CEUR-WS.org/Vol-3366/paper-03.pdf

A Large-Scale Dataset for Known-Item Question Performance Prediction

Maik Fröbe[1,*], Eric Oliver Schmidt[2] and Matthias Hagen[1]

[1]*Friedrich-Schiller-Universität Jena, 07743 Jena, Germany*
[2]*Martin-Luther-Universität Halle-Wittenberg, 06120 Halle (Saale), Germany*

Abstract

Searchers who cannot resolve a known-item information need using a search engine (e.g., as the searcher only remembers vague details about a movie from some years ago) might post a respective question on a question answering platform, hoping that the discussion with other people can help to identify the item. To foster research on applying query performance prediction to known-item information needs, especially in the light of the upcoming tip-of-my-tongue known-item retrieval track at TREC 2023, we build a large-scale dataset of 1.28 million known-item questions (47 % have an identified answer) from the r/tipofmytongue subreddit. As the "performance" of a known-item question, we use the time it took the community to solve the question (or the absence of a solution) and evaluate the effectiveness of seven standard pre-retrieval query performance predictors in a pilot study. Not surprisingly, none of the tested predictors can really assess the performance of known-item questions.

Keywords

Query Performance Prediction, Known-Item Search, Re-Finding, Tip of my Tongue, Reddit

1. Introduction

Many queries submitted to search engines are navigational queries in the sense that searchers want to find a specific item (for the early days of web search, Broder [1] stated an amount of more than 20 %). Among the navigational queries, known-item and re-finding queries (i.e., accessing an item known to exist or even previously accessed), can be particularly challenging when a searcher is unable to recall a suitable identifier for the item [2, 3]. Actually, many people who post known-item questions on Reddit state that their previous attempts with traditional search engines were unsuccessful [4]. Still, known-item information needs posted on question answering platforms have hardly been part of retrieval system evaluations (e.g., at TREC tracks); the main focus usually are informational queries. Since also the previous work on query performance prediction (QPP) [5, 6, 7] used the standard retrieval collections, the effectiveness of query performance prediction on difficult known-item information needs (often expressed as rather verbose questions with explanations) has not been analyzed yet.

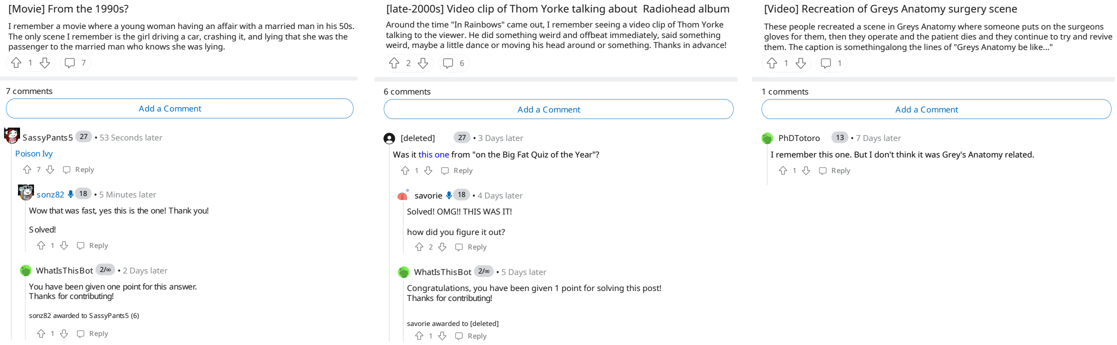
QPP++ 2023: Query Performance Prediction and Its Evaluation in New Tasks, co-located with The 45th European Conference on Information Retrieval (ECIR) April 2, 2023, Dublin, Ireland

maik.froebe@uni-jena.de (M. Fröbe); matthias.hagen@uni-jena.de (M. Hagen)
https://jena.webis.de/ (M. Fröbe); https://jena.webis.de/ (M. Hagen)
0000-0002-1003-981X (M. Fröbe); 0000-0002-9733-2890 (M. Hagen)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Figure 1: Examples of known-item questions: the left question is answered within a minute, the middle question within three days, and the rightmost question has not been answered after a week.



