ColBERT-Based User Profiles for Personalized Information Retrieval

Aleena Ahmad*, Gibson Nkhata†, Abdul Rafay Bajwa‡, Hannah Marsico†, Bryan Le†, Susan Gauch†

* *School of Electrical Engineering and Computer Science, National University of Sciences and Technology(Pakistan)*
email: alahmad.bscs21seecs@seecs.edu.pk

† *School of Electrical Engineering and Computer Science, University of Arkansas(USA)*
email: (gnkhata, hmarsico, bryanle, sgauch)@uark.edu

‡ *Syed Babar Ali School of Science and Engineering, Lahore University of Management Sciences(Pakistan)*
email: 25100194@lums.edu.pk

UNIVERSITY OF ARKANSAS

IARIA

Presenter Biography

- Aleena Ahmad is an undergraduate student at the National University of Sciences and Technology (NUST), Pakistan, pursuing a Bachelor of Science degree in Computer Science.

Overview

- Aims and Contributions
- Our approach
- Experiments
- Results
- Discussion
- Conclusion
- Limitations and Future Work

Aims and Contributions

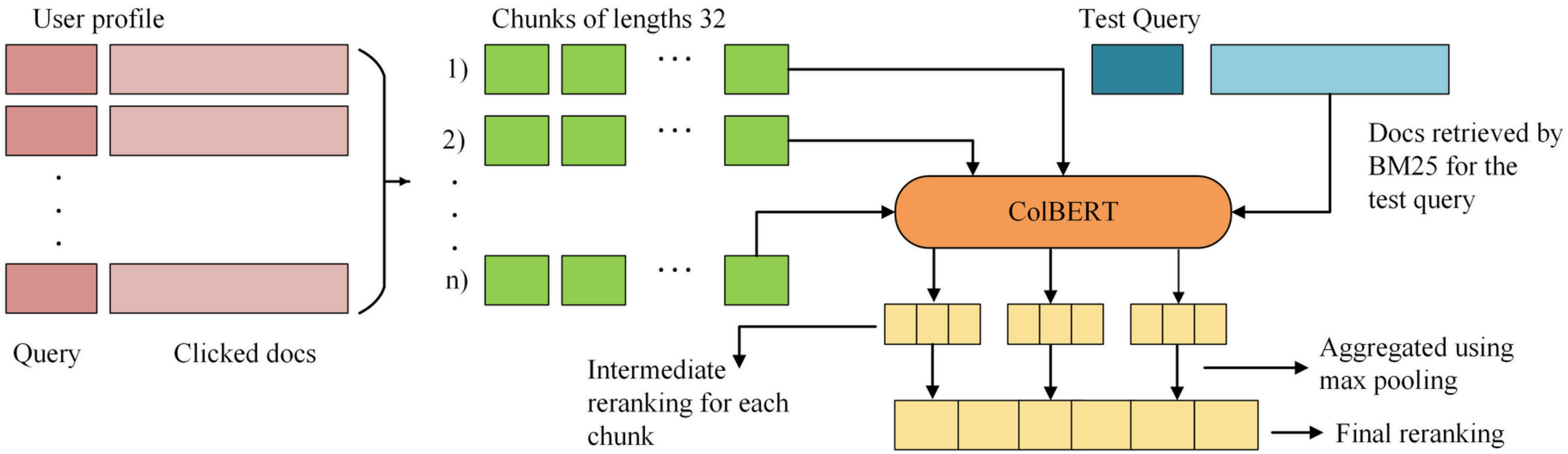
- Personalized information retrieval (PIR) is the process of tailoring search results to match user preferences.
- Traditional PIR approaches range from using feature engineering to ontologies to capture user interests.
- Recently, deep learning based approaches have emerged which rely primarily word embeddings for query expansion, by selecting representative terms from user profiles.
- These approaches face challenges such as choosing descriptive keywords and omitting crucial information.
- We propose an approach which:
  - fully encodes user profiles using contextual embeddings
  - tests a recency-frequency weighting mechanism which adjusts query influence based on its repetition frequency and time.

Our Approach

Our approach consists of the following components:

1. **Profile Generation:** a user profile consists of
   - title and full text of previously clicked documents
   - previously issued queries

