Information Detrieval

IR Project Final Presentation

Group 41

Aditya Pratap Singh, Arnav Goel, Ashutosh Gera, Medha Hira, Nalish Jain, Shikhar Sharma

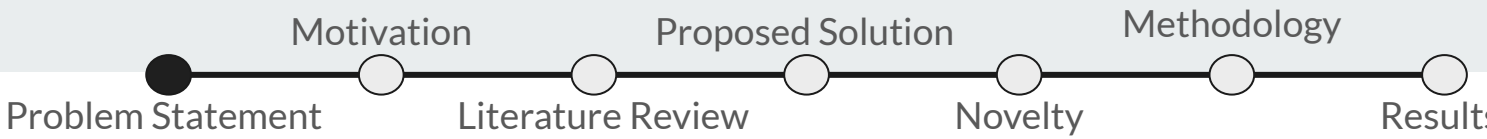


Table of Contents

Problem Statement

- **Challenge of Fragmented User Experience:** Users face significant hurdles in accessing relevant information and personalized assistance across different digital platforms due to the lack of a unified assistant, impacting productivity and accessibility.

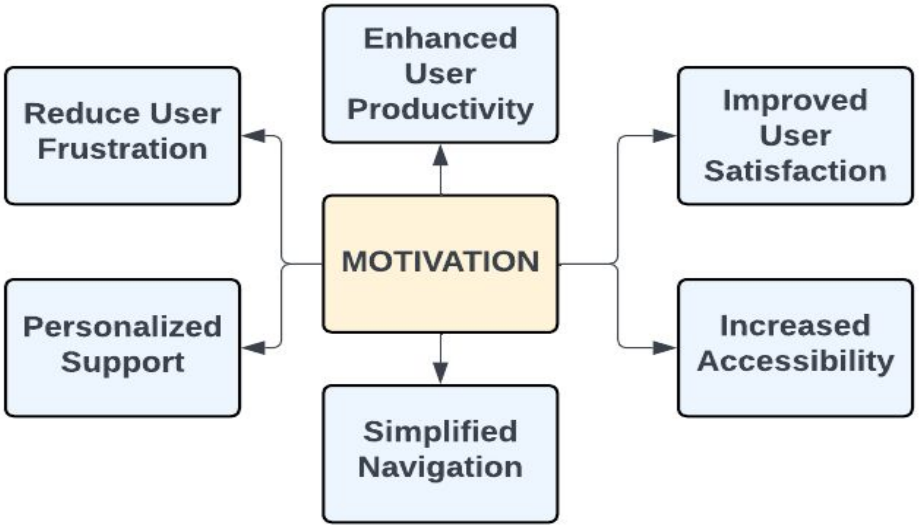
- **Reliability in Chat-bots:** Our browser extension addresses the common issue of information inaccuracy in chat-bots, known as hallucination, by using controlled history exposure (CHE) and aggregating data from diverse sources to enhance the accuracy and reliability of responses.

Motivation

Addressing the lack of a unified intelligent assistant across website tabs is crucial for enhancing user experience. Here are the issues caused by its absence:

- **Fragmented User Experience**: The absence of a consistent assistant leads to disjointed interactions, reducing productivity and user satisfaction.
- **Accessibility Concerns**: Without a unified system, users struggle to navigate and access information across multiple tabs, negatively impacting accessibility.
- **Complex Online Landscape**: Navigating a vast array of digital information is daunting without a central assistant, increasing user frustration with limited chatbot responses.
- **Need for Personalized Assistance**: Users require tailored support during online interactions; without it, they receive less relevant help.

