
John Wick: Chapter 4 Po...
John Mellencamp Just ...
John Wick 3 Was Writte...
John Kerry - Age, Bio, Bi

Figure 5

Figure 5 shows for the picture page, I use web scraping to fetch images from Bing's website.

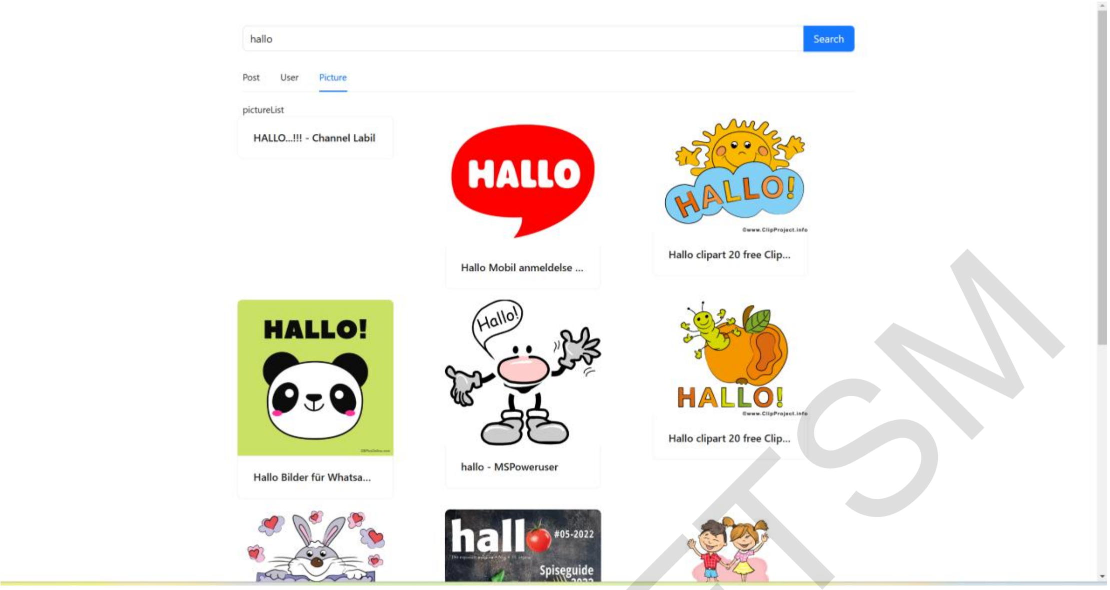

hallo
Search
Post
User
Picture
pictureList
HALLO...!!! - Channel Labil
HALLO
Hallo Mobil anmeldelse ...
HALLO!
Hallo clipart 20 free Clip...
HALLO!
Hallo Bilder für Whatsa...
Hallo!
hallo - MSPoweruser
HALLO!
Hallo clipart 20 free Clip...
hall
#05-2022
Spiseguide

Figure 6

Figure 6 shows some images fail to load due to hotlink protection

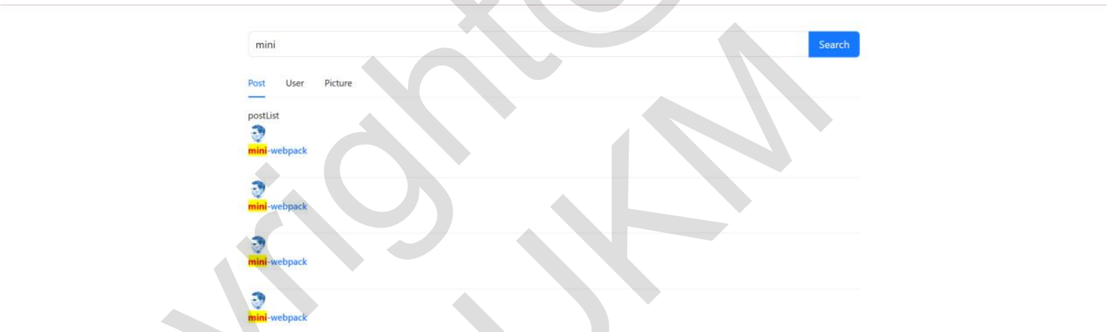


Figure 7

Figure 7 shows this function is keywords highlighting and it is combine the backend and frontend to implement

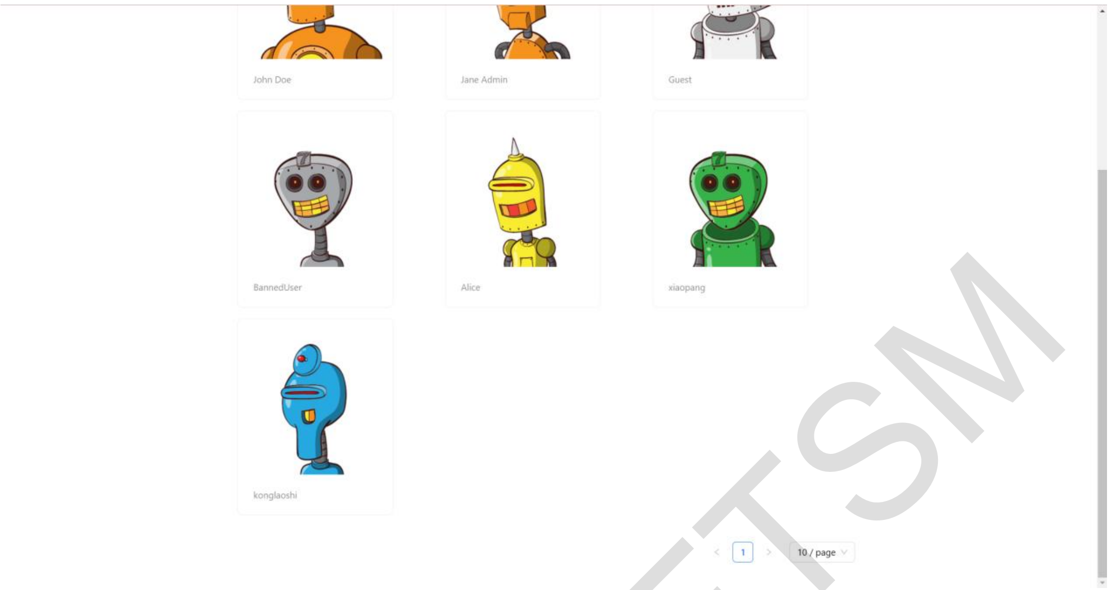


Figure 8

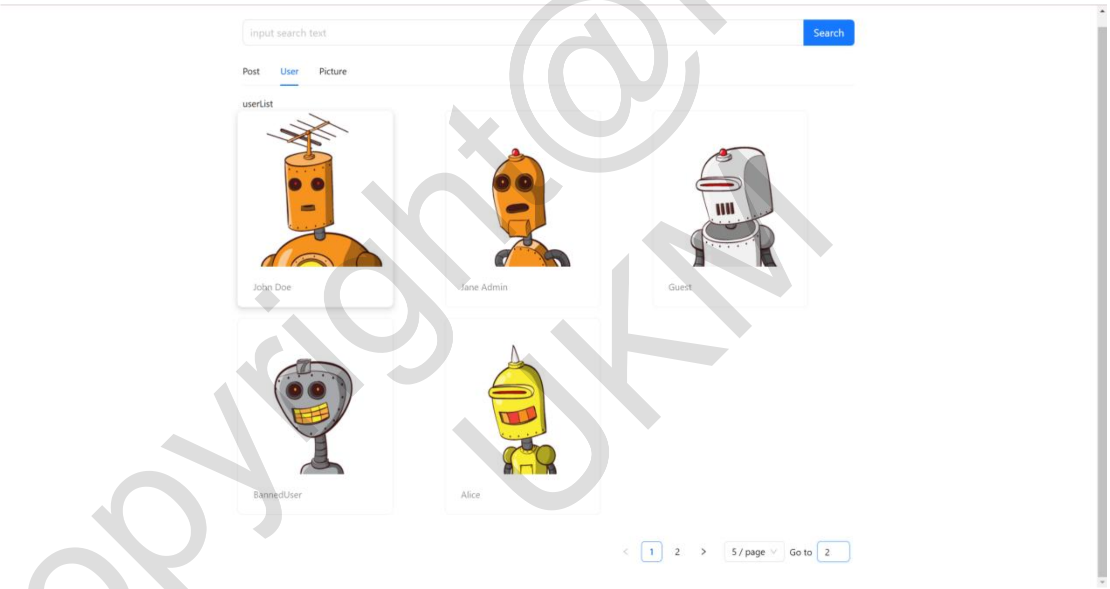


Figure 9

Figure 8 and Figure 9 shows At the bottle of the page,there is a pagination function that provides users to divide post,user,picture into page of 5,10 or 20 items per page base on user selection.users also can directly jump to a specific page by the selected number

