Retrieving Webpages Using Online Discussions

Kevin Ros
kjros2@illinois.edu
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Matthew Jin
mjin11@illinois.edu
Department of Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

Jacob Levine
jlevine4@illinois.edu
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

ChengXiang Zhai
czhai@illinois.edu
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, Illinois, USA

ABSTRACT

Online discussions are a ubiquitous aspect of everyday life. An Internet user who interacts with an online discussion may benefit from seeing hyperlinks to webpages relevant to the discussion because the relevant webpages can provide added context, act as citations for background sources, or condense information so that conversations can proceed seamlessly at a high level. In this paper, we propose and study a new task of retrieving relevant webpages given an online discussion. We frame the task as a novel retrieval problem where we treat a sequence of comments in an online discussion as a query and use such a query to retrieve relevant webpages. We construct a new data set using Reddit, an online discussion forum, to study this new problem. We explore and evaluate multiple representative retrieval methods to examine their effectiveness for solving this new problem. We also propose to leverage the comments that contain hyperlinks as training data to enable supervised learning and further improve retrieval performance. We find that results using modern retrieval methods are promising and that leveraging comments with hyperlinks as training data can further improve performance. We release our data set and code to enable additional research in this direction.

CCS CONCEPTS

• **Information systems** → **Web and social media search**.

KEYWORDS

information retrieval; discussion forums; hyperlink prediction

ACM Reference Format:
Kevin Ros, Matthew Jin, Jacob Levine, and ChengXiang Zhai. 2023. Retrieving Webpages Using Online Discussions. In *Proceedings of the 2023 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '23), July 23, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3578337.3605139

ICTIR '23, July 23, 2023, Taipei, Taiwan

ACM ISBN 979-8-4007-0073-6/23/07…$15.00
https://doi.org/10.1145/3578337.3605139

1 INTRODUCTION

Online discussion forums are a ubiquitous aspect of everyday life. It is estimated that 55% of Americans have posted a comment online and that 78% of Americans have read online comments [36]. Popular social media platforms, such as Reddit, facilitate discussion forum commenting and browsing for millions of daily users [7]. Discussion forums are also crucial for online education. Organizations such as Khan Academy [22] and Campuswire [5] provide forums for students to discuss lecture content. And with the recent shift towards online learning due to the COVID-19 pandemic, many traditional classrooms have leveraged online discussion forums to support asynchronous student engagement.

In online discussion forums, hyperlinks in comments play an important role of providing context for the discussions. Much like a citation in a research paper, a hyperlink to an external webpage can provide a source for justifying a claim, which is crucial for combating misinformation and referencing different perspectives. Additionally, a hyperlinked webpage may help prevent confirmation bias, as readers of a conversation can simply follow the hyperlink to the webpage for a deeper understanding of the discussed topic. Moreover, a hyperlinked webpage can help condense background content so that discussions can take place at a higher level, thus alleviating the need to discuss redundant information.

Despite this importance, there is no clear standard for when a discussion forum user should add a hyperlink to a webpage in an online comment. As a result, valuable context may be left out of discussions. This may impact discussion participants, who may misunderstand the original comment due to the lack of context. And this may negatively impact readers with different levels of background knowledge, as they may miss crucial parts of the conversation. Although users generally have the ability to add hyperlinks to their posted comments, this does not happen very often. In fact, of the 83 million comments posted on Reddit during September of 2017, only 4.5 million comments contained a hyperlink.[1]

To see an example of a scenario where having a hyperlink would be useful, consider the comments depicted the left half of Figure 1. The discussion is centered around various characteristics of New Jersey. In the last comment, a URL was added by the user to provide a deeper context for the claim about New Jersey superfund sites. Clearly, this link provides utility to those who read the comment

[1] From our analysis.

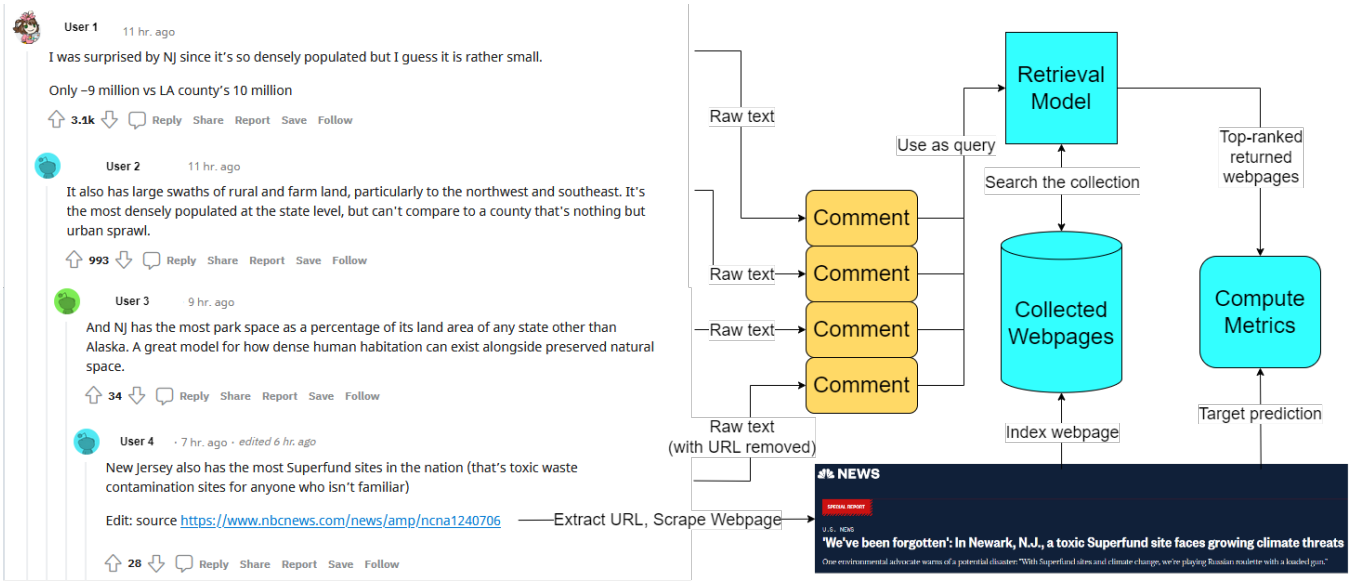

Figure 1: An overview of our problem setting. The left half of this image shows a typical comment thread on Reddit, and the right half of this image shows how we use the comment thread in our retrieval models. We remove the URL and collect the URL's webpage text. We then index all collected webpages into a corpus, and use the comment chain (without the URL) as a query to retrieve the previously-hyperlinked webpage.