[Movie] From the 1990s?

I remember a movie where a young woman having an affair with a married man in his 50s. The only scene I remember is the girl driving a car, crashing it, and lying that she was the passenger to the married man who knows she was lying.

7 comments

SassyPants5 27 • 53 Seconds later

Poison Ivy

sonz82 18 • 5 Minutes later

Wow that was fast, yes this is the one! Thank you!

Solved!

WhatIsThisBot 2/∞ • 2 Days later

You have been given one point for this answer.
Thanks for contributing!

sonz82 awarded to SassyPants5 (6)

[late-2000s] Video clip of Thom Yorke talking about Radiohead album

Around the time "In Rainbows" came out, I remember seeing a video clip of Thom Yorke talking to the viewer. He did something weird and offbeat immediately, said something weird, maybe a little dance or moving his head around or something. Thanks in advance!

6 comments

[deleted] 27 • 3 Days later

Was it this one from "on the Big Fat Quiz of the Year"?

savorie 18 • 4 Days later

Solved! OMG!! THIS WAS IT!

how did you figure it out?

WhatIsThisBot 2/∞ • 5 Days later

Congratulations, you have been given 1 point for solving this post!
Thanks for contributing!

savorie awarded to [deleted]

[Video] Recreation of Greys Anatomy surgery scene

These people recreated a scene in Greys Anatomy where someone puts on the surgeons gloves for them, then they operate and the patient dies and they continue to try and revive them. The caption is somethingalong the lines of "Greys Anatomy be like..."

1 comments

PhDTotoro 13 • 7 Days later

I remember this one. But I don't think it was Grey's Anatomy related.

Retrieving answers to known-item questions will be the focus of the upcoming ToT track at TREC 2023.[1] The task can be very challenging since known-item questions often do not mention a suited identifier for the item or only a loosely similar one [8, 9], and since information in the question might even be wrong [10]. As training data, the organizers of the ToT track announced a dataset by Bhargav et al. [8]: 15,000 questions with linked known items (focus: movies and books) extracted from the subreddit r/tipofmytongue/.[2] Having run an adapted version of Bhargav et al.'s approach, our resulting TOMT-KIS dataset (tip-of-my-tongue known-item search) contains 1.28 million known-item questions (47 % with an identified answer) and is freely available under a permissive license.[3]

As an indicator for the "performance" of a known-item question, we simply use the time elapsed to answer the question. For example, the left question in Figure 1 was answered within a minute and thus performed better than the middle question (answered after three days) that again performed better than the rightmost question (not answered after a week). In a pilot study with seven standard pre-retrieval query performance predictors, we find that none of the predictors can really assess the performance of known-item questions.

2. A Large-Scale Dataset of Known-Item Questions

To create our TOMT-KIS dataset, we have crawled all questions and discussions from the tip-of-my-tongue subreddit. Note that answers are not explicitly tagged in the subreddit discussions, but the guidelines state that the asker should reply with "Solved!" to a post with the correct answer. In creating the dataset that the TREC 2023 ToT track organizers suggest as training data, Bhargav et al. [8] focused on questions for which the asker replied exactly with "Solved!" to some post and from these they only kept the questions where the answer contains exactly one link to a Wikipedia, IMDb, or GoodReads page. Using this rather restrictive approach, only 15,000 questions with a specified known-item answer were identified.

[1] https://trec-tot.github.io/
[2] https://www.reddit.com/r/tipofmytongue/
[3] Data and code for TOMT-KIS: https://github.com/webis-de/QPP-23

Table 1
Excerpt from TOMT-KIS for the examples from Figure 1. For questions a moderator tagged as solved, our precision-oriented heuristic extracted the post with the answer. The data is stored in JSONL format (127 attributes; excerpt of 6 attributes shown in tabular form for readability).