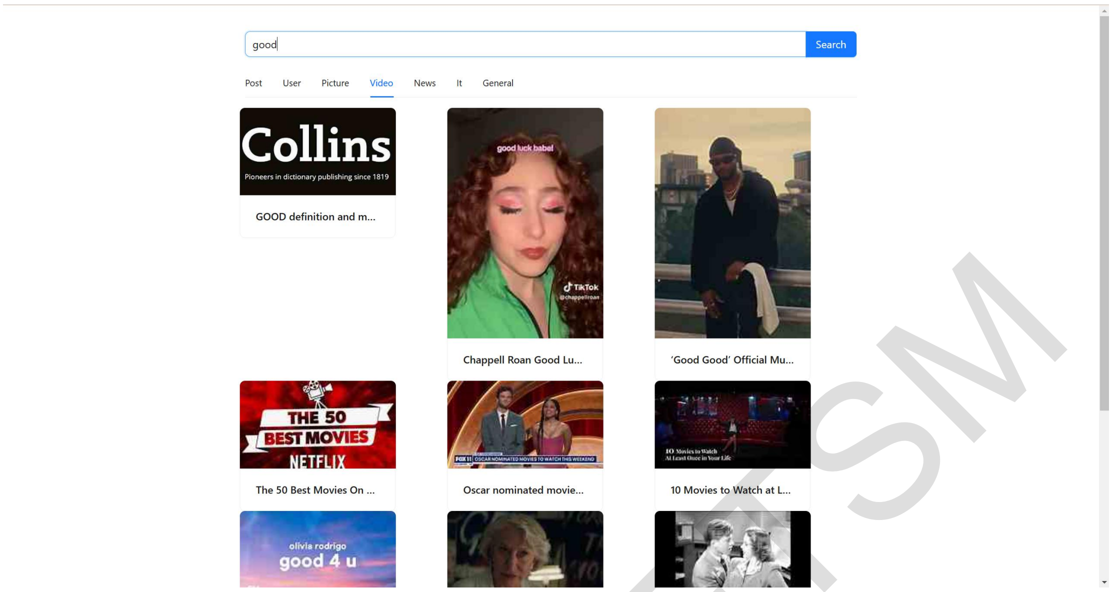


Figure 10

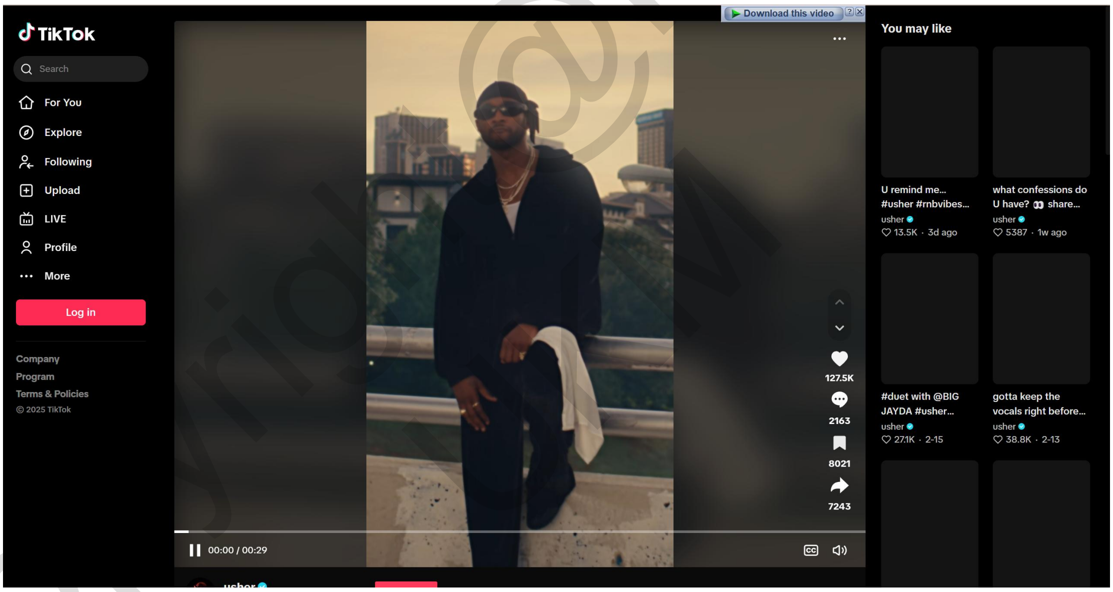
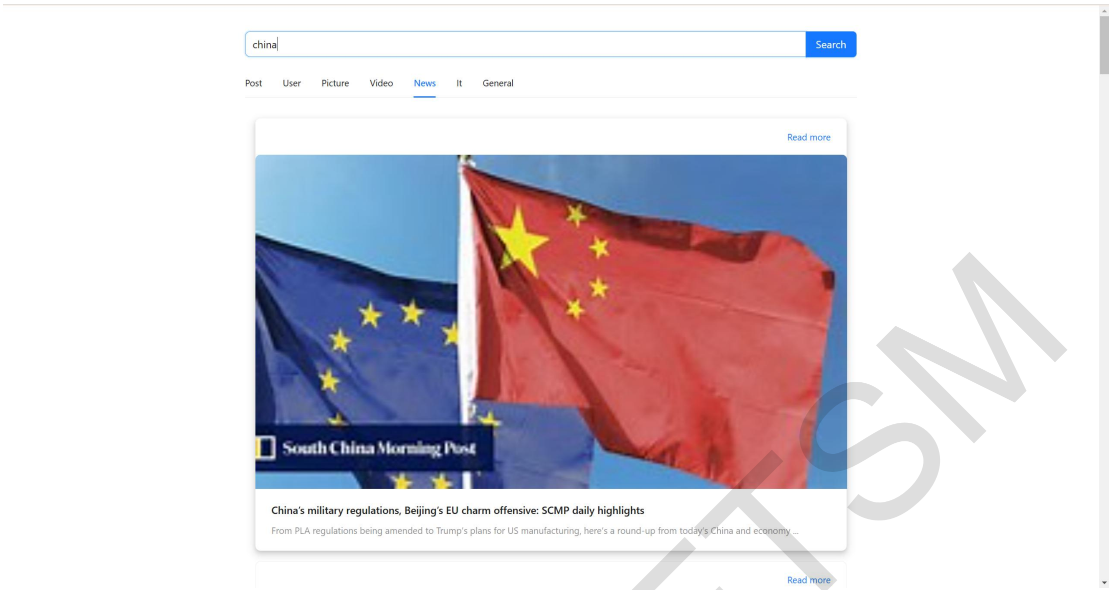