and are interested in learning more about the claim, as they do not need to independently search for this information. Note, however, that the three prior comments do not contain hyperlinks. If a reader wanted to learn more about the park space to land area ratio of New Jersey (or to simply verify the claim), then the reader would need to open a new tab, overcome any domain-specific lack of knowledge, translate the general information need into a concrete search engine query, and browse possibly-related results for more information. Not only does the burden of finding this information fall on the reader, but it also must be repeated for every reader who is interested in the topic. In addition to the third comment, the first and second comment would also benefit from additional context regarding population density and urban sprawl, respectively. In summary, we suspect that, although some comments contain a hyperlink, there are many millions more (like those depicted in Figure 1) which can benefit from additional context. This, combined with the previously discussed benefits of hyperlinks in comments, motivates the need for further research.

Therefore, we study the task of automatically retrieving relevant webpages using online discussions. We frame this task as a problem of information retrieval: given a comment thread, the goal is to search over a collection of webpages using the thread as a query and return the webpage that is most likely to be useful for forum users who are interacting with the sequence of comments. For this task, the research challenges include novel data set construction, query formulation, and respective performance analysis on various representative baseline retrieval models. We study all of these aspects in this paper.

As depicted in Figure 1, we construct a new data set using Reddit, a popular online discussion forum. Concretely, we treat a comment chain (with the removed hyperlink) as a query and its respective hyperlinked webpage as the ground-truth retrieval target. We then attempt to retrieve the webpage using the comment chain from a large corpus of collected webpages. With this constructed data set, we explore and evaluate multiple baseline representative retrieval methods to measure their effectiveness in our problem setting. We also leverage this collected data as training data and explore the performance of neural retrieval methods fine-tuned on our collected data set. We make the following contributions:

(1) We propose and formalize the novel problem of retrieving webpages to construct hyperlinks using online discussion forum conversations.
(2) We collect, examine, and release a new data set that simulates relevance measures of hyperlinked webpages to discussion forum conversations.
(3) We establish the first benchmark for this new task by quantitatively and qualitatively evaluating multiple representative retrieval models on the collected data set using various contextual query settings that correspond to different real-world application scenarios.
(4) We show that the naturally available user-created hyperlinks can be leveraged as "free" training data to further improve performance.

Our code and data set are publicly available online.[2]

2 PROBLEM DEFINITION

We study how to predict the relevance of a webpage to a comment thread, and we frame this task as one of information retrieval.

[2]https://github.com/kevinros/contextAndConnections

Formulated as a retrieval problem, we consider a comment thread to be a query $q = \{c_1 \rightarrow c_2 \rightarrow \cdots \rightarrow c_n\}$ where comment c_i is a reply to comment c_{i-1}. Search happens over a collection of webpages $W = \{w_1, w_2, \ldots, w_m\}$. If there exists a webpage $w_h \in W$ that is relevant to comment c_n, then the goal is to retrieve webpage w_h from corpus W given comment-chain query q. A relevant webpage is one that can provide useful information to users who are interacting with the comment chain.

Because the query in our retrieval setting is unconventional, a highly interesting and novel research question is how to represent the query. Thus, we frame our investigation around various types of query representations by comparing how different representations of query q affect retrieval performance. We study three different contextual settings:

(1) **Full**: given all comments in the chain, retrieve the webpage relevant to the last comment. Formally, given query $\{c_1 \rightarrow c_2 \rightarrow \cdots \rightarrow c_n\}$, retrieve webpage w_h.
(2) **Last**: given the last comment, retrieve the webpage relevant to this comment. Formally, given query $\{c_n\}$, retrieve webpage w_h.
(3) **Proactive**: given all comments in the chain except for the last comment, retrieve the webpage relevant to the removed comment. Formally, given query $\{c_1 \rightarrow c_2 \rightarrow \cdots \rightarrow c_{n-1}\}$, retrieve webpage w_h.

We make a distinction among these three settings to better understand the context needed for effective retrieval. The **Full** setting provides the most informative context, but retrieval models may not be able to separate out the earlier, potentially unrelated comments. The **Last** setting offers the most specific context but at the potential expense of useful information (e.g., if the last comment is very short). And the **Proactive** setting represents the case where the information need is anticipated from the prior context.

Additionally, we make these distinctions to simulate different real-world application scenarios with variable amounts of query information. Specifically, the **Full** setting could simulate the automated addition of citations of relevant hyperlinks to an existing discussion thread to facilitate readers in understanding the context of the discussion. The **Last** setting could simulate the recommendation of a relevant hyperlink to cite when a user is composing a comment (the last comment) or arriving at the comment from a search engine. And the **Proactive** setting could simulate the automated generation of a brief comment with references to a relevant hyperlink to continue an existing discussion.

3 BUILDING THE DATA SET

As there was no available data set for evaluating the proposed new task, our first challenge was to create a data set for evaluation. Ideally, we would have had real users evaluate the usefulness of a recommended webpage, but that would have been labor-intensive and would have prevented us from performing experiments at scale. To address this challenge, we leveraged the hyperlinks in existing conversations as an approximation for relevance judgments.

In the case where a user explicitly added a hyperlink to a comment, it is reasonable to assume that the user felt that the hyperlinked webpage was relevant to the conversation. This assumption enabled us to automatically evaluate our task without requiring manual relevance judgements. In other words, to simulate the notion of relevance, we define the relevance between a comment chain and a webpage to be when the last comment in the comment chain contains a hyperlink to the webpage. We recognize that our assumption of relevance is not perfect; different users may be looking for different webpages due to different goals, background knowledge, etc. However, limiting our assumption of relevance to the hyperlink present in a comment during evaluation still allows us to make meaningful comparisons of different retrieval algorithms.

