SoftwareX 27 (2024) 101872

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original Software Publication

End-to-end vertical web search pseudo relevance feedback queries recommendation software

Tajmir Khan [a], Umer Rashid [a,*], Abdur Rehman Khan [b], Naveed Ahmad [c], Mohammed Ali Alshara [c]

[a] Department of Computer Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan
[b] School of Computer Science, Centre for Data Science, Queensland University of Technology, Brisbane, QLD 4001, Australia
[c] College of Computer and Information Sciences, Prince Sultan University, Riyadh 12435, Saudi Arabia

ARTICLE INFO

Keywords:
Pseud relevance feedback
Web search engine
Exploratory search
Information retrieval

ABSTRACT

Users' web information needs are increasingly exploratory, seeking to navigate unfamiliar domains and discover knowledge. However, existing search engines struggle with ambiguous queries, leading to irrelevant results. To address this, we propose an architecture that autonomously extracts domain knowledge from initial queries. This system transforms queries into a semantic model, suggesting relevant queries to aid exploration. Evaluated against Google, our architecture achieved 89 % accuracy in automated query recommendation, with 85.4 % system usability and an "A" ranking. This approach enhances exploratory search by facilitating clearer, more effective query formulation, thus improving information retrieval for users.

Metadata

Nr	Code metadata description	*Please fill in this column*
C1	Current code version	*V1*
C2	Permanent link to code/repository used for this code version	https://github.com/HQuser/Pseudo-feedback-software
C3	Permanent link to reproducible capsule	https://codeocean.com/capsule/6163061/tree/v1
C4	Legal code license	MIT License.
C5	Code versioning system used	Git
C6	Software code languages, tools and services used	Python 3
C7	Compilation requirements, operating environments and dependencies	Torch, protobuf, sentencepiece, transformers, google-search-results, sentence-transformers, python-dateutil, beautifulsoup4, sumy
C8	If available, link to developer documentation/manual	NA
C9	Support email for questions	umerrashid@qau.edu.pk

1. Motivation and significance

The multimedia information on the web is proliferating [1]. Zettabytes of data is generated yearly, and the data expansion is nearly exponential [2]. To organize the multimedia information on the web, the existing web search engines such as Google, Yahoo, and Bing, etc., assemble the media-specific content into disjoint repositories called verticals [3]. The user query is dispatched to the disjoint verticals such as web, image, video, and news, etc., to retrieve the relevant results [4]. To further enhance the retrieval efficacy, the search engines blend top-few search results into initial Search Engine Result Page (SERP) [5]. This paradigm best suits the lookup searches where the user query is well-articulated and top-few search results may suffice the users' information needs [6].

At present, the users' information need on the web are becoming exploratory [7]. The resultant users' information needs have become open-ended and multifaceted with a goal to explore unfamiliar domain and discover knowledge [7,8]. Contrarily, existing web search engines require users to formulate well-articulated queries for retrieval of relevant information [9]. Whereas in exploratory searches, the users' information needs are ill-defined due to lack of prior domain knowledge. As a result, existing web users often issue short-length ambiguous queries that results in retrieval of irrelevant results [10].

Existing literature solves this problem by systematically assisting a user through enhanced queries based on user's intent (known as relevance feedback) [11]. These approaches primarily work on the notion of taking the users' feedback implicitly or explicitly [12]. The former aims to understand the user intent from search behaviour, such as

* Corresponding author.
E-mail address: umerrashid@qau.edu.pk (U. Rashid).

https://doi.org/10.1016/j.softx.2024.101872
Received 13 May 2024; Received in revised form 15 July 2024; Accepted 29 August 2024
Available online 6 September 2024

click-through rates or dwell time on search results, while the latter involves direct user input, such as rating search results or providing explicit feedback. The implicit feedback approaches suffer from degraded accuracy of suggested queries whereas the explicit feedback approaches add undue burden to the user [13]. To maintain a balance, we propose an architecture that works on the premises of pseudo-relevance feedback; wherein the initial intent is taken to retrieve the relevant results and based on the retrieved corpus, the system further suggests the exploration paths via well-articulated queries.