Figure 11

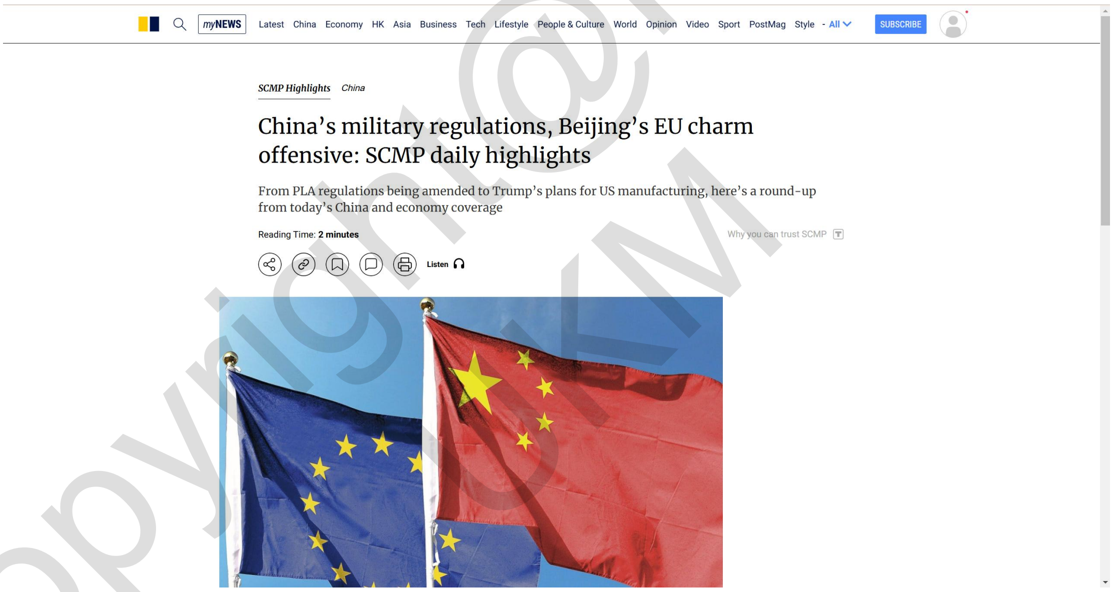
The Figure 10 and Figure 11 show the search function from the video tag.when user input good SearchText and can get the result.user can click the picture to open the video website.



China’s military regulations, Beijing’s EU charm offensive: SCMP daily highlights

From PLA regulations being amended to Trump’s plans for US manufacturing, here’s a round-up from today’s China and economy ...

Figure 12



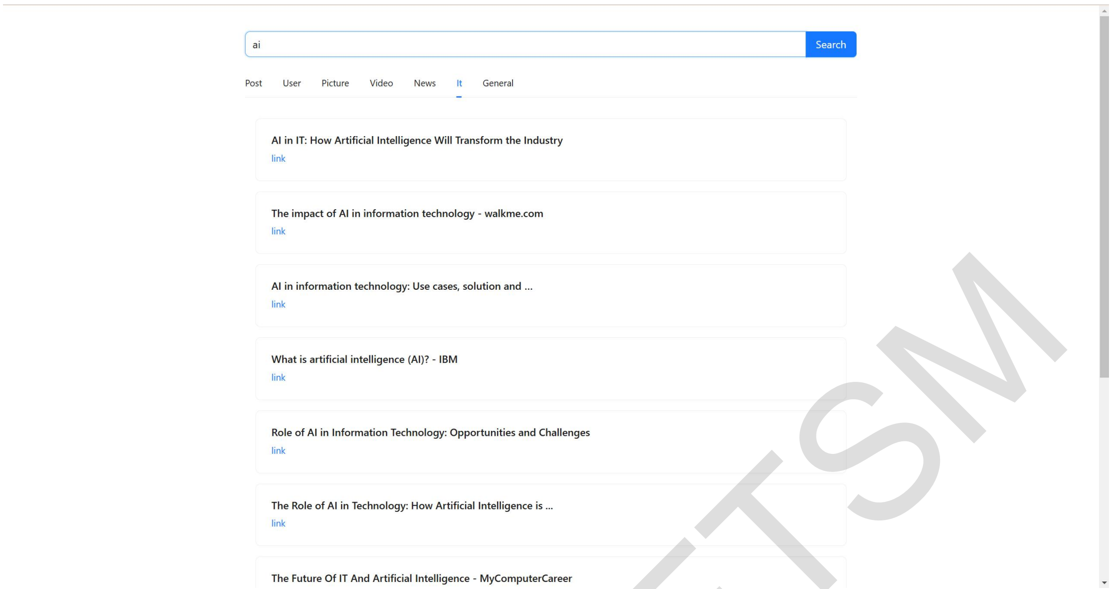
China’s military regulations, Beijing’s EU charm offensive: SCMP daily highlights

From PLA regulations being amended to Trump’s plans for US manufacturing, here’s a round-up from today’s China and economy coverage

Reading Time: 2 minutes

Figure 13

Figure 12 and Figure 13can search the content from the news tag and when click the read more,the page will redirect to the news website



AI in IT: How Artificial Intelligence Will Transform the Industry
link

The impact of AI in information technology - walkme.com
link

AI in information technology: Use cases, solution and ...
link

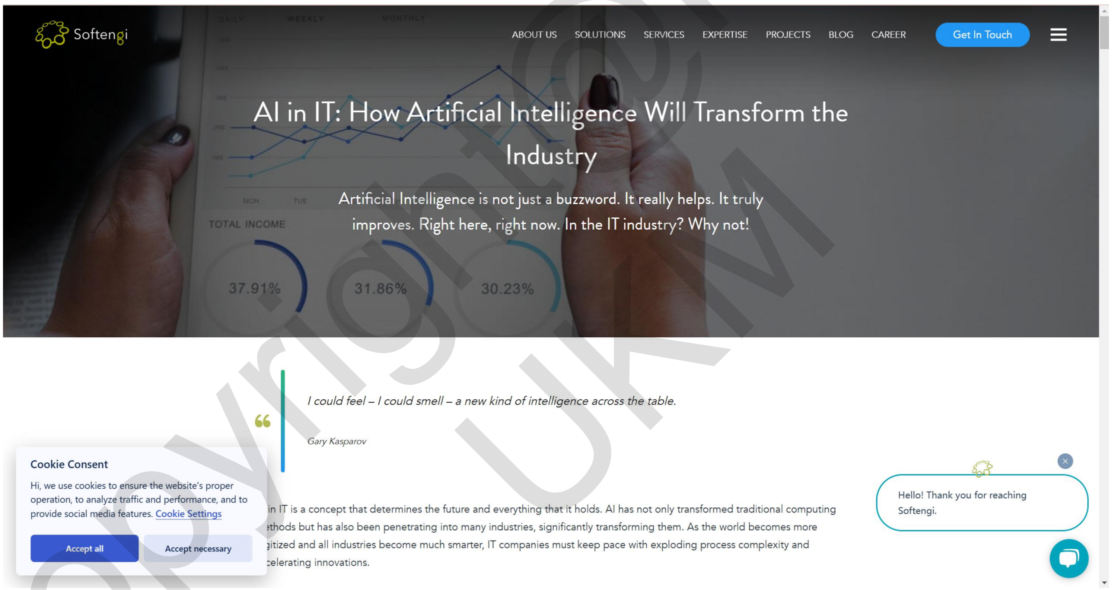
What is artificial intelligence (AI)? - IBM
link

Role of AI in Information Technology: Opportunities and Challenges
link

The Role of AI in Technology: How Artificial Intelligence is ...
link

The Future Of IT And Artificial Intelligence - MyComputerCareer

Figure 14



AI in IT: How Artificial Intelligence Will Transform the Industry

Artificial Intelligence is not just a buzzword. It really helps. It truly improves. Right here, right now. In the IT industry? Why not!

I could feel – I could smell – a new kind of intelligence across the table.