We used Reddit as the basis for constructing a data set for our task as Reddit is one of the largest publicly-available discussion forums, and it contains numerous comments across varied content. Without loss of generality, we began with all of the comments from September of 2017 [30]. In total, there were approximately 83 million candidate comments.

3.1 Collecting the Webpages

For each comment, we checked to see if it contained a URL from a set of pre-defined domain names (e.g., en.wikipedia.org). A full list of the selected domain names is presented in the two "Domain" columns of Table 1. We defined this set by choosing the most frequent domain names present in the Reddit data set which were likely to have significant textual content at the respective webpages, such as Wikipedia pages or news articles. We explicitly ignored hyperlinks back to Reddit as our focus was on retrieving content from external sources. However, we acknowledge that this direction is interesting and encourage future work to compare this setting to retrieving content from external sources.

Following the discovery of a comment containing a URL from a valid domain, we applied a few additional filters. First, we ignored the comment if it was a root comment. This guaranteed that there would be some context, as many root comments tended to reply to the post title or description, which we ignored for this exploration. Second, we ignored comments that were posted by specific bots that re-comment non-mobile versions of URLs. Third, we ignored any URLs that ended with common non-text endings ("jpg", "png", "gif", and "pdf"). And after some preliminary analysis, we found that there were many duplicate non-mobile and mobile Wikipedia URLs, so we converted the latter into the former.

With this filtered URL candidate set, we then scraped the HTML from each respective webpage. The webpage scraping was done serially in a random order to avoid overloading any particular domain. Each request timed out after two seconds. If the request succeeded, then from the resulting HTML, we extracted all plain text content. Next, we applied some basic filtering to help ensure that the resulting cleaned text was of high quality. Specifically, we checked that the ratio of special to non-special characters was less than 0.2 and that the extracted text length was greater than 100 characters. We also performed a brief manual inspection of the 1,000 shortest extracted texts. If a URL failed to pass through any of the aforementioned filters, then we removed it and any associated comments from our candidate set. In total, our final webpage corpus consisted of 98,231 unique URLs. Table 1 lists the number of URLs per domain.

Domain	URL Count	Domain	URL Count
en.wikipedia.org	63,656	www.latimes.com	908
www.theguardian.com	4,935	www.slate.com	855
www.forbes.com	2,414	www.nbcnews.com	784
www.businessinsider.com	2,039	abcnews.go.com	720
www.cnn.com	2,031	www.wired.com	587
www.npr.org	1,965	www.pbs.org	577
www.dailymail.co.uk	1,866	www.investopedia.com	577
www.telegraph.co.uk	1,834	www.vox.com	541
www.independent.co.uk	1,640	www.theonion.com	382
www.huffingtonpost.com	1,616	www.foxnews.com	358
www.espn.com	1,486	www.thesun.co.uk	256
www.theatlantic.com	1,448	www.cnbc.com	103
www.bbc.com	1,438	insider.foxnews.com	36
www.bbc.co.uk	1,136	www.chicagotribune.com	11

Table 1: The domains and the respective webpage counts in the final constructed corpus. Domains were chosen by examining the overall domain frequency in comments and by the likelihood of the webpages containing textual content. In total, 98,231 unique URLs were selected.

3.2 Constructing the Queries and Relevance Judgments

For each comment containing a hyperlink which successfully passed the aforementioned filters, we removed the hyperlink (the URL and any associated markdown) from the comment and reconstructed the comment's ancestor chain to the root comment. The comment reconstruction was done via the *parent_id* field present in the Reddit data set. This entire comment path, from the root comment to the comment containing the (removed) hyperlink, is what we considered a single query q, as described in Section 2. For each query, the removed hyperlink's webpage was assumed to be the query's relevant webpage.

Overall, the construction process resulted in 158,997 queries. Figure 2 depicts three aggregate measures of the queries. The left-most subplot shows how the number of comments per query is skewed left. Note that there are queries longer than 10 comments, but they are not depicted in the figure (the decreasing trend continues). In total, across the 158,997 queries, there were 825,751 individual comments, including possible duplicates. The center subplot shows the distribution over the number of words per query, excluding any outliers. Note that we added the separator "<C>" between each comment so that our retrieval models could distinguish between individual comments. The right-most subplot shows the distribution over the number of words per comment, excluding any outliers. The center and right-most subplot indicate that the vast majority queries are longer than traditional search engine queries [31], which distinguishes our problem setting from a traditional retrieval setting. We truncated queries to the final 500 words, in order to avoid query tokenization limits of Lucene (see Section 4.1). Qualitatively, many of the collected queries (comments) discuss news and news-related events. This is due to the URL selection criteria described in Section 3.1, as many of the collected webpages are news articles.

Type	Training	Validation	Testing
Queries	128,404	15,344	15,249
Webpages	78,637	9,849	9,745

Table 2: The training, validation, and testing splits of the queries (comment chains) and respective webpages.

We chose a random 80%-10%-10% train-validation-test split. The random seed was set to 100. Table 2 describes the query and webpage counts per split. Having more queries than webpages in each split reflects that, in some cases, multiple queries were mapped to the same webpage. This appears to happen frequently on Reddit as different comment threads can reference the same URL. We do not believe that this query-webpage count imbalance affected our evaluation because each query, regardless of split, was searched over the entire webpage corpus (all 98,231 webpages). Within each of these splits, we further constructed three different query data sets, corresponding to the Full, Last, and Proactive settings described in Section 2. Each setting had the same training, validation, and testing query split.

4 METHODOLOGY

In this section, we cover the retrieval models used in our exploration. As this work is the first exploration of a new task, our goal was to establish the first comprehensive benchmark for this task to facilitate further development of more advanced methods. To this end, we systematically evaluated multiple representative retrieval models in each of the aforementioned query settings.