Question				Answer	
Title	Content	UTC	Solved	Content	UTC
[Movie] From the 1990s?	I remember a movie where a young woman having an affair with a married man [...]	1576446820	✓	[Poison Ivy] (https://www.imdb.com/title/tt0105156/)]	1576446873
[late-2000s] Video clip of Thom Yorke talking about Radiohead album	Around the time "In Rainbows" came out, I remember seeing a video clip of Thom Yorke [...]	1598985699	✓	Was it [this one] (https://m.youtube.com/watch?v=rrnNn0mt8h4) from [...]	1599242625
[Video] Recreation of Greys Anatomy surgery scene	These people recreated a scene in Greys Anatomy [...]	1479737186	✗	—	—

Answer Identification Analyzing different questions from the subreddit, we observed that askers often do not reply to the correct answer with "Solved!" but with other posts (e.g., "This was it!"), and we found a further metadata field (maintained by moderators but not used by Bhargav et al.) that indicates whether a question is solved. We thus adapted the answer identification method of Bhargav et al. to recall more questions with answers. For questions that the moderator metadata indicates as solved, we use four manually created lexical patterns as a rule-based answer identification method: if a post from the asker in reply to a post a contains 'solved', 'thank', 'yes', or 'amazing', we view the post a as the answer. On 50 random questions and 50 questions for which this heuristic identified an answer (none of them used to develop the rules), the achieved precision is 92 % at a recall of 78 %.

Dataset Description Our TOMT-KIS dataset is available in JSONL format. For each question, we include all the attributes available in the crawled data and add the chosen answer when our heuristic could extract it. Table 1 shows an excerpt for the questions from Figure 1.

Overall, TOMT-KIS includes 127 attributes for each question, such as `title`, `content`, `created_utc` (indicating the posted question's timestamp), and `link_flair_text` (indicates whether the question is solved; set by moderators). The complete tree of the discussion on each question is stored in the `comments` field. To simplify subsequent processing steps, we run our precision-oriented answer identification heuristic on questions tagged as solved by a moderator and add four "new" attributes when the heuristic could identify an answer: (1) `answer_detected` is a Boolean flag indicating whether our heuristic could extract an answer, (2) `solved_utc` specifies the timestamp when the identified answer was posted, (3) `chosen_answer` contains the extracted answer, and (4) `links_on_answer_path` contains all links to Reddit-external pages that were found in posts between the question and the post with the answer (this can be used for future retrieval experiments like retrieving known-item candidates from a web crawl [8]).

Dataset Analysis Many questions in the tip-of-the-tongue subreddit have assigned tags that roughly describe an assumed category of the known item. Table 2 shows the 60 most frequent tags in our TOMT-KIS dataset, excluding the [tomt] tag itself (all questions have it).

Table 2
The 60 most frequent tags in our TOMT-KIS dataset; we later merged similar ones to larger categories.

Rank	Tag	Count
1	[song]	161,385
2	[movie]	158,543
3	[video]	67,642
4	[music]	47,591
5	[book]	47,578
6	[2000s]	38,200
7	[game]	32,961
8	[2010s]	22,122
9	[tv show]	17,731
10	[music video]	17,434
11	[website]	14,140
12	[video game]	13,007
13	[cartoon]	10,774
14	[tv]	9,395
15	[youtube video]	8,107
16	[1990s]	7,784
17	[youtube]	7,636
18	[anime]	7,284
19	[show]	7,215
20	[word]	6,394
21	[meme]	5,789
22	[reddit post]	5,748
23	[comic]	5,503
24	[gif]	5,383
25	[image]	4,894
26	[film]	4,320
27	[short story]	4,287
28	[2000s]	4,282
29	[quote]	4,269
30	[pc game]	4,115
31	[picture]	4,095
32	[commercial]	4,012
33	[90s]	3,560
34	[videogame]	3,371
35	[2000s?]	3,355
36	[1980s]	3,028
37	[webcomic]	2,955
38	[story]	2,872
39	[subreddit]	2,782
40	[toy]	2,705
41	[early 2000s]	2,695
42	[2010s]	2,646
43	[flash game]	2,584
44	[animation]	2,189
45	[band]	2,175
46	[reddit]	2,088
47	[tv series]	2,069
48	[album]	2,022
49	[computer game]	2,012
50	[2010s?]	1,998
51	[2000's]	1,977
52	[manga]	1,939
53	[movie/tv]	1,934
54	[short film]	1,917
55	[2000s-2010s]	1,890
56	[youtube channel]	1,860
57	[documentary]	1,758
58	[tiktok]	1,713
59	[movies]	1,655
60	[2020]	1,623

Besides [tomt], a question has between 0 and 14 tags (average: 0.89), some of which directly express uncertainty (e.g., [2010s?] vs. [2010s]).