Gary Kasparov

Cookie Consent

Hi, we use cookies to ensure the website's proper operation, to analyze traffic and performance, and to provide social media features. Cookie Settings

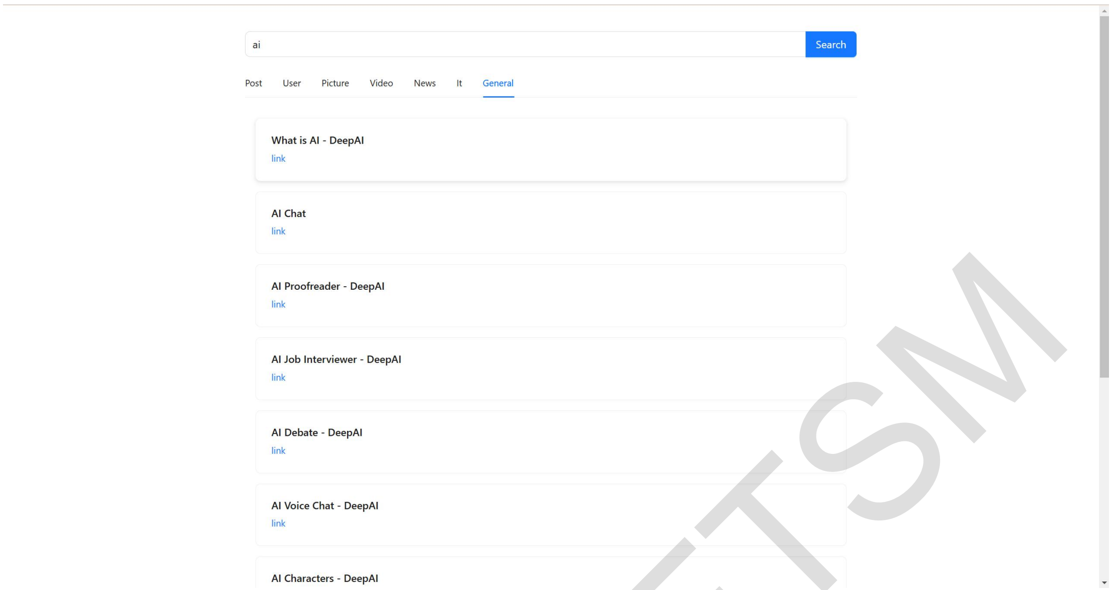
Accept all | Accept necessary

in IT is a concept that determines the future and everything that it holds. AI has not only transformed traditional computing ethods but has also been penetrating into many industries, significantly transforming them. As the world becomes more gitized and all industries become much smarter, IT companies must keep pace with exploding process complexity and celerating innovations.

Hello! Thank you for reaching Softengi.

Figure 15

Figure 14 and Figure 15 show search the content from the IT(information technology)tags,all the content about the IT and click the link also can redirect the website



What is AI - DeepAI
link

AI Chat
link

AI Proofreader - DeepAI
link

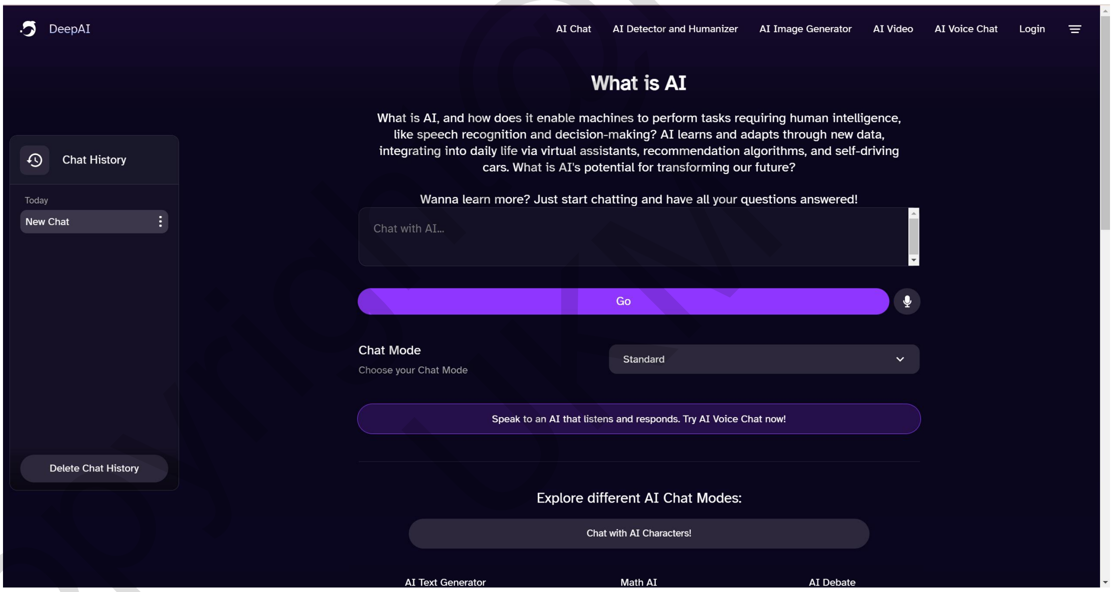
AI Job Interviewer - DeepAI
link

AI Debate - DeepAI
link

AI Voice Chat - DeepAI
link

AI Characters - DeepAI

Figure 16



What is AI

What is AI, and how does it enable machines to perform tasks requiring human intelligence, like speech recognition and decision-making? AI learns and adapts through new data, integrating into daily life via virtual assistants, recommendation algorithms, and self-driving cars. What is AI's potential for transforming our future?

Wanna learn more? Just start chatting and have all your questions answered!

Chat Mode
Choose your Chat Mode

Speak to an AI that listens and responds. Try AI Voice Chat now!

Explore different AI Chat Modes:

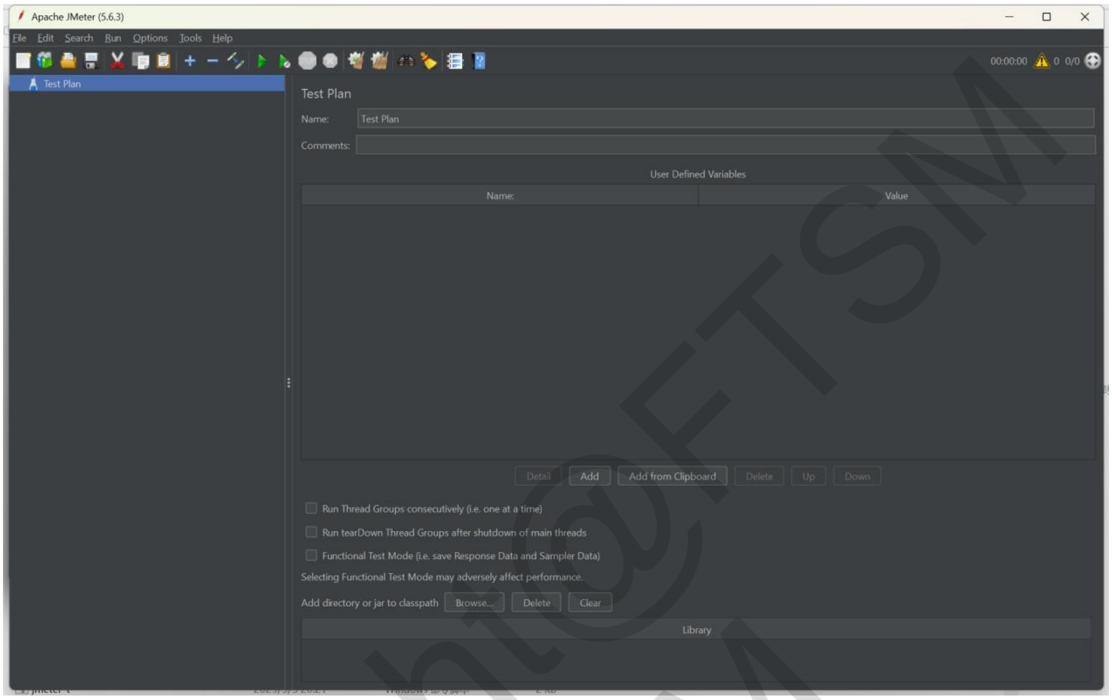
Figure 17

Figure 16 and Figure 17 show the search content from the general tags.the result will return the general post and the link when click the link also can redirect

Pengujian Kebolehgunaan

Testers accessed the application through the front-end Vue3 web interface and verified interaction with the Spring Boot back-end APIs.Testers performed search queries using valid and invalid inputs.Functional tests included UI interactions (input fields, buttons) and backend response validation.Performance tests measured response time, scalability, and resource consumption.