4.1 BM25 and BM25+RM3

We implemented BM25 [34] and BM25 with RM3 pseudo-feedback [1] via Pyserini [25], an open-source retrieval toolkit built on Lucene, for the initial baselines. We kept most of the default indexing settings (porter stemming, etc.) and added our own stopwords. The list of stopwords can be found in our repository. The parameters for

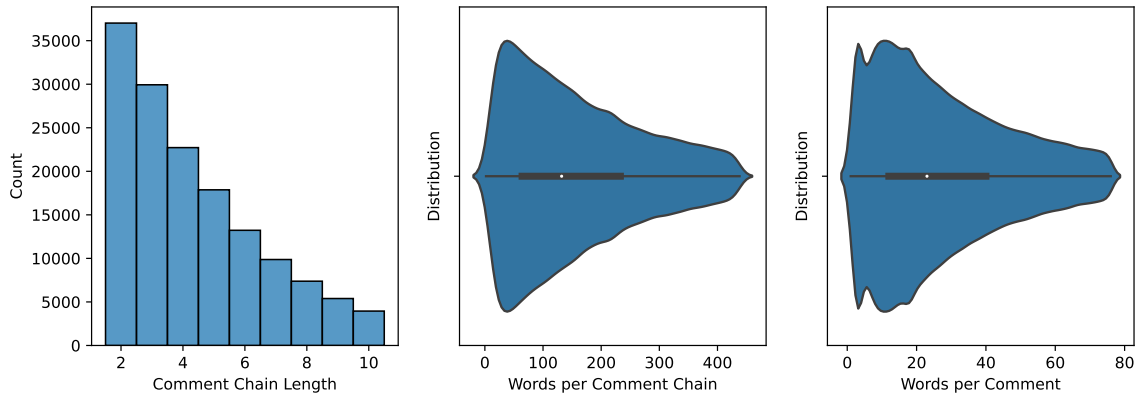

Figure 2: An overview of our collected queries, in terms of the comments. For the first plot, there were additional comment chains with length greater than 10, but they are not shown. The general decreasing trend continues. For the second and third plot, the comments were split by white space, and the resulting "words" were counted.

BM25 and BM25 with RM3 were manually tuned on the validation set.

4.2 The Contriever

The Contriever [15] is a transformer-based dense passage retriever that has been pretrained in an unsupervised fashion using the Inverse Cloze Task [21] and independent cropping. Intuitively, the Contriever learns to predict the information that is missing from a text-based context (e.g., a missing sentence from a paragraph). We selected the Contriever as a baseline because it shares similarities to our setting: in essence, our problem setting aims to retrieve the missing webpage from the discussion context.

Because the model has a token input limit, we reversed the comment chain so that, in the case of truncation, the oldest comments are removed first. This is based on the assumption that the comment containing the removed URL and its closest ancestors are more relevant to the webpage than the older ancestors. Moreover, we encoded the beginning tokens (up to the model length limit) of each webpage in our corpus. This was motivated by the empirical intuition that a significant amount of webpage content is present in the initial page tokens (e.g. titles, Wikipedia summaries, news article overviews, etc.).

4.3 Semantic Search

Next, we tested the effectiveness of an out-of-the-box semantic search neural model. Semantic search models embed both queries and webpages into a vector space, and they attempt to minimize the distance (or maximize the similarity) between a query and its relevant webpages while maximizing the distance (or minimizing the similarity) between a query and its irrelevant webpages. We selected msmarco-distilbert-cos-v5, a pre-trained BERT-based Transformer model provided by Sentence-BERT [32] that has been fine-tuned on the MS MARCO data set [27]. We selected this model due to its strong performance on the MS MARCO data set and due to its small, distilled size. We encoded queries and webpages as described in Section 4.2.

4.4 Semantic Search with Hyperlink Data

Our problem setting is quite different from that which the msmarco-distilbert-cos-v5 model was originally trained for, mostly due to the differences between Reddit comment chains and search engine queries. Therefore, we also fine-tuned msmarco-distilbert-cos-v5 on our collected data set in each problem setting. For our training loss function, we selected multiple negative ranking loss [14] with cosine similarity as the utility function. Each query was paired with one negative example, which was chosen as the highest-ranked non-relevant webpage returned by the best-performing BM25 run. If no webpage or only the relevant webpage was returned, then we randomly selected the negative sample. We encoded queries and webpages as described in Section 4.2.

5 EVALUATION

5.1 Evaluation Metrics

All models are trained and evaluated in each of the three settings using the precision at one (P@1) and the mean reciprocal rank at 10 (MRR@10) on the validation data set. Because some distinct webpages may be highly similar to a target (relevant) webpage and such similar webpages may also be regarded as relevant to some extent, we further examined the models' performances with respect to a cluster version of MRR@10, called "C-MRR@10", where "C" stands for "Cluster". Specifically, we computed the TF-IDF score vector for every webpage and mapped each webpage to its cluster (all other webpages with a similarity of 0.95 or greater). We considered a returned webpage relevant to a query if the returned webpage was the original, target webpage or in the query's original webpage's cluster. This reduced the confounding effects of similar webpages (which we discovered during the qualitative analysis, discussed in Section 5.3).

5.2 Parameters and Hyperparameters

5.2.1 BM25. The two parameters for BM25, k_1 and b, were selected based on the performance on the validation set. It was clear early on that larger values of both parameters offered better performance in the Full and Proactive settings, likely due to the long lengths of the queries. Thus, we tested variations of $k_1 \in \{2, 3, 4, 5, 6, 7, 8, 9\}$ and $b \in \{0.5, 0.75, 0.9, 0.99\}$. The best-performing parameters for the Full setting were $k_1 = 8$ and $b = 0.99$, for the Last setting, $k_1 = 4$ and $b = 0.9$, and for the Proactive setting, $k_1 = 7$ and $b = 0.99$. Each run took approximately 0.5-2 hours.

5.2.2 BM25 + RM3. For each setting, we used the best-performing BM25 parameters. The RM3 addition added three parameters: the original query weight oqw, the number of feedback documents fb_{doc}, and the number of feedback terms fb_{term}. During our experiments, we quickly found that the best-performing parameters minimized the effects of feedback. Thus, we tested variations of $oqw \in \{0.5, 0.8, 0.9\}$, $fb_{doc} \in \{1, 3, 10\}$, and $fb_{term} \in \{5, 10, 20\}$. The best-performing parameters in all settings were $oqw = 0.9$, $fb_{doc} = 1$, and $fb_{term} = 10$. Each run took between 0.5-2 hours.

5.2.3 Semantic Search and Contriever. There were no hyperparameters to tune for the Semantic Search and Contriever models. Encoding all websites and queries for each setting took approximately one hour, and inference took approximately five minutes.