While the existing pseudo queries recommendation approaches are proposed and evaluated from the system metrics e.g., precision and recall [14–19]. Such systems use offline dataset and may lack effectiveness when deployed in real environment. There exists a lack of approaches that focus user-in-the-loop. Hence, a research gap exists to design and evaluate the pseudo relevance feedback recommender systems that capture the users' realistic behaviour and tailor their searching experience. Therefore, to overcome the aforementioned challenges, we designed the proposed approach from the user-centric perspectives.

Specifically, the proposed architecture acts as a generic middle-layer that autonomously extracts the domain knowledge from the users' initial query intent from the search engine and suggest the well-articulated queries based on Lasswell communication model of 5W's. The 5W models defines the questions starting with (Who, What, Where, When, and Why) and has shown proven effectiveness in unfamiliar exploratory environments [20,21]. Such pattern enhances a search engine's contextual knowledge for retrieving relevant information (e.g., entity, place, date, etc.) when adequately augmented with domain-specific keywords. Otherwise, due to lack of user domain knowledge, it becomes challenging to assemble manually. The proposed system architecture was evaluated against the Google (baseline) search engine on general exploratory topics achieved 89 % accuracy scores for the automated queries recommendation, with 85.4 % system usability (N=37), and "A"-grade ranking. To the best of our knowledge, we are the first to propose a generalized pseudo relevance feedback approach that practically considers the user-in-the-loop during the framework design and evaluation process.

2. Software description

The proposed vertical web search pseudo relevance feedback queries recommendation software works on the premises of the user's initial intent. This intent is systematically processed, and the list of well-articulated queries are generated to the users ranked by the high relevancy to the initial intent. The recommended query can further be passed to the proposed software iteratively to facilitate guided search journey. The complete details of the internal software working are outlined in Algorithm 1.

2.1. Software architecture

The proposed software architecture acts like a package that can be employed as a sub-component of any search system. The internal working of proposed architecture is divided into three components. Mainly, scrapper (Fig. 1a), summarizer (Fig. 1b), and queries generator (Fig. 1c) component. The detailed working of each component is provided in the subsequent subsection.

Software functionalities*:*

2.1.1. Scrapper

This component takes in the user's initial query intent f(q), which is passed to the various verticals = f(get_all_verticals(q)) of the search engine in real-time. The search results from the verticals are retrieved and necessary preprocessing f(preprocess(verticals)) is applied to remove non-word characters. These verticals are then aggregated into a single processed dictionary. In our experiments, the maximum limit to fetch the search results from each vertical was 100. Therefore, we used

Algorithm 1
End to End Pseudo Relevance Feedback Queries Generation Model

```
Input: user initial query intent (q)
Output: List of sorted recommended queries
function preprocess(data)
    processed ← filter(data, /[^A-Z\s]+/g)
    return processed
function extract_text (data)
    title ← preprocess(data["title"])
    description ← preprocess(data["description"])
    return title + description
function get_all_verticals(q)
    verticals ← get_verticals(q, ["web", "image", "news", "video"])
    for each v ∈ vertical:
        for each r ∈ result:
            aggregated[r] ← {title(r), description(r), url(r), thumbnail(r)}
    return aggregated
function generate_summary(aggregated)
    summary[] ← ∅
    for each d ∈ aggregated:
        summary[d] ← LSA(d["title"] + d["description"])
    return summary
function generate_questions(summary)
    questions[] ← ∅
    for each s ∈ summary:
        candidate_queries[] ← 5W_model(summary)
    return filter(questions, /^(what|why|who|how|where)/)
function rank_queries(candidate_queries, q)
    sorted_recommended_queries[] ← ∅
    query_transformer ← BERT(q)
    for each c ∈ candidate_queries:
        sentence_transformer[c] ← BERT(c)
        cosine_similarity ← cosine(sentence_transformer[c], query_transformer)
    sorted_recommended_queries[c] ← sort(cosine_similarity, descending=true)
    return sorted_recommended_queries
```

the top-100 documents from each vertical as our pseudo relevance feedback.