The summarized results of the testing phase are as follows:User interface worked correctly and the API responded within 1 second under the long query search,And this project uses the Jmeter to stress test.



Test Plan

Name: Test Plan

Comments:

User Defined Variables

Name:	Value

☐ Run Thread Groups consecutively (i.e. one at a time)

☐ Run tearDown Thread Groups after shutdown of main threads

☐ Functional Test Mode (i.e. save Response Data and Sampler Data)

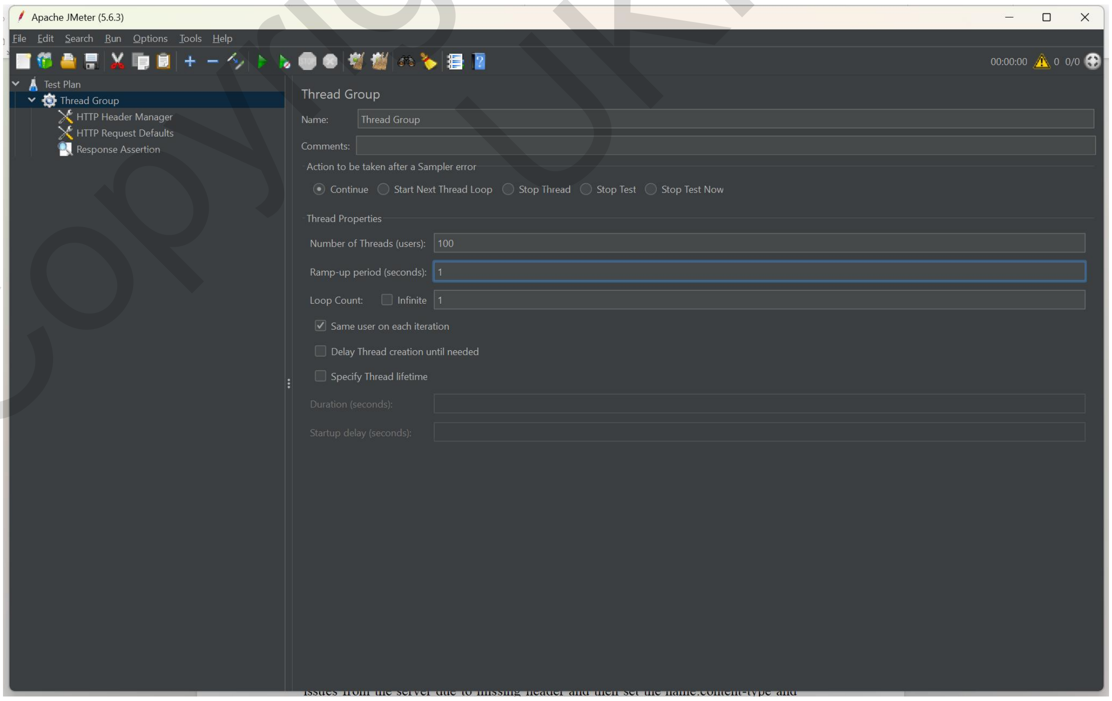
Selecting Functional Test Mode may adversely affect performance.

Add directory or jar to classpath

Library

Figure 1

Figure 1 shows this is jmeter default page and this software can test easily


Apache JMeter (5.6.3)
File Edit Search Run Options Tools Help
00:00:00 0 0/0
Test Plan
Thread Group
HTTP Header Manager
HTTP Request Defaults
Response Assertion
issues from the server due to missing header and then set the name.content-type and

Thread Group

Name: Thread Group

Comments:

Action to be taken after a Sampler error

● Continue ○ Start Next Thread Loop ○ Stop Thread ○ Stop Test ○ Stop Test Now

Thread Properties

Number of Threads (users): 100

Ramp-up period (seconds): 1

Loop Count: ☐ Infinite 1

☑ Same user on each iteration

☐ Delay Thread creation until needed

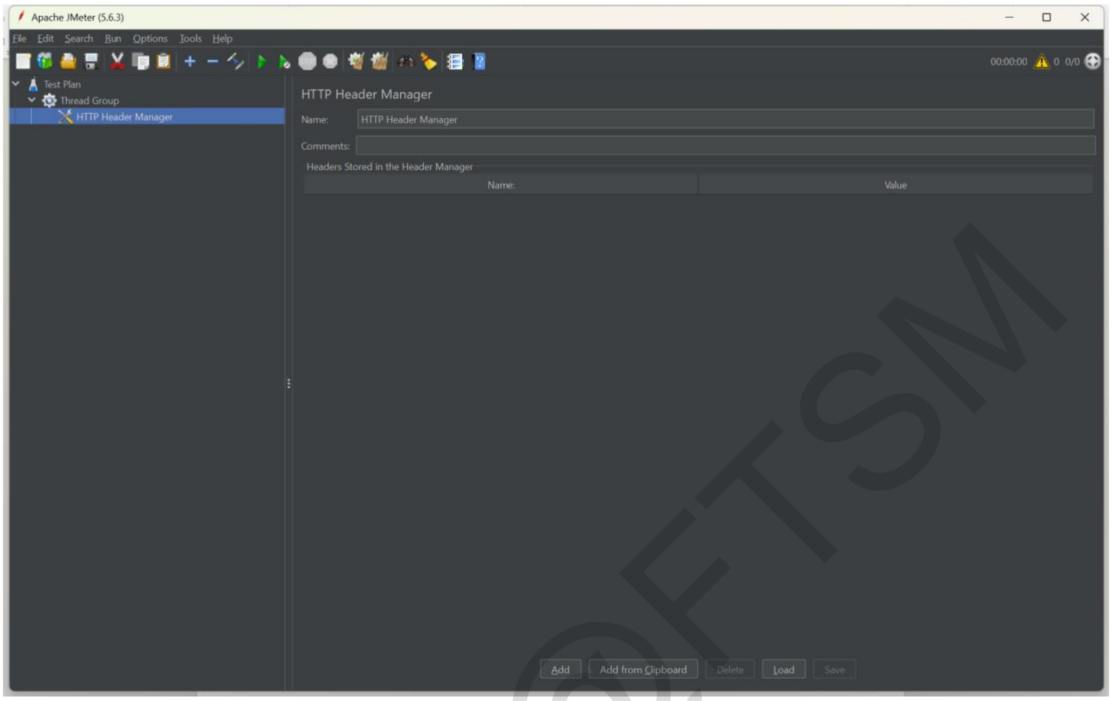
☐ Specify Thread lifetime

Duration (seconds):

Startup delay (seconds):

Figure 2

Figure 2 shows This project uses thread group to stress test.jmeter set 100 number of Threads which means 100(users) to request in 1 second.