5.2.4 Semantic Search with Hyperlink Data. Training per setting took approximately five hours on a single NVIDIA RTX A5000 GPU. We selected a batch size of 20 and trained for five epochs. Each model was evaluated on the validation set every 3,500 steps (twice per epoch: once approximately halfway through the epoch, and once at the end of the epoch), and the best-performing model was saved. We mostly used the default hyperparameters from when the model was trained on the MS MARCO data set, and all of our settings used the same hyperparameters. The maximum sequence length was set to 512, mean pooling was used, the optimizer was set to Adam [18], the number of warm-up steps was set to 300, the learning rate was set to $2e^{-5}$, and a single negative example per query was used, which was determined by the best-performing BM25 run.

5.3 Quantitative Analysis

Table 3 contains the quantitative results of our experiments. The first column, titled "Setting", delineates each of our query settings, and the second column, titled "Model", contains the respective models per setting. Columns three and four denote the performance of each model on the validation and test set. We now describe the main findings, which are in bold font below.

The setting choice played a large role in the performance of the models. The models in the Full and Last settings had similar scores, with the best-performing run resulting from interpolating the BM25 and Semantic Search Hyperlink runs. The Proactive setting had the lowest scores for every model, with the best-performing run also resulting from the interpolation. The lower scores in the Proactive setting than those in the other two settings were expected due to the larger semantic gap between the query and the relevant webpage(s) in the Proactive setting. In other words, the topics in comment chain discussions may have progressed enough so that retrieval models are generally unable to anticipate topic direction changes. These scores suggest that the Proactive setting is a quite challenging retrieval task with room for further research.

Across all settings, with respect to the other methods, BM25 was a strong baseline. Similar to Semantic Search, BM25 also performed better in the Last setting compared to the Full setting. We hypothesize that this was because the last comment (which originally contained the URL) offered a more focused representation of what the correct webpage should be. Note that BM25 + RM3 performed worse than BM25 in all settings. We believe that this was due to having only one relevance judgement per query, as the RM3 parameter selections tended to favor minimizing the effects of the feedback weighting (i.e., favoring original query terms).

Through fine-tuning, the Semantic Search Hyperlink model was able to effectively use hyperlinked webpages as training data. The Semantic Search Hyperlink model outperformed many of the non-interpolated methods in each setting. This is encouraging for future work because of the abundance of hyperlink data. Moreover, the Semantic Search Hyperlink model performed best in the Full setting, indicating that it was able to effectively use the entire comment chain as a query. Interestingly, the Semantic Search model performed slightly better in the Last setting compared to the Full setting. We attribute the lower performance of Semantic Search in the Full setting to the structural differences between MS MARCO queries and our comment-based queries. The Contriever performed worse than the Semantic Search. This is slightly surprising, as one may expect the pre-training setting of the Contriever to better match the problem settings proposed in this paper.

The interpolation of BM25 and Semantic Search Hyperlink performed best. Due to the strong performance of BM25 and Semantic Search Hyperlink as well as their complementary nature, we hypothesized that it may be beneficial to combine them. We thus interpolated BM25 with Semantic Search Hyperlink as follows. The normalized BM25 run scores were multiplied by α and the normalized Semantic Search with Hyperlink Data scores were multiplied by $1 - \alpha$, where $\alpha \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$. The resulting scores were then summed per returned webpage. Note that only the top 10 returned webpages for each run were interpolated. For the Full setting, $\alpha = 0.2$, for the Last setting, $\alpha = 0.5$, and for the Proactive setting, $\alpha = 0.8$. The results of such an interpolation run are shown in the table as "Interpolated". We see that our hypothesis is supported by the results because the interpolation was the best-performing run in all settings. Overall, the performance indicates that existing retrieval models in the Full or Last settings would be reasonably effective in a practical application, and that interpolated methods are likely to perform best. The performance in the Proactive setting, however, provides strong evidence against any practical applications in this setting.

Accounting for similar webpages increases performance. This can be seen via the "C-MRR@10" column for each of the validation and test runs. Interestingly, the best-performing individual model in this case was BM25. However, this may be due to the TF-IDF clustering methodology choice. Future work may consider explicitly optimizing for such a measure.

There was not a large difference in performance distribution over Wikipedia versus non-Wikipedia domains. Our collected webpage corpus is skewed towards the Wikipedia domain.

Setting	Model	Validation			Test		
		P@1	MRR@10	C-MRR@10	P@1	MRR@10	C-MRR@10
Full	BM25	0.2010	0.2762	0.2946	0.1944	0.2709	0.2902
	BM25 + RM3	0.2000	0.2597	0.2771	0.1935	0.2547	0.2730
	Contriever	0.1206	0.1743	0.1872	0.1210	0.1744	0.1884
	Semantic Search	0.1636	0.2212	0.2333	0.1643	0.2214	0.2327
	Semantic Search Hyperlink	0.2430	0.3121	0.3318	0.2387	0.3061	0.3254
	Interpolated	**0.2752**	**0.3409**	**0.3668**	**0.2662**	**0.3321**	**0.3594**
Last	BM25	0.2292	0.3027	0.3489	0.2218	0.2968	0.3437
	BM25 + RM3	0.2275	0.2910	0.3154	0.2206	0.2864	0.3127
	Contriever	0.1541	0.2135	0.2312	0.1531	0.2141	0.2320
	Semantic Search	0.1712	0.2252	0.2383	0.1774	0.2294	0.2409
	Semantic Search Hyperlink	0.2158	0.2752	0.2911	0.2163	0.2746	0.2906
	Interpolated	**0.2566**	**0.3372**	**0.3656**	**0.2547**	**0.3376**	**0.3687**
Proactive	BM25	0.0840	0.1262	0.1350	0.0819	0.1223	0.1326
	BM25 + RM3	0.0830	0.1192	0.1275	0.0816	0.1168	0.1246
	Contriever	0.0466	0.0733	0.0781	0.0509	0.0774	0.0831
	Semantic Search	0.0676	0.0967	0.1020	0.0665	0.0967	0.1026
	Semantic Search Hyperlink	0.0597	0.0927	0.0999	0.0589	0.0909	0.0993
	Interpolated	**0.0945**	**0.1355**	**0.1451**	**0.0939**	**0.1345**	**0.1442**

Table 3: Columns three and four list the performance of each model on our validation set and test set, respectively. The Full setting and the Last setting offer a clear contextual advantage to the Proactive setting. We discuss the statistical significance in the last paragraph of Section 5.3.