2.1.2. Summarizer

The summarizer module extracts the pseudo relevance feedback from the retrieved information. Mainly, the title and description f (extract_text(verticals)) from the retrieved search results are extracted. Each search result is now represented as a document (d). Each document is passed to the latent semantic analyzer d = f(LSA(d)), wherein, the document is summarized based on the most important terms inside. Afterwards, all the summarized documents s = {∀r ∈ d} are aggregated in a single dictionary to facilitate queries generation. Since the existing search engines are optimized from the precision perspectives [2], at this stage, we refrained to filter out the irrelevant search results to retain maximum diversity in queries generation process. Otherwise, the resultant corpus could limit the possible number of search directions provided to the user. However, this decision may result in less generation of queries less relevant the user intent. To overcome this challenge, we passed the documents to our subsequent module to determine the relevance according to the user intent while maintaining diversity, explained in following subsection.

2.1.3. Queries generator

The summarized documents(s) are forwarded to the end-to-end questions generation model that works on the text-to-text transfer transformer architecture. This model outputs the list of candidate recommendation queries c = f(5W_model(s)) from diverse corpus that adhere to the principle of Lasswell's 5W communication model. To further filter the irrelevant queries, the recommended queries are subsequently passed to f(rank_queries(c, q)) in which transforms the queries via bidirectional encoder representations from transformers = f(BERT (c)) to determine the relevancy to the user intent (q) wherein each recommended query and the user initial intent I = f(BERT(q)) are enclosed in embeddings. Finally, a pairwise semantic similarity = f (cosine(transformers$_i$, I)) comparison is performed between intent and

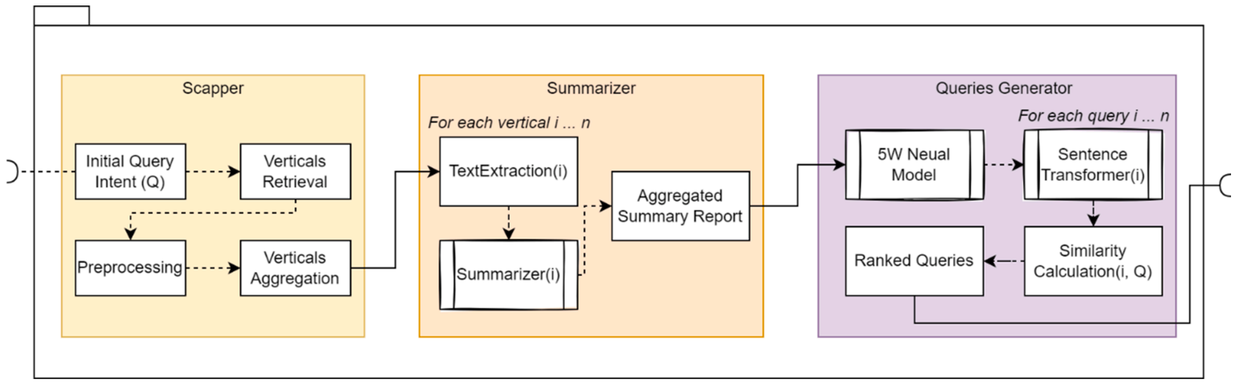

Fig. 1. Proposed software architecture consisting of three components; (a) Scrapper, (b) Summarizer, and (c) Queries Generator component. The rectangular boxes represent routines whereas the double lined rectangular boxes represent machine learning model. The dotted arrow lines represent dependency, solid lines represent information passing, and the moon line shows coupling.

the recommended queries are sorted in descending order of recommended queries = f(sort(similarity, descending=true)).

The proposed software can be employed as an addon to a search system. In the illustrative example (Fig. 2), a user first enters their search intent (Fig. 2 (a)). This search intent is passed to the proposed architecture, which performs scrapping from the web search engine and retrieves the resultant search results (Fig. 2 (b)). Meantime, the user initial intent is summarized, and the queries are generated via the summarizer and queries generator component of the proposed architecture. The generated queries are subsequently presented to the user (Fig. 2 (c)). The scrapper component saves the contextual information of each search result, allowing user to browse any information snippet (Fig. 2 (d)). In case a user is unsatisfied with the retrieved search results quality, they can choose a query from recommendation panel (Fig. 2 (e)). The same process is performed with the user chosen query as depicted in (Fig. 2 (a), (b), and (c)). This allows users to iteratively explore the search results space by generating personalized queries according to updated user information needs (Fig. 2 (g) and (h)).