HTTP Header Manager

Name: HTTP Header Manager

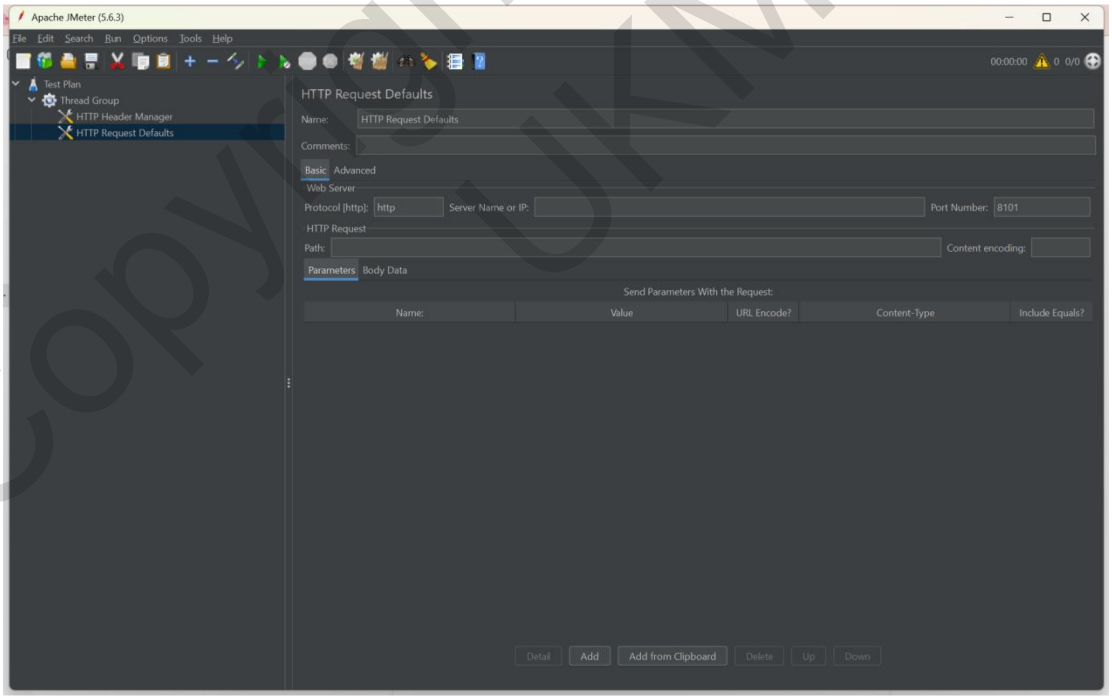
Comments:

Headers Stored in the Header Manager

Name:	Value

Figure 3

Figure 3 shows in first step,Jmeter neet to set a HTTP Header Manager.This prevents errors or request issues from the server due to missing header and then set the name:content-type and value: application/json to avoid 415 unsupport media type.



HTTP Request Defaults

Name: HTTP Request Defaults

Comments:

Web Server

Protocol [http]: http Server Name or IP: Port Number: 8101

HTTP Request

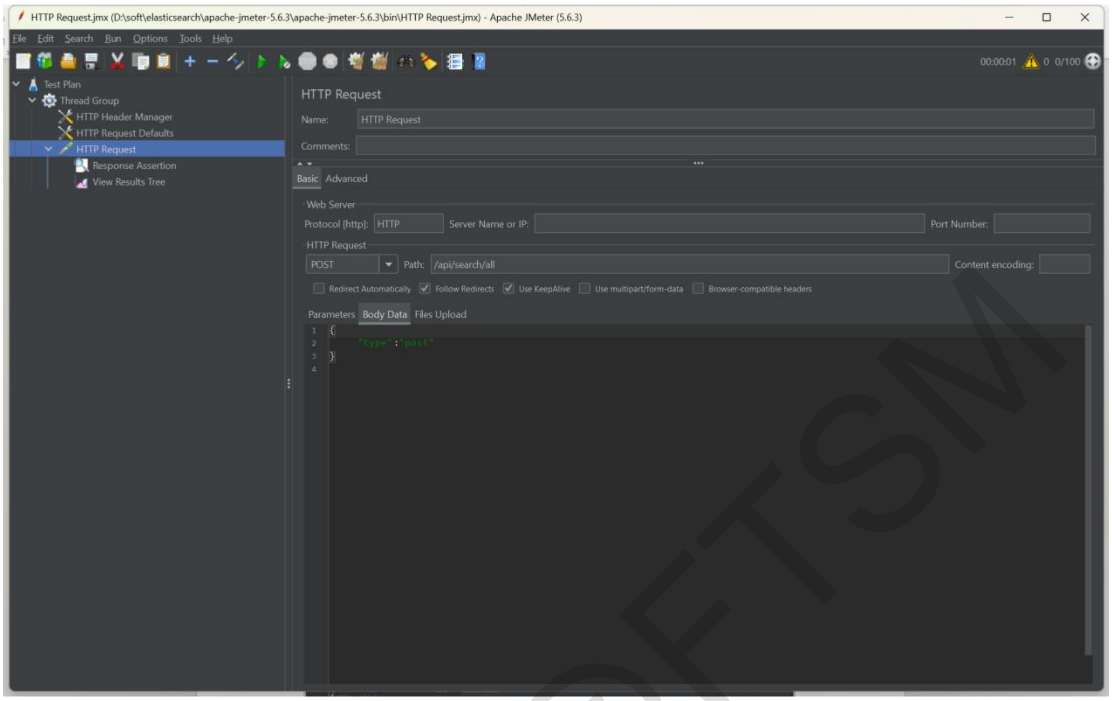
Path: Content encoding:

Send Parameters With the Request:

Name:	Value	URL Encode?	Content-Type	Include Equals?

Figure 4

Figure 4 shows in the second step,Jmeter neet to set a HTTP Request Defaults.The purpose of setting HTTP Request Defaults is to provide a default port for all http requests.



HTTP Request.jmx (D:\soft\elasticsearch\apache-jmeter-5.6.3\apache-jmeter-5.6.3\bin\HTTP Request.jmx) - Apache JMeter (5.6.3)

HTTP Request

Name: HTTP Request

Comments:

Web Server

Protocol [http]: HTTP　Server Name or IP:　Port Number:

HTTP Request

POST　Path: /api/search/all　Content encoding:

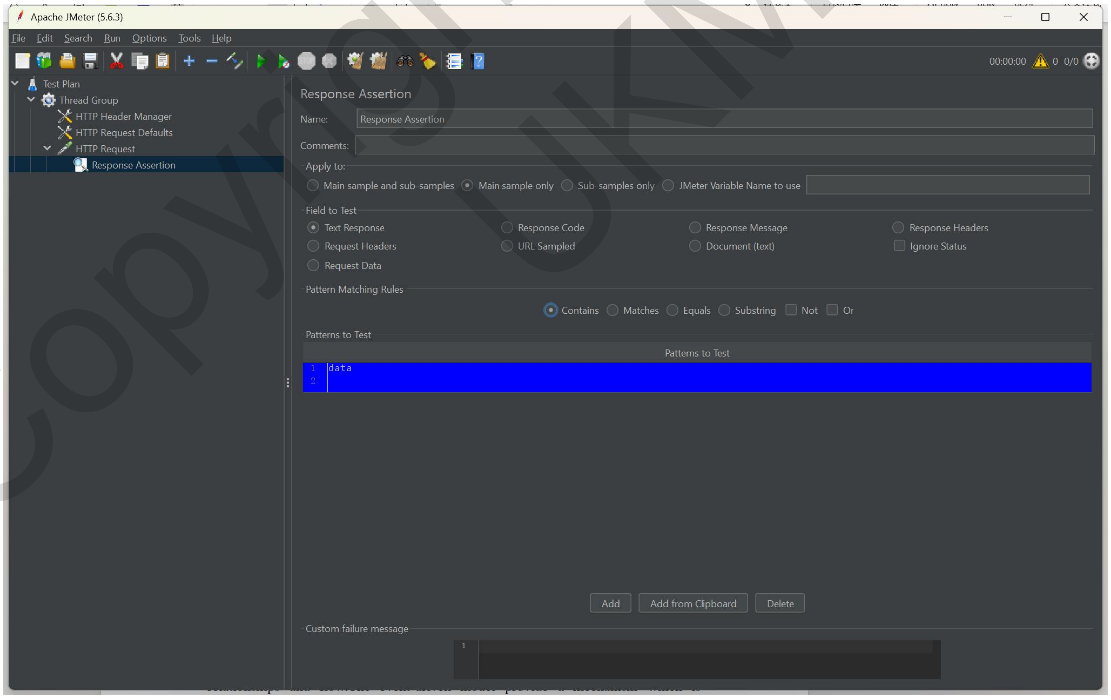
□ Redirect Automatically　☑ Follow Redirects　☑ Use KeepAlive　□ Use multipart/form-data　□ Browser-compatible headers