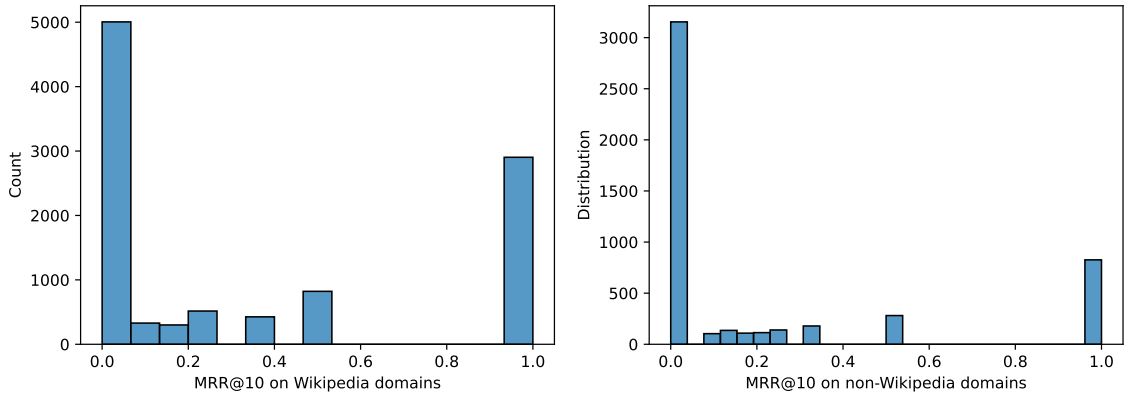

Figure 3: MRR@10 on Wikipedia and non-Wikipedia domains for the Semantic Search Hyperlink model in the Full setting. The performance distributions appear to be bimodal, with a large spike near zero and near one.

Naturally, we would like to ensure that our performance measurements are not heavily affected by this skew. In Figure 3, we show the MRR@10 over Wikipedia and non-Wikipedia domains for the Semantic Search Hyperlink model in the Full setting. The distributions appear to be similar, as both share large concentrations of performance near zero and near one.

The best-performing run in each setting was statistically significant at $p = 0.0001$ with respect to the other runs in the setting. Moreover, each model was significantly different across settings at $p = 0.0001$, except for the Semantic Search model in the Full and Last setting ($p = 0.0986$), and the Interpolated BM25 + Semantic Search Hyperlink model in the Full and Last setting ($p = 0.2371$). To test the significance of our findings, we used a relative t-test via the Scipy Python package [38] on the MRR@10.

5.4 Qualitative and Error Analysis

We performed some additional qualitative and error analysis on queries from the validation set. First, we examined the set of URLs in the Full setting where the relevant webpage was not returned in the top 10 results by both BM25 and Semantic Search Hyperlink, thus representing the most difficult queries. For these queries, we observed many cases where the relevant webpage was very similar to the returned webpages. For example, a discussion thread about Donald Trump claiming he was wiretapped returned various news

articles reporting on this claim. The webpage relevant to this discussion was about Susan Rice trying to "unmask" Trump associates in relation to the wiretap, whereas the articles returned by the retrieval models were about other events surrounding this wiretap (e.g., the Justice Department's report of no evidence surrounding previous wiretappings of Trump tower). In another example, the conversation was centered around the Black Lives Matter movement, police brutality, and individuals who have been killed by the police. The ground truth webpage was a Wikipedia article about one of these individuals. The Semantic Search Hyperlink model, although missing the ground truth, still returned multiple Wikipedia pages of individuals killed by police. These finding suggest that a topic-based or timeline-based coalescing of webpages may be helpful for providing relevant results.

After examining the Proactive setting, we found, for many cases where the retrieval models did not return a relevant webpage, that the returned webpages were topically similar to the parent comments themselves. In other words, the models had trouble "anticipating" information needs. This issue was likely magnified when the last comment (which was removed) introduced significant or unpredictable changes in topic.

Finally, we noticed some general trends that could be addressed with more robust preprocessing. There were instances where different URLs pointed to the same content. For example, a Wikipedia URL may have a fragment that resolved at a particular subsection of the article, or a URL may redirect to another webpage. As a result, some queries might have artificially lower performance metrics, as each query was only assigned one ground truth. This motivated us to use the "C-MRR@10" measure, as it accounts for all cases in evaluation where the different URLs point to the same or similar webpages. Future work may consider constructing more fine-grained target webpages which only contain the specific subsection referenced by the URL fragment.

6 RELATED WORK

6.1 Proactive Search

In a traditional information retrieval setting, an explicit user query is used to predict relevant content. On the other hand, in the setting of proactive or anticipatory search, pre-search context is used to predict relevant content [10, 24, 33]. The notion of pre-search context is extremely broad; it can include explicitly relevant documents [10], global search history logs [24], short-term personal logs [37], previously read news articles [19], the current written text [20, 26], time and location [41], or conversations [3, 29].

Conversations themselves can offer rich contextual clues that can be used to anticipate the information needs of the participants. For example, Twitter user often explicitly and implicitly express their needs about mobile applications via "user status text" updates [29]. The authors framed this need as a retrieval problem where the query is the status update and attempted to retrieve relevant mobile applications. Specifically, they leverage a collection of explicit intentions to predict the true intention using an implicit intention. More generally, extracting entities from spoken conversations for proactively retrieving relevant information has also been studied [3]. The authors found that their proactive search system helped conversation participants fact-check discussion points with low mental effort and that the returned results tended to influence the conversation. Proactive search can also be applied to various domains, such as argument retrieval [35]. Finally, similar to our work is the idea of retrieving sentences for the next turn in an open-ended dialogue [13]. Here, the authors use the prior turns as a query to retrieve sentences such that these sentences can be used to generate the next dialogue turn. This is most similar to our Proactive setting. However, our primary goal is to retrieve webpages for adding context to a conversation for a user, rather than as input for a generative language model.