2.2. Impact

Existing web search systems rely on well-articulated queries for retrieval of relevant search results [22]. Contrarily, web users having exploratory information needs face challenges in articulating the queries [4]. Therefore, the proposed architecture takes in the user's pseudo intent and operates it as relevance feedback for systematic generation of well-articulated queries without added cognitive constrained to the users.

This architecture was evaluated empirically and useably [4]. Empirically, two human experts were recruited that annotated the relevancy and accuracy of the proposed queries in binary format regardless of information on data source. The truth values revealed 89 % queries accuracy. Afterwards, usability evaluation was performed on the proposed architecture by practically deploying the software (N=37) users by the researchers in [4]. The within-subjects evaluation parameters included standard System Usability Scale (SUS), After-Scenario Questionnaire (ASQ) and custom-defined subjective queries recommendation satisfaction scale. Overall, the proposed software achieved 2 points higher system usability, 10 % more after-scenario satisfaction and found 16 % better query assistance when compared with the baseline (Google) system. Furthermore, the user behavioural analysis revealed 20 % better user exploration, and 60 % less query reformulations and 66 % less clicking efforts at the users' end.

The existing queries recommendation approaches are proposed and evaluated from the system metrics e.g., precision and recall [14–19]. Such system models are trained on offline dataset and might not be

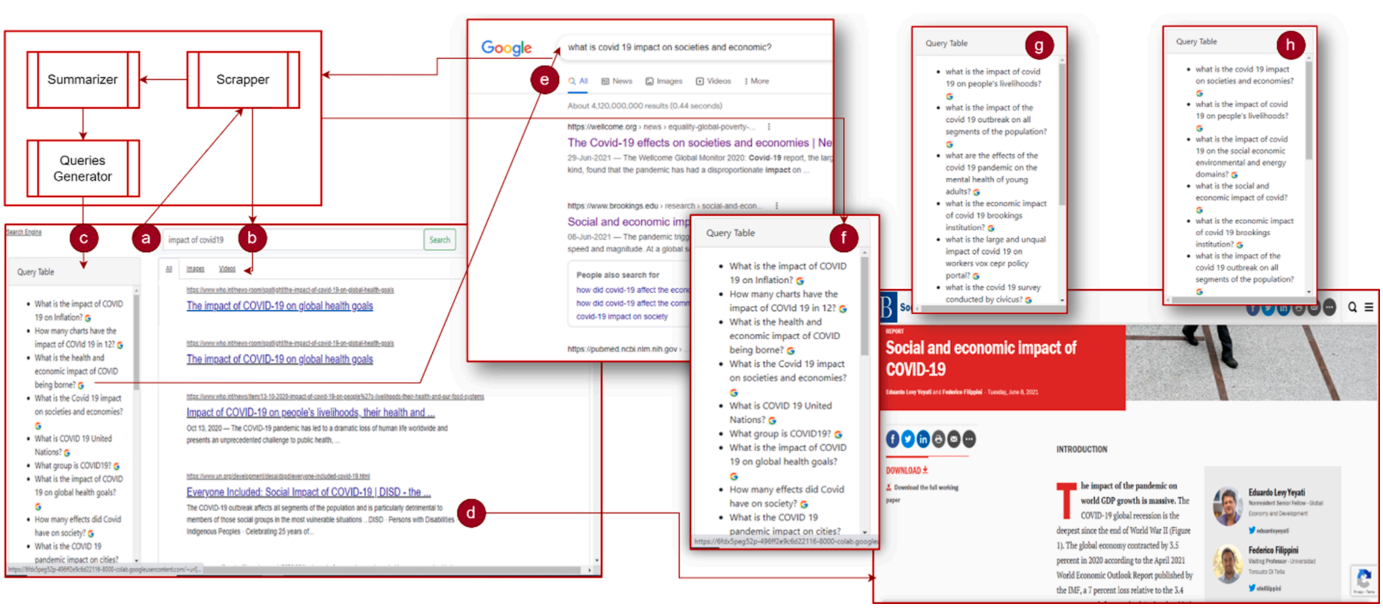

Fig. 2. Illustrative example of end-to-end vertical web search pseudo relevance feedback queries recommendation software in a demonstrative search system.