For further analyses, we manually merged the original tags to form larger categories (e.g., combining [song], [music], etc. to a "Music" category). Table 3 shows general statistics for all questions from our new TOMT-KIS dataset and for the four most popular merged categories. As a proxy for the "performance" of a known-item question, we use the time elapsed until the solution was posted (columns Δ_T; lower is better). Obviously, the performance varies between years and between categories. For instance, relatively more known-item questions in the movies category are solved than in the music category (52 % for movies vs. 46% for music) and the average time until an answer is posted is lower (9 hours for movies vs. 14 hours for music). Since query performance prediction is evaluated via measuring the correlation of a predictor's query ranking (by predicted performance) to the ground truth ranking [5, 7, 11], we build training, validation, and tests sets for the set of all questions and for the shown four most popular categories by each time sampling 100 questions for validation and 100 questions for test while keeping all other questions as training data.

Table 4 shows length statistics of the titles and contents of the questions, and of the identified answers in our TOMT-KIS dataset as the average number of characters and words (whitespace tokenization). Overall, the titles of questions are much shorter (14 words on average across all categories) than the explanations in the content field (81 words on average), while the identified answers again are rather short (12 words on average). There are some notable differences between categories like the titles for movie questions being longer (15 words on average) or the answers in the book category being longer (16 words on average).

Table 3
Overview of our TOMT-KIS dataset by year: number of questions (#), proportion of questions that are solved as identified by our heuristic (Solv.), and average time in hours to solve a question (Δ_T); for all questions combined and for the four most frequent categories (Movies, Music, Books, Games).

Year	All Questions			Movies			Music			Books			Games		
	#	Solv.	Δ_T	#	Solv.	Δ_T	#	Solv.	Δ_T	#	Solv.	Δ_T	#	Solv.	Δ_T
2009	1,045	0.01	2.11	47	0.00	—	36	0.03	0.22	18	0.00	—	9	0.00	—
2010	3,861	0.02	96.34	268	0.03	671.28	97	0.02	0.49	97	0.00	—	25	0.08	0.09
2011	24,544	0.02	10.88	2,702	0.02	2.95	1,591	0.02	2.01	834	0.03	99.47	246	0.03	1.07
2012	52,356	0.34	22.58	7,672	0.38	13.25	6,231	0.35	72.53	2,294	0.33	15.81	1,491	0.42	18.05
2013	84,675	0.48	9.04	11,332	0.53	23.26	12,564	0.46	6.83	2,894	0.50	14.41	2,326	0.56	5.92
2014	95,949	0.42	6.30	13,747	0.47	4.30	14,961	0.42	7.25	2,878	0.44	9.33	2,422	0.48	13.76
2015	110,609	0.43	10.07	16,775	0.47	5.21	18,462	0.43	12.81	3,392	0.42	10.07	2,798	0.46	9.78
2016	97,984	0.47	6.17	16,021	0.50	4.78	19,143	0.46	5.53	2,996	0.44	5.54	2,380	0.49	4.44
2017	100,888	0.46	3.76	17,081	0.51	3.40	19,532	0.44	4.24	3,215	0.45	5.79	2,272	0.48	4.25
2018	124,126	0.48	4.47	21,058	0.53	3.75	21,400	0.45	6.78	4,530	0.49	5.13	2,736	0.47	5.17
2019	132,977	0.52	11.13	26,491	0.58	6.31	25,029	0.48	13.90	5,670	0.57	9.96	3,772	0.49	10.67
2020	176,154	0.50	17.30	34,962	0.55	11.91	28,091	0.49	22.03	7,079	0.54	22.81	4,547	0.47	18.37
2021	143,675	0.52	24.78	30,008	0.57	17.72	22,126	0.52	26.81	5,837	0.54	23.59	4,104	0.49	49.64
2022	130,582	0.50	8.82	27,907	0.56	5.73	19,437	0.50	11.61	5,840	0.51	9.12	3,831	0.46	15.12
Total	1,279,425	0.47	11.78	226,071	0.52	8.99	208,700	0.46	14.38	47,574	0.49	13.43	32,959	0.48	16.33

Table 4
Avg. lengths (characters / words) of the title and content of a question, and of the identified answer.