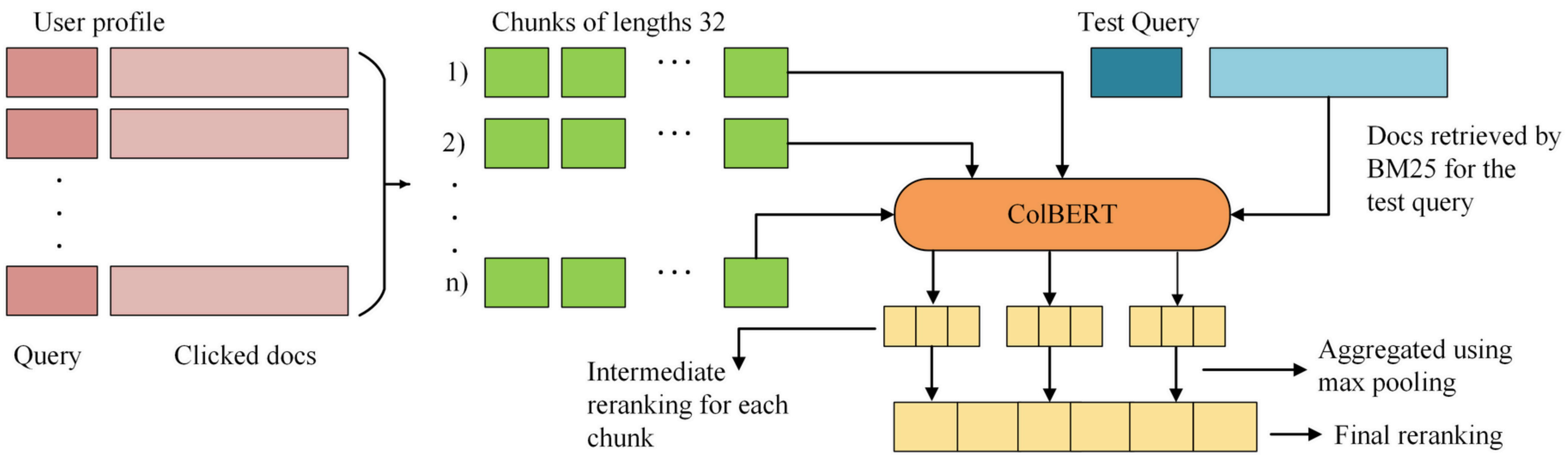
- These are aggregated to create a user profile, which is divided into chunks of 32 tokens (ColBERT's default limit).
- The chunks are represented as word embeddings by ColBERT.



Our Approach

2. **Reranking through ColBERT:**

- for each test query, top-k candidate documents are retrieved using BM25.
- each profile chunk reranks each candidate document individually → generates a candidate document list
- final ranked document list for the profile is obtained by aggregating each chunk's list through max-pooling



Frequency-Recency Approach

- Adjusts influence of each query based on:
  - Recency: recent queries get higher weight through exponential decay
  - Frequency: frequently repeated queries get higher weight (log scale)
- Weighting formula:

$$w_i = w_{\text{recency},i} \cdot w_{\text{frequency},i} = e^{-\alpha \cdot \Delta t_i} \cdot \log(1 + \beta \cdot f_i)$$

where

- α and β are parameters which are fine-tuned experimentally
- t_i is the time difference between test and past query
- f_i is the number of times query *i* has been issued by the user

This weight is multiplied with each chunk's ranking.

Experiments

We conducted experiments on two publicly available datasets:

1. **Personalized Results Reranking Benchmark (PRRB):**
   - a multi-domain dataset proposed by Bassani et al.
   - contains a total of 1.9 million queries divided across four domains.
   - The PERSON methodology was used for the construction of the dataset: a paper's title is considered as query, one of the authors considered as the user, and the referenced papers are considered relevant documents.

TABLE VI
PRRB STATISTICS FOR SAMPLE 1000.

Domain	# Users	# User Docs	# Docs/User (Avg)
Computer Science	881	78 516	104
Physics	803	61 394	104
Psychology	937	70 399	84
Political Science	837	38038	52

Experiments

2. American Online Logs for Personalized Search (AOL4PS):

- large-scale dataset constructed from the AOL logs
- Each query record has an associated user id, timestamp, session information, URLs of the top ten retrieved documents, and the index position of the clicked document.

TABLE I
AOL4PS Statistics

Metric	Value
Total number of records	1,339,101
Number of users	12,907
Average number of records per user	103.75
Unique records per user (mean)	47.23

Experiments

2. American Online Logs for Personalized Search (AOL4PS):

- A significant portion of the URLs included in the dataset were no longer available, document content could not be parsed → many queries had missing candidate documents.
- These queries could not be used for profile creation. Preprocessing was done to remove such queries and users for which, consequently, lacked sufficient valid queries.
- The remaining users were categorized into buckets based on the amount of past queries available for profile creation, allowing users with varying length of profiles to be tested.

TABLE IV
USER BUCKETS AND ASSOCIATED PROFILE LENGTHS

Bucket	Tested Profile Length
10-15	5
16-25	10
26-35	20
36-45	30
46 and above	40

TABLE V
USER BUCKET STATISTICS

Bucket	# Users	# Train Records	# Test Records
10-15	56	280	295
16-25	157	1570	858
26-35	43	860	229
36-45	18	540	94
46 and above	75	3000	569

Results: PRRB

Baselines:

- **BM25:** Ranks documents using tf-idf scoring with length normalization.
- **ColBERT-PRF:** Expands queries by clustering ColBERT embeddings from feedback docs and selecting high-idf centroid terms.
- **Query Expansion with Email Search (QEES):** Expands queries by selecting user terms with highest softmax-weighted similarity to query terms.
- **Query Expansion with Enriched User Profiles (QEEUP):** Expands queries using user terms most similar to the summed query embedding.
- **Personalized Query Expansion with Contextual Word Embeddings (PQEWC):** Uses HDBSCAN clustering to select user terms for query expansion.