Body Data

```
{
    "type":"post"
}
```

Figure 5

Figure 5 shows in third step,Jmeter set a HTTP request to simulate a post request which use HTTP protocol and 8101 port.HTTP request contains a post request path:/api/search/all the back-end @postmapping and body data:{“type”:”post”} to specify the request payload



Apache JMeter (5.6.3)

Response Assertion

Name: Response Assertion

Comments:

Apply to:

○ Main sample and sub-samples　● Main sample only　○ Sub-samples only　○ JMeter Variable Name to use

Field to Test

● Text Response　○ Response Code　○ Response Message　○ Response Headers

○ Request Headers　○ URL Sampled　○ Document (text)　□ Ignore Status

○ Request Data

Pattern Matching Rules

● Contains　○ Matches　○ Equals　○ Substring　□ Not　□ Or

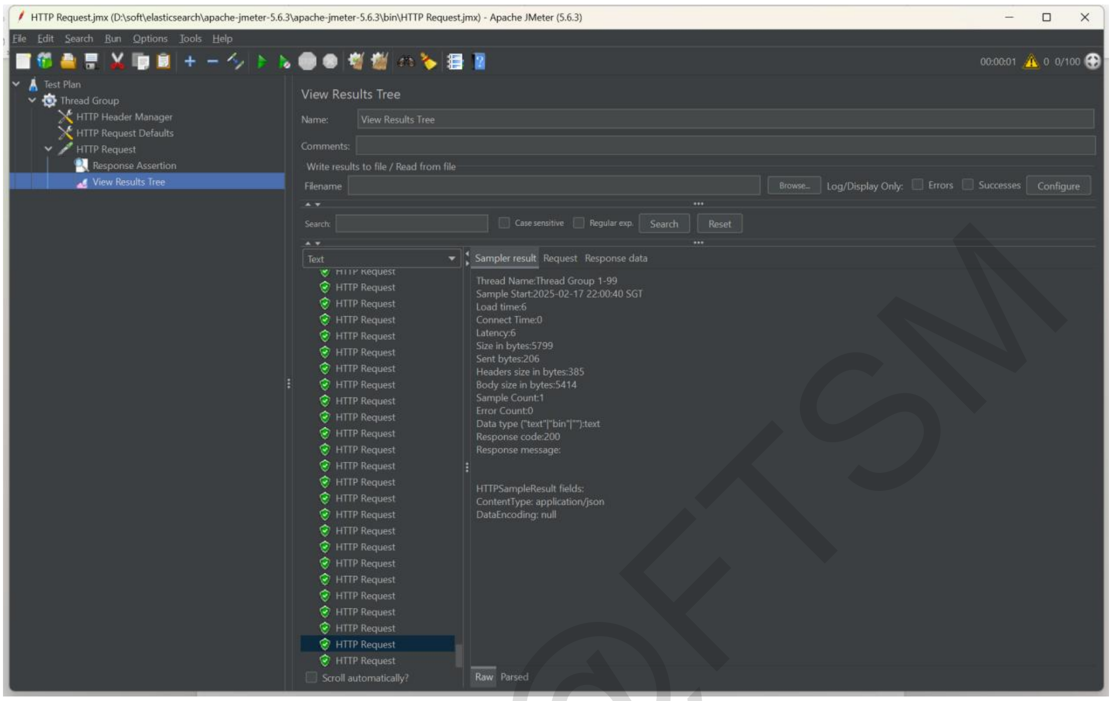
Patterns to Test

```
data
```

Custom failure message

Figure 6

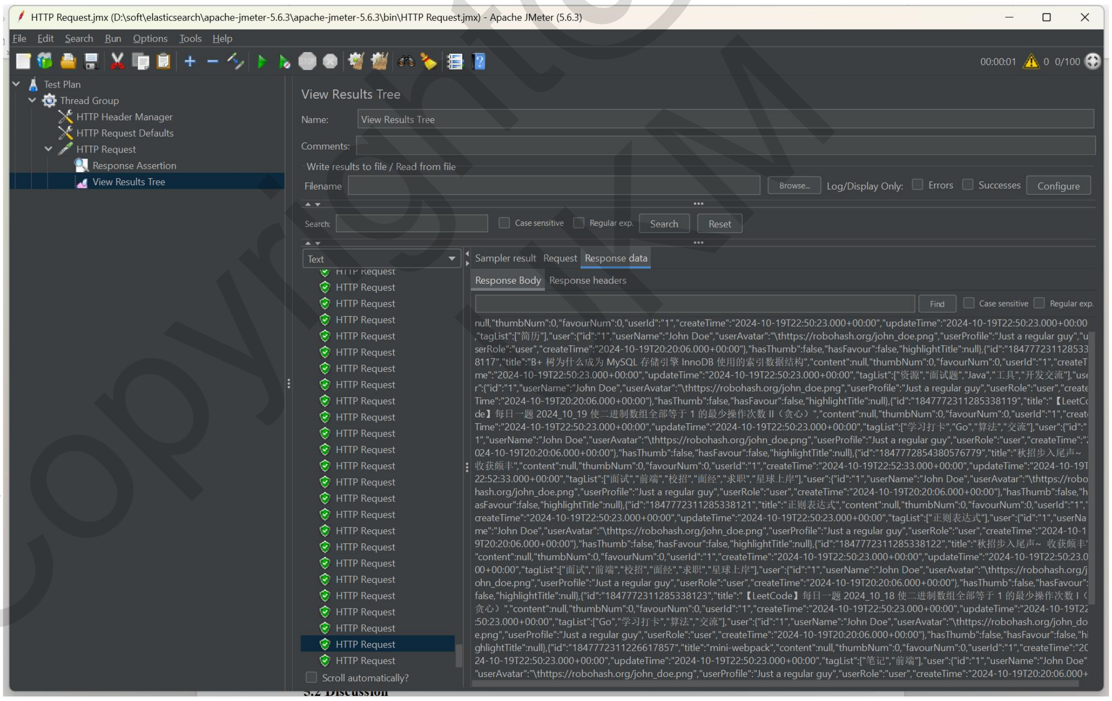
Figure 6 shows in fourth step,Jmeter should set a response assertion in the HTTP request and pattern to test: data to ensure the result return the correctly data structure.response assertion ensures that the server response meets expectations.Contain rule peforms partial match.



```
Thread Name:Thread Group 1-99
Sample Start:2025-02-17 22:00:40 SGT
Load time:6
Connect Time:0
Latency:6
Size in bytes:5799
Sent bytes:206
Headers size in bytes:385
Body size in bytes:5414
Sample Count:1
Error Count:0
Data type ("text"|"bin"|""):text
Response code:200
Response message:


HTTPSampleResult fields:
ContentType: application/json
DataEncoding: null
```