applicable to the latest or different datasets. Employing such systems in our evaluation could result in dataset-induced bias [23]. Conversely, use of offline dataset in proposed approach may have hindered capturing the users' realistic behaviour during interaction with the proposed system [24]. Therefore, to overcome the aforementioned challenges, we opted to use real dataset for multiple purposes. Firstly, this allows the proposed approach to be designed from the user-centric perspectives. Secondly, this allowed the generic framework design conception of the proposed approach without relying on a specific dataset or system. Finally, this allowed the proposed approach to be directly applicable to the most widely used search engines that have the ability to provide queries recommendation to the users. To the best of our knowledge, we are the first to propose a generalized pseudo relevance feedback approach that practically considers the user-in-the-loop during the framework design and evaluation process.

Hence, the implications of the proposed software are broad and appliable to any search system. For this purpose, the proposed architecture is designed from the modularity perspectives that can be implement regardless of domain implications. The proposed pseudo relevance feedback approach depends on the retrieved corpus rather than the user query. Therefore, regardless of the type of user query, the proposed approach would have performed in the similar manners. Furthermore, each module can be modified or fine-tuned according to the system requirements. This includes further varying the information scrapping, changing the model parameters, and utilizing the interim results to optimize exploratory outcomes, as discussed in (Section 3). The publicizing of this software will further lead to employment of the proposed software in diverse domains yielding enhanced user exploratory outcomes.

3. Conclusions

Initializing exploratory search can be challenging for the users due to lack of prior domain knowledge. To assist users in exploring unknown domains, we proposed an end-to-end modular software architecture. The software architecture consisted of three major processes. Firstly, user initial intent as pseudo relevance feedback was extracted and the corpus was retrieved from the web search engine in real-time. Afterwards, the corpus was summarized and fed to the 5W deep neural model to generate suggestive queries. These queries were subsequently transformed into sentence embeddings to perform pair-wise semantic similarity and filter irrelevant queries. Finally, the queries were presented based on descending order of closeness to the user's initial intent. The empirical and usability evaluation of the proposed architecture revealed 89 % query generation accuracy and 85.4 % overall usability, respectively. In future, we are interested in advancing the proposed software to incorporate context extraction from multiple modalities (acoustic and visual) and incorporate user logs to provide persistent and resumable queries assistance across different search domains.

CRediT authorship contribution statement

Tajmir Khan: Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Umer Rashid:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Abdur Rehman Khan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Formal analysis, Data curation. **Naveed Ahmad:** Writing – review & editing, Resources, Project administration, Funding acquisition. **Mohammed Ali Alshara:** Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data/code associated with this research has been published publically in CodeOceans and Github.

Acknowledgements

The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication. They would also like to thank Prince Sultan University for their support.