More generally, our work is similar to the existing work insofar as we use a conversation as pre-search context. This is related to conversational retrieval tasks such as TREC CAsT (Conversational Assistance Track) [28]. In these tasks, the user directly interacts with the retrieval system through turn-based dialogues. In our setting, however, the retrieval system is "observing" the dialogue among individuals and proactively retrieving content according to the dialogue. We are unaware of any work that frames comment threads as queries and uses them to retrieve webpages relevant to the comment threads.

6.2 Citation and Hyperlink Recommendation

Hyperlinks added in online discussions can be thought of as informal citations to external content. There has been much work on examining citation recommendation, albeit in an academic setting [2]. A key insight to citation recommendation is the use of the context in which the citation is introduced [8]. Many recent content-based citation recommendation approaches have represented citation context via deep representations which are learned by GRUs [4], encoder-decoders (TDNN-RNN) with attention [9], LSTMs [40], or Transformers [16]. Work has also been done for recommending hyperlinks on Twitter [11]. Note that many of these approaches use user information (e.g., via previous publications, personal citation networks, or collaborative filtering) to recommend citations or hyperlinks. Moreover, many approaches leverage explicit networks or interaction graphs to recommend content.

Unlike many of the aforementioned approaches, we do not use any user representations in our framework, nor do we use any specific graph structure. One benefit of our retrieval approach is its generalization - webpages can be added to the candidate corpus without any prior user interactions, thus helping mitigate the cold-start problem. And with respect to citation recommendation, another difference is that scientific literature articles are well-written with rich and semantically coherent content, whereas the comments in a forum are informal, short, and not always coherent, adding additional challenges for retrieval. Additionally, we do not restrict ourselves to scientific citations, but rather we collect and retrieve general hyperlinked webpages.

6.3 News Recommendation

There has been prior work on news recommendation [17, 39]. Most of this work is focused on modeling a user's need and generating personalized recommendations. While our problem setup is similar to recommendation, we go beyond general news recommendation to retrieve general webpages with a focus on the context of online forum discussions. The closest work in news recommendation to

ours is the use of forum discussions for recommendation [23], in which discussion forum comments are used as additional features for general recommendation of news articles. In addition to recommending not just news, our work also studies different ways to construct a query, and leverages cited hyperlinks for both scalable evaluation without requiring manual relevance judgments.

7 CONCLUSION

We have introduced a novel task of retrieving relevant webpages using online discussions and framed the problem as a novel retrieval problem with various ways to construct a query, including using the full comment thread, only the last comment, and the full comment thread excluding the last comment. We created a new data set based on Reddit by leveraging cited hyperlinked webpages in user comments as simulated relevance judgments to enable quantitative evaluation of the task. We studied the effectiveness of both representative state-of-the-art retrieval algorithms and a new idea of fine-tuning using the naturally available hyperlink citations in forums. The results show that the popular BM25 retrieval algorithm works well for this new task, but more advanced semantic search algorithms based on pre-trained language models can more effectively make use of the full comment threads to improve search results. The results also show that the proposed idea of fine-tuning with user-cited hyperlinked webpages works very well and enables neural algorithms to generally outperform traditional retrieval algorithms. This approach seems particularly promising given the abundance of hyperlinks on the Internet and the potential to use larger language models. The best performance in all settings was achieved by combining BM25 with such a fine-tuned semantic search algorithm, suggesting that the traditional models and the new neural ranking models might have complementary benefits.

The positive results of using naturally available hyperlinked webpages for training are quite encouraging from application perspective as it means the algorithms can possibly increase its performance over time as we accumulate increasingly more cited hyperlinks. The proposed techniques can be used immediately to help build novel real-world applications that can help retrieve relevant context for online discussions.

8 LIMITATIONS AND FUTURE WORK

As a preliminary exploration of a new task, our work has the following limitations, which are also opportunities for future research.

8.1 Improvement of the Data Set

While our assumption of cited hyperlinked webpages being relevant is reasonable and enabled us to have meaningful comparison of different retrieval algorithms without requiring expensive human judgments, it is desirable to expand such a notion of a single ground truth to a cluster or collection of relevant webpages (e.g., news articles about the same topic), which may lead to more precise quantitative evaluation. How to automatically construct such as data set without (much) manual labeling is an interesting challenge for future research. Moreover, our corpus was constructed by only including webpages which were mentioned in at least one comment and we restricted ourselves to a single platform, with hand-selected domains. The majority of our collected webpages were from Wikipedia, a single domain. In a more realistic setting, the corpus would be much larger, diverse, and possibly contain webpages that were not mentioned in any comment. Even though the latter was somewhat approximated via our query search over all training, validation, and testing splits, it could be expanded by collecting more webpages (e.g., from the Common Crawl corpus [6]).

8.2 Improvement of Models

Model-wise, it is unclear how to handle comments longer than model input length limits, or how to effectively encode long webpages. Regarding the latter, we showed that a simple truncation-based encoding produced meaningful results. We suspect that a more robust solution, perhaps including webpage structure, will lead to significant improvement in the retrieval performance. Recent developments in improving context window length (e.g., [12]) may also be applicable. Moreover, the idea of leveraging cited hyperlinked webpages for supervised learning opens up additional opportunities of exploring even more effective fine-tuning methods in the future, such as graph-based citation approaches.

8.3 Multi-Modality

We restricted our exploration to a handful of text-based domains. Expanding to multi-modal queries and documents, although difficult, would create much more pragmatic and rich problem setting. For example, including PDFs and lecture videos may directly benefit those in an academic setting. Alternatively, examining hyperlinks which link back to Reddit may help learn relationships among various conversation threads. In a similar fashion, future work could expand the number of comments considered by collecting additional data from Reddit or from other domains.

8.4 User Studies

Any application of this work would certainly benefit from user studies, direct feedback, and result diversification, as there may be various nuances that differ from the laboratory setting that we used in our evaluation. Through user feedback, this analysis could be extended to evaluation over *any* online comment, rather than just those with a hyperlink.

8.5 Applications

Perhaps the most exciting areas for future work are the various applications of this research direction in domains beyond social media. For example, such an approach could be used to augment classroom discussion forms with relevant course material, or to provide additional context for employees using a company-based messaging service. Naturally, these domains would introduce additional research challenges, necessitating further empirical and user studies.