Figure 7



```
null,"thumbNum":0,"favourNum":0,"userId":"1","createTime":"2024-10-19T22:50:23.000+00:00","updateTime":"2024-10-19T22:50:23.000+00:00
,"tagList":["简历"],"user":{"id":"1","userName":"John Doe","userAvatar":"\thttps://robohash.org/john_doe.png","userProfile":"Just a regular guy","u
serRole":"user","createTime":"2024-10-19T20:20:06.000+00:00"},"hasThumb":false,"hasFavour":false,"highlightTitle":null},{"id":"184777231128533
8117","title":"B+ 树为什么成为 MySQL 存储引擎 InnoDB 使用的索引数据结构","content":null,"thumbNum":0,"favourNum":0,"userId":"1","createT
me":"2024-10-19T22:50:23.000+00:00","updateTime":"2024-10-19T22:50:23.000+00:00","tagList":["资源","面试题","Java","工具","开发交流"],"use
r":{"id":"1","userName":"John Doe","userAvatar":"\thttps://robohash.org/john_doe.png","userProfile":"Just a regular guy","userRole":"user","create
Time":"2024-10-19T20:20:06.000+00:00"},"hasThumb":false,"hasFavour":false,"highlightTitle":null},{"id":"1847772311285338119","title":"【LeetCo
de】每日一题 2024_10_19 使二进制数组全部等于 1 的最少操作次数 II（贪心）","content":null,"thumbNum":0,"favourNum":0,"userId":"1","creat
Time":"2024-10-19T22:50:23.000+00:00","updateTime":"2024-10-19T22:50:23.000+00:00","tagList":["学习打卡","Go","算法","交流"],"user":{"id":"
1","userName":"John Doe","userAvatar":"\thttps://robohash.org/john_doe.png","userProfile":"Just a regular guy","userRole":"user","createTime":"2
024-10-19T20:20:06.000+00:00"},"hasThumb":false,"hasFavour":false,"highlightTitle":null},{"id":"1847772854380576779","title":"秋招步入尾声~
收获颇丰","content":null,"thumbNum":0,"favourNum":0,"userId":"1","createTime":"2024-10-19T22:52:33.000+00:00","updateTime":"2024-10-19T
22:52:33.000+00:00","tagList":["面试","前端","校招","面经","求职","星球上岸"],"user":{"id":"1","userName":"John Doe","userAvatar":"\thttps://robo
hash.org/john_doe.png","userProfile":"Just a regular guy","userRole":"user","createTime":"2024-10-19T20:20:06.000+00:00"},"hasThumb":false,"h
asFavour":false,"highlightTitle":null},{"id":"1847772311285338121","title":"正则表达式","content":null,"thumbNum":0,"favourNum":0,"userId":"1","
createTime":"2024-10-19T22:50:23.000+00:00","updateTime":"2024-10-19T22:50:23.000+00:00","tagList":["正则表达式"],"user":{"id":"1","userNa
me":"John Doe","userAvatar":"\thttps://robohash.org/john_doe.png","userProfile":"Just a regular guy","userRole":"user","createTime":"2024-10-1
9T20:20:06.000+00:00"},"hasThumb":false,"hasFavour":false,"highlightTitle":null},{"id":"1847772311285338122","title":"秋招步入尾声~ 收获颇丰"
"content":null,"thumbNum":0,"favourNum":0,"userId":"1","createTime":"2024-10-19T22:50:23.000+00:00","updateTime":"2024-10-19T22:50:23.0
00+00:00","tagList":["面试","前端","校招","面经","求职","星球上岸"],"user":{"id":"1","userName":"John Doe","userAvatar":"\thttps://robohash.org/j
ohn_doe.png","userProfile":"Just a regular guy","userRole":"user","createTime":"2024-10-19T20:20:06.000+00:00"},"hasThumb":false,"hasFavour":
false,"highlightTitle":null},{"id":"1847772311285338123","title":"【LeetCode】每日一题 2024_10_18 使二进制数组全部等于 1 的最少操作次数 I（
贪心）","content":null,"thumbNum":0,"favourNum":0,"userId":"1","createTime":"2024-10-19T22:50:23.000+00:00","updateTime":"2024-10-19T22
:50:23.000+00:00","tagList":["Go","学习打卡","算法","交流"],"user":{"id":"1","userName":"John Doe","userAvatar":"\thttps://robohash.org/john_do
e.png","userProfile":"Just a regular guy","userRole":"user","createTime":"2024-10-19T20:20:06.000+00:00"},"hasThumb":false,"hasFavour":false,"hi
ghlightTitle":null},{"id":"1847772311226617857","title":"mini-webpack","content":null,"thumbNum":0,"favourNum":0,"userId":"1","createTime":"20
24-10-19T22:50:23.000+00:00","updateTime":"2024-10-19T22:50:23.000+00:00","tagList":["笔记","前端"],"user":{"id":"1","userName":"John Doe",
"userAvatar":"\thttps://robohash.org/john_doe.png","userProfile":"Just a regular guy","userRole":"user","createTime":"2024-10-19T20:20:06.000+
```

Figure 8

Figure 7 and Figure 8 shows in final step,jmeter set the view results tree to get the sampler result and the response data.The result displays the test absolutely correct.

Cadangan Penambahbaikan

To enhance the Aggregated Search Engine, several improvements can be implemented, including optimized data synchronization to ensure real-time indexing, AI-driven search ranking algorithms for more accurate and personalized results, and advanced filtering options to allow users to refine searches more effectively. Additionally, integrating a caching mechanism using Redis can significantly reduce database load and improve response times, while enhanced security measures such as multi-factor authentication and encrypted query logging can strengthen data protection. These enhancements will make the system more intelligent, adaptive, and scalable, ensuring efficient query handling and a superior user experience.

KESIMPULAN

The Aggregated Search Engine successfully integrates multiple technologies, including Spring Boot, Vue 3, MySQL, and Elasticsearch, to provide a high-performance, scalable, and user-friendly search platform. By aggregating results from various sources and implementing keyword highlighting, real-time suggestions, and categorized search results, the system enhances search efficiency and user experience. Performance testing confirmed fast response times and reliable scalability, ensuring the system can handle concurrent searches effectively. While the project achieved its core objectives, further improvements such as AI-driven ranking, enhanced real-time data synchronization, advanced filtering, and caching mechanisms can further optimize performance and accuracy. With these enhancements, the system has the potential to evolve into a more intelligent and adaptive search engine, meeting the growing demands of modern information retrieval.

PENGHARGAAN

The authors of this study would like to express their utmost respect and deep gratitude to Assoc. Prof. Ts. Dr. Abdul Hadi Abd Rahman for his invaluable guidance, encouragement, and unwavering support throughout the course of this research/project. His expertise and insightful feedback have greatly contributed to the success of this work.Additionally, the authors would also like to extend their heartfelt appreciation to all individuals who have directly or indirectly contributed to the successful completion of this project. Their assistance, advice, and support were truly invaluable. Without their help, this project would not have been able to progress smoothly.May God bless them all and grant them the best in return.

RUJUKAN

Achsas, S. (2022, March). Academic aggregated search approach based on BERT language model. In *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)* (pp. 1-9). IEEE.

Chen, J., Huang, X., & Li, Y. (2024). Dynamic supplementation of federated search results for reducing hallucinations in LLMs.

Erdmann, A., Arilla, R., & Ponzoa, J. M. (2022). Search engine optimization: The long-term strategy of keyword choice. *Journal of Business Research, 144*, 650-662. https://doi.org/10.1016/j.jbusres.2022.01.065

Hatcher, W. G., Qian, C., Gao, W., Liang, F., Hua, K., & Yu, W. (2021). Towards efficient and intelligent Internet of Things search engine. *IEEE Access, 9*, 15778-15795. https://doi.org/10.1109/ACCESS.2021.3052759

Lewandowski, D., & Schultheiß, S. (2022). Public awareness and attitudes towards search engine optimization. *Behaviour & Information Technology, 42*(8), 1025–1044. https://doi.org/10.1080/0144929X.2022.2056507

Oliveira, B., & Teixeira Lopes, C. (2023, March). The evolution of web search user interfaces—An archaeological analysis of Google search engine result pages. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval* (pp. 55-68).

Whitley, S. C., Chakravarty, A., & Wang, P. (2025). Positive emotions during search engine use: How you feel impacts what you search for and click on. *Journal of Marketing, 89*(1), 77-93. https://doi.org/10.1177/00222429241263012

Xiong, H., et al. (2024). When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing, 17*(6), 4558-4577. https://doi.org/10.1109/TSC.2024.3451185

SONGYUFENG(A178877)
Assoc. Prof. Ts. Dr. Abdul Hadi Abd Rahman
Fakulti Teknologi & Sains Maklumat
Universiti Kebangsaan Malaysia