TABLE VII
EXPERIMENTAL RESULTS ON PRRB.

Model	Computer Science		Physics		Psychology		Political Science	
	MAP@100	MRR@10	MAP@100	MRR@10	MAP@100	MRR@10	MAP@100	MRR@10
BM25	0.1511	0.4826	0.1295	0.5551	0.2122	0.6297	0.1713	0.5430
ColBERT-PRF	0.1856	0.5682	0.1877	0.6150	0.2192	0.6253	0.1642	0.5351
QEES	0.1813	0.5632	0.1783	0.6118	0.2142	0.6285	0.1598	0.5305
QEEUP	0.1818	0.5686	0.1805	0.6256	0.2137	0.6276	0.1549	0.5285
PQEWC	0.1903	0.5766	0.1917	0.6381	0.2230	0.6421	0.1724	0.5510
Ours	**0.2026**	**0.5871**	**0.1919**	**0.6495**	**0.2278**	**0.6493**	**0.1840**	**0.5694**

Results: AOL4PS

Previously created user buckets used, tested with:

- **BM25**
- **ColBERT Non Personalized:** ColBERT embedding representations with original query for reranking
- **Personalization Approach:** ColBERT embedding representations with user profile for reranking
- **Recency-Frequency Approach:** ColBERT embedding representations with user profile with time and frequency information for reranking

TABLE VIII
EXPERIMENTAL RESULTS ON AOL4PS.

Buckets	BM25		ColBERT Non Personalized		Personalization Approach		Recency-Frequency	
	MRR@10	MAP@1	MRR@10	MAP@1	MRR@10	MAP@1	MRR@10	MAP@1
10-15	0.3311	0.1559	0.3671	0.1322	**0.5723**	0.2780	0.5285	**0.2848**
16-25	0.3249	0.1480	0.3789	0.1317	**0.5822**	**0.2984**	0.5304	0.2879
26-35	0.3188	0.1222	0.3906	0.1354	**0.5921**	**0.3188**	0.5186	0.2751
36-45	0.4160	0.2553	0.3425	0.0957	**0.6447**	**0.3404**	0.5111	0.2660
46-above	0.3452	0.1706	0.3684	0.1255	**0.6463**	**0.3667**	0.5102	0.2647

Results: AOL4PS

- To judge the impact of increasing profile length, the personalization approach was tested further with different profile lengths.
- For each bucket, the profile lengths of lower buckets were tested along with the original associated length.

TABLE IV
USER BUCKETS AND ASSOCIATED PROFILE LENGTHS

Bucket	**Tested Profile Length**
10-15	5
16-25	10
26-35	20
36-45	30
46 and above	40

TABLE IX
TESTING WITH DIFFERENT PROFILE LENGTHS.

Buckets	**5**		**10**		**20**		**30**		**40**	
	MRR@10	MAP@1	MRR@10	MAP@1	MRR@10	MAP@1	MRR@10	MAP@1	MRR@10	MAP@1
10-15	**0.5723**	**0.2780**	-	-	-	-	-	-	-	-
16-25	0.4433	0.0991	**0.5822**	**0.2984**	-	-	-	-	-	-
26-35	0.4203	0.1004	0.4438	0.1223	**0.5921**	**0.3188**	-	-	-	-
36-45	0.4247	0.1064	0.4404	0.1383	0.4679	0.1596	**0.6447**	**0.3404**	-	-
46-above	0.4718	0.1176	0.4735	0.1255	0.4877	0.1353	0.4918	0.1294	**0.6463**	**0.3667**

Discussion

- The personalized with ColBERT approach consistently outperforms the baseline in each test setup.
- As observed in the testing of AOL4PS sample, increasing the length of the profile increases MRR and MAP scores, indicating better performance with greater user data.
- The recency-frequency approach offers improvements against BM25 and ColBERT Non-Personalized approach. However, it does not surpass simple ColBERT personalization.
- It can be noted that the recency-frequency approach performs better with smaller user profiles, which may imply that, with sufficient user data, the need to integrate recency and frequency information diminishes.

Conclusion

- We presented a novel PIR approach that encodes entire user profiles using contextual word embeddings and re-ranks BM25 retrieved documents.
- Additionally, we tested a frequency-recency weighting mechanism to study the impact of temporal proximity and repetition on personalization performance.
- Our results confirm the effectiveness of our approach is confirmed. ColBERT-based personalization with entire user profiles is reinforced as an effective approach.
- The recency-frequency approach offers improvements relative to the baseline, however it benefits users with limited search history more.

Limitations and Future Work

Limitations:

- Our testing with AOL4PS involved a small user sample due to limited valid URLs.
- Session information not considered in profile generation or reranking.

Future Work:

- Testing on a larger sample of users for AOL4PS - through increased URL scraping from different geographic locations.
- Improvements in time and memory considerations of responses.