Motivation


Enhanced User Productivity
Reduce User Frustration
Improved User Satisfaction
MOTIVATION
Personalized Support
Increased Accessibility
Simplified Navigation

Literature Review

Advancements in Digital Assistance

- **Generalized Co-Pilots:** Microsoft Copilot and Perplexity AI highlighted as key players enhancing user browsing and searching experiences.
- **Conversational Information Seeking (CIS):** Innovations enabling direct conversation with search engines, fundamentally changing information retrieval dynamics. **(Ren et al, SIGIR 2021)**
- **Enhancement Through Retrieval-Augmented Models:** Implementing Retrieval-Augmented Generative (RAG) models enhances domain-specific question answering by dynamically using external knowledge sources for more accurate and relevant responses **(Izacard and Grave, EACL 2021)**

Literature Review

Innovative Technologies in AI and Data Extraction

- **Data Extraction Techniques:** Integration of NLTK's punkt tokenizer and Body Text Extraction (BTE) algorithms for precise content extraction from web pages **(Bevendorff et al, SIGIR 2023)**
- **Multimodal Large Language Models (MM-LLM)**: Introduction of NExT-GPT which leverages multimodal adaptors for diverse interactions, enhancing AI's understanding across different media **(Wu et al, 2023)**
- **Retrieval-Augmented Question Answering:** Usage of Retrieval-Augmented Generative (RAG) models to improve accuracy in domain-specific question answering by incorporating external knowledge sources **(Anand et al, BDA 2023 ; Lewis et al, CoRR 2020)**

Proposed Solution

- **Web Scraping:**

  The `scrape_websites` method fetches web content using GET requests, parsing HTML with BeautifulSoup 4. It employs domain-specific CSS selectors for content extraction, which includes text, images, and their ALT texts. Cleaned text and HTML are stored locally, alongside metadata in JSON format that details the URL and page title after processing with NLP techniques like stopword removal and lemmatization.

- **Pinecone Initialization:**

  This stage involves setting up Pinecone indices, a vector database service used for storing and searching high-dimensional vectors derived from the web content and image ALT texts.

Proposed Solution

- **Index Initialization:**

  Utilizes an ingestion pipeline to transform text content into vectors using a sentence splitter and an embedding model. These vectors are then stored in a separate vector index. This setup supports both pre-filtering and post-filtering methods for efficient querying, with metadata enrichment aiding the first stage of filtering.