References

[1] Rashid U, Saleem K, Ahmed A. MIRRE approach: nonlinear and multimodal exploration of MIR aggregated search results. Multimed Tools Appl 2021;80(13):20217–53.
[2] Khan AR, Rashid U, Saleem K, Ahmed A. An architecture for non-linear discovery of aggregated multimedia document web search results. PeerJ Comput Sci 2021;7:e449.
[3] Khan AR, Rashid U, Ahmed N. An explanatory study on user behavior in discovering aggregated multimedia web content. IEEE Access May 2022;10:56316–30. https://doi.org/10.1109/access.2022.3177597.
[4] Khan T, Rashid U, Khan AR. End-to-end pseudo relevance feedback based vertical web search queries recommendation. Multimed Tools Appl 2024. https://doi.org/10.1007/s11042-024-18559-4.
[5] Rashid U, Saddal M, Farooq G, Khan MA, Ahmad N. An SUI-based approach to explore visual search results cluster-graphs. PLoS One 2023;18(1):e0280400.
[6] Khan AR, Rashid U, Ahmed N. AMED: Aggregated multimedia exploratory and discovery search software. SoftwareX 2023;21:101312.
[7] La Y, Yuan L. Computing and explaining knowledge relevance in exploratory search environment. In: 2022 7th International Conference on Cyber Security and Information Engineering (ICCSIE); 2022. p. 73–81.
[8] Raouf Ahmad Rather SMCLIKRH, Hollebeek Linda D. Exploring Tourists' virtual reality-based brand engagement: a uses-and-gratifications perspective. J Travel Res 2023. https://doi.org/10.1177/00472875231166598.
[9] Khan AR, Rashid U. A relational aggregated disjoint multimedia search results approach using semantics. In: 2021 International Conference on Artificial Intelligence, ICAI 2021; 2021. p. 62–7. https://doi.org/10.1109/ICAI52203.2021.9445229.
[10] H. Kaur, D. Downey, A. Singh, E. Y.-Y. Cheng, D. Weld, and J. Bragg, "FeedLens: polymorphic lenses for personalizing exploratory search over knowledge graphs," 2022. doi:10.1145/3526113.3545631.
[11] Ahmed A, Malebary SJ. Query expansion based on top-ranked images for content-based medical image retrieval. IEEE Access 2020;8:194541–50. https://doi.org/10.1109/ACCESS.2020.3033504.
[12] Balakrishnan V, Ahmadi K, Ravana SD. Improving retrieval relevance using users' explicit feedback. Aslib J. Inf. Manag. 2015.
[13] E. Sørig, N. Collignon, R. Fiebrink, and N. Kando, "Evaluation of rich and explicit feedback for exploratory search," 2019.
[14] X. Zhang et al., "Selecting query-bag as pseudo relevance feedback for information-seeking conversations," *arXiv Prepr. arXiv2404.04272*, 2024.
[15] Inje B, Nagwanshi K, Rambola R. Document retrieval using clustering-based Aquila hash-Q optimization with query expansion based on pseudo relevance feedback. Int J Comput Appl 2024:1–12.
[16] Wang J, Pan M, He T, Huang X, Wang X, Tu X. A pseudo-relevance feedback framework combining relevance matching and semantic matching for information retrieval. Inf Process Manag 2020;57(6):102342.
[17] C. Li et al., "NPRF: A neural pseudo relevance feedback framework for ad-hoc information retrieval," *arXiv Prepr. arXiv1810.12936*, 2018.
[18] Li H, Mourad A, Koopman B, Zuccon G. How does feedback signal quality impact effectiveness of pseudo relevance feedback for passage retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2022. p. 2154–8.
[19] Li H, Mourad A, Zhuang S, Koopman B, Zuccon G. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. ACM Trans Inf Syst 2023;41(3):1–40.
[20] Wei H, Chen J, Gan X, Liang Z. Eight-Element communication model for internet health rumors: a New exploration of lasswell's '5W communication model. Healthcare 2022;10(12):2507.
[21] L. Enrico Lopez, D. K. Cruz, J. C. Blaise Cruz, and C. Cheng, "Transformer-based End-to-End question generation," *arXiv e-prints*, p. arXiv–2005, 2020.

[22] Naseer S, Rashid U, Saddal M, Khan AR, Abbas Q, Daadaa Y. WSREB mechanism: web search results exploration mechanism for blind users. Appl Sci 2023;13(19): 11007.

[23] Rehman SU, Waqas M, Tu S, Koubaa A, ur Rehman O, Ahmad J, Hanif M, Han Z. Deep Learning Techniques for Future Intelligent Cross-Media Retrieval," *CISTER-Research Centre in Realtime and Embedded Computing Systems*. 2020.

[24] Jabbar A, Iqbal S, Tamimy MI, Rehman A, Bahaj SA, Saba T. An Analytical Analysis of Text Stemming Methodologies in Information Retrieval and Natural Language Processing Systems. 11. IEEE Access; 2023. p. 133681–702.