9 ACKNOWLEDGEMENTS

This work is supported in part by the IBM-Illinois Discovery Accelerator Institute and by the National Science Foundation under Grant No. 1801652.

REFERENCES

[1] Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. *Computer Science Department Faculty Publication Series* (2004), 189.

[2] Zafar Ali, Irfan Ullah, Amin Khan, Asim Ullah Jan, and Khan Muhammad. 2021. An overview and evaluation of citation recommendation models. *Scientometrics* 126, 5 (2021), 4083–4119.

[3] Salvatore Andolina, Valeria Orso, Hendrik Schneider, Khalil Klouche, Tuukka Ruotsalo, Luciano Gamberini, and Giulio Jacucci. 2018. Investigating proactive search support in conversations. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 1295–1307.

[4] Trapit Bansal, David Belanger, and Andrew McCallum. 2016. Ask the GRU: Multi-task learning for deep text recommendations. In *proceedings of the 10th ACM Conference on Recommender Systems*. 107–114.

[5] Campuswire. 2022. *Commenting on discussion posts.* http://web.archive.org/web/20221002030422/https://campuswire.com/chatrooms

[6] Common Crawl. 2022. *Common Crawl.* http://web.archive.org/web/20221014025949/https://commoncrawl.org/

[7] Brian Dean. 2021. *Reddit User and Growth Stats (Updated Oct 2021).* http://web.archive.org/web/20221005051600/https://backlinko.com/reddit-users

[8] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the association for information science and technology* 65, 9 (2014), 1820–1833.

[9] Travis Ebesu and Yi Fang. 2017. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. 1093–1096.

[10] Desmond Elliott and Joemon M Jose. 2009. A proactive personalised retrieval system. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 1935–1938.

[11] Dehong Gao, Renxian Zhang, Wenjie Li, and Yuexian Hou. 2012. Twitter hyperlink recommendation with user-tweet-hyperlink three-way clustering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2535–2538.

[12] Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916* (2021).

[13] Itay Harel, Hagai Taitelbaum, Idan Szpektor, and Oren Kurland. 2022. A Dataset for Sentence Retrieval for Open-Ended Dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2960–2969.

[14] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652* (2017).

[15] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118* (2021).

[16] Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics* 124, 3 (2020), 1907–1922.

[17] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems–Survey and roads ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Weize Kong, Rui Li, Jie Luo, Aston Zhang, Yi Chang, and James Allan. 2015. Predicting search intent based on pre-search context. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 503–512.

[20] Markus Koskela, Petri Luukkonen, Tuukka Ruotsalo, Mats Sjöberg, and Patrik Floréen. 2018. Proactive information retrieval by capturing search intent from primary task context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 3 (2018), 1–25.

[21] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300* (2019).

[22] Marcia Lee Lee and Kitt Hirasaki. 2012. *Commenting on discussion posts.* http://web.archive.org/web/20221206102201/https://blog.khanacademy.org/commenting-on-discussion-posts/

[23] Qing Li, Jia Wang, Yuanzhu Peter Chen, and Zhangxi Lin. 2010. User comments for news recommendation in forum-based social media. *Information Sciences* 180, 24 (2010), 4929–4939.

[24] Daniel J Liebling, Paul N Bennett, and Ryen W White. 2012. Anticipatory search: using context to initiate search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 1035–1036.

[25] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2356–2362.

[26] Petri Luukkonen, Markus Koskela, and Patrik Floréen. 2016. LSTM-based predictions for proactive information retrieval. *arXiv preprint arXiv:1606.06137* (2016).

[27] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

[28] Paul Owoicho, Jeffrey Dalton, Mohammad Aliannejadi, Leif Azzopardi, Johanne R Trippas, and Svitlana Vakulenko. 2023. TREC CAsT 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation. In *Proceedings of the NIST Text Retrieval Conference (TREC 2022). TREC'22*. 1–11.

[29] Dae Hoon Park, Yi Fang, Mengwen Liu, and ChengXiang Zhai. 2016. Mobile app retrieval for social media users via inference of implicit intent in social media text. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 959–968.

[30] PushShift.io. 2021. *PushShift.io.* https://files.pushshift.io/

[31] Jan Heinrich Reimer, Sebastian Schmidt, Maik Fröbe, Lukas Gienapp, Harrisen Scells, Benno Stein, Matthias Hagen, and Martin Potthast. 2023. The Archive Query Log: Mining Millions of Search Result Pages of Hundreds of Search Engines from 25 Years of Web Archives. *arXiv preprint arXiv:2304.00413* (2023).

[32] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

[33] Bradley James Rhodes and Pattie Maes. 2000. Just-in-time information retrieval agents. *IBM Systems journal* 39, 3.4 (2000), 685–704.

[34] Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*. Springer, 232–241.

[35] Kevin Ros, Carl Edwards, Heng Ji, and Cheng Xiang Zhai. 2021. Team Skeletor at Touché 2021: Argument Retrieval and Visualization for Controversial Questions. In *CEUR Workshop Proceedings*, Vol. 2936. CEUR-WS, 2441–2454.

[36] Natalie Jomini Stroud, Emily Van Duyn, and Cynthia Peacock. 2016. *Survey of Commenters and Comment Readers.* http://web.archive.org/web/20221129215636/https://mediaengagement.org/research/survey-of-commenters-and-comment-readers/

[37] Yury Ustinovskiy and Pavel Serdyukov. 2013. Personalization of web-search using short-term browsing context. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1979–1988.

[38] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[39] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2021. Personalized news recommendation: A survey. *arXiv preprint arXiv:2106.08934* (2021).

[40] Libin Yang, Yu Zheng, Xiaoyan Cai, Hang Dai, Dejun Mu, Lantian Guo, and Tao Dai. 2018. A LSTM based model for personalized context-aware citation recommendation. *IEEE access* 6 (2018), 59618–59627.

[41] Hamed Zamani, Michael Bendersky, Xuanhui Wang, and Mingyang Zhang. 2017. Situational context for ranking in personal search. In *Proceedings of the 26th International Conference on World Wide Web*. 1531–1540.