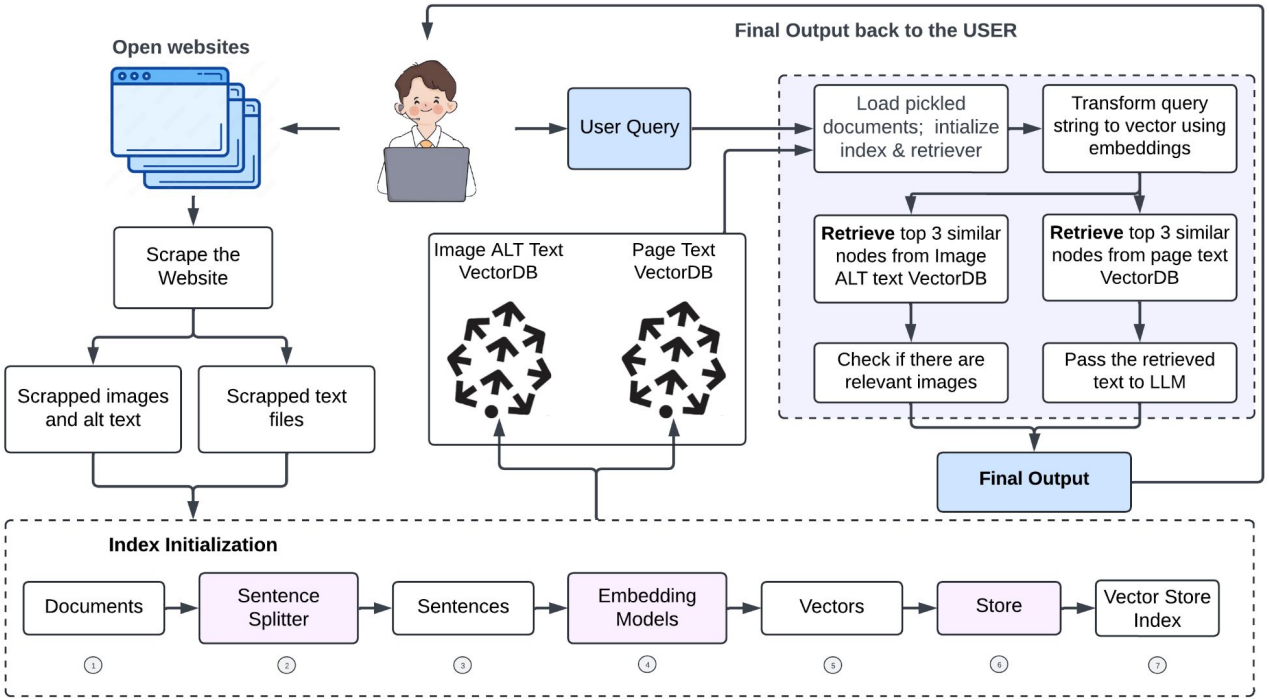
- **Query Execution:**

  The `run_query` method manages the querying process by loading pre-pickled documents or pickling them as required, initializing the index and retriever, and using the retriever to find the most similar vectors to the query. For image ALT texts, vectors are filtered further based on relevance.

- **Output Generation:**

  Results from the page text vectors are sent to a Large Language Model (LLM) to generate contextual responses. The final output to users includes both text and relevant images retrieved from the indices.

Proposed Solution - Pipeline



Novelty

- **Enhanced Contextual Understanding:** Leverages real-time data scraping and a history of up to three previous tabs to provide a deeply personalized and contextually aware user experience tailored to specific user actions and preferences.
- **Multimodal Integration:** Incorporates both input and output modality options, supporting a robust multimodal retrieval system that enhances the way users interact with and receive information, adapting to varied user needs and preferences.
- **Personalization through Controlled History Exposure:** Employs a sophisticated mechanism to selectively utilize user's browsing history, enhancing the personalization of content and responses by aligning with the user's immediate and past interactions with various web pages.

Methodology and Experiments

Query Generation

Data Collection:

- Websites from the IIITD domain, Taj and legal law were scraped and saved as .txt files.
- Each website provided content on topics like B.Tech Project, Guest House, Hostels, Mess, Student Affairs, etc.
- The corpus comprised 10 documents for each website, forming the basis for experimentation.

Query Generation:

- Each .txt file was processed through OpenAI's ChatGPT chatbot with the zero-shot prompt.
- The prompt instructed the chatbot to generate 5 question-answer pairs based on the given text.

Methodology and Experiments

Query Generation

Selection of Queries:

- From the generated pairs, the top 2 queries(1 for legal law) were selected for each document. The total number of queries is 50.
- These selected queries were then pooled together for testing our retrieval system.

Methodology and Experiments

LLM Benchmarking

Dataset and Technique:

- We utilize the WikiQA dataset to evaluate the various large language models.
- This involves feeding the LLMs with a query, general instructions, and the retrieved context, employing the RAG technique to enhancerelevance and accuracy

Models Benchmarked:

- Mixtral-7x8b, LLaMa-3-8b, Gemma-7b, and GPT-3.5-turbo

Methodology and Experiments

LLM Benchmarking

Procedure:

- For each query, retrieve the most relevant Wikipedia articles.
- The retrieved pages serve as the context for the LLMs, along with some general instructions.

Evaluation Metrics:

- BLEU-score, METEOR-score, and BERTscore (precision, recall and f1)

Results - Query Generation

K	Avg Precision	
1	0.85	
2	0.5	
3	0.33	
	MAP	0.925

Table 1: Pipeline Results using Cosine Similarity (IIIT D)

K	Avg Precision	
1	0.88	
2	0.47	
3	0.31	
	MAP	0.916

Table 2: Pipeline Results using Cosine Similarity (Taj)

K	Avg Precision	
1	0.93	
2	0.68	
3	0.52	
	MAP	0.96

Table 4: Updated Combined Pipeline Results for all 3 websites

Results - LLM Benchmarking

Model	BLEU	METEOR	B-P	B-R	B-F1
Mixtral-8x7b	0.013	0.183	0.55	0.675	0.605
LLAMA-3-8b	0.012	0.171	0.595	0.672	0.628
Gemma-7b	0.009	0.113	0.627	0.660	0.641
GPT-3.5-turbo	**0.021**	**0.232**	**0.654**	**0.713**	**0.687**

Table 5: Performance comparison of different LLMs on various metrics. B-P: BERT-Precision, B-R: BERT-Recall, B-F1: BERT-F1.

THANK YOU!

ANY QUESTIONS?