Length	All Questions			Movies			Music			Books			Games		
	title	cont.	answ.	title	cont.	answ.	title	cont.	answ.	title	cont.	answ.	title	cont.	answ.
# Char.s	83.42	420.70	41.58	92.67	483.90	43.77	77.17	365.92	37.13	92.54	614.41	58.01	80.86	520.73	37.37
# Words	13.86	81.67	11.83	15.14	95.08	11.11	12.65	67.16	11.75	15.11	118.41	16.05	13.07	100.10	10.59

3. Known-Item Question Performance Prediction on TOMT-KIS

We conduct a pilot study on TOMT-KIS with the seven pre-retrieval query performance predictors implemented in the qpptk toolkit[4]—including the state-of-the-art max-var approach [6, 7]. In the experiments, we use the Robust04 CIFF index[5] from the Open-Source IR Replicability Challenge [12] to compute corpus-relative values that the predictors need.

Following previous practice on evaluating query performance prediction [6], Table 5 shows the rank correlations of the predictors to the ground truth in form of Kendall's τ, Spearman's ρ, and Pearson's r (a score of 1 indicates perfect correlation, 0 indicates random correlation, and -1 indicates perfect inverse correlation). In all scenarios, the existing predictors only achieve correlation scores close to 0 which indicates that they are not suited for known-item question performance prediction. For the development of more effective known-item question performance prediction, our new TOMT-KIS dataset can form an ideal starting point.

[4] https://github.com/Zendelo/QPP-EnhancedEval/tree/main/code/qpptk
[5] https://github.com/osirrc/ciff/blob/master/README.md

Table 5

Effectiveness of the seven pre-retrieval predictors implemented in qpptk (max-idf, avg-idf, scq, avg-scq, var, max-var, avg-var) on our five known-item question test sets (all questions, movies, music, books, games). Correlation to the ground truth rankings given as Kendall's τ, Spearman's ρ, and Pearson's r.

Category	max-idf			avg-idf			scq			avg-scq			var			max-var			avg-var		
	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r
All	.06	.06	.07	.02	.02	.07	.11	.16	.09	-.05	-.06	.02	.12	.18	.10	.08	.10	.03	-.06	-.07	.00
Movies	-.06	-.08	-.20	-.21	-.31	-.16	.03	.05	-.02	-.03	-.04	.04	.04	.06	-.00	.04	.06	.04	-.06	-.07	.01
Music	.04	.05	.02	.05	.07	.09	.07	.09	.02	-.06	-.08	-.01	.07	.09	.02	.01	.01	-.02	-.06	-.07	-.05
Books	.16	.18	.10	-.09	-.13	-.13	.04	.06	.01	-.11	-.16	-.15	.02	.03	.01	.00	-.00	-.03	-.11	-.16	-.16
Games	.15	.17	.18	.04	.06	.03	.05	.06	.03	-.06	-.09	-.07	.05	.07	.04	.05	.06	.05	-.05	-.05	.05

4. Conclusion

We have constructed the new large-scale TOMT-KIS dataset[6] for known-item question performance prediction by crawling the complete tip-of-my-tongue subreddit (1.28 million questions; 47 % with heuristically identified answers). As a proxy for the performance of a question, we use the time elapsed until the solving answer was posted. In a pilot study, none of the existing pre-retrieval query performance predictors implemented in the qpptk toolkit could really predict a known-item question's performance. Known-item question performance prediction thus is still not "solved" and forms an interesting subject for future research—with our dataset as a possible starting point. Other interesting directions could be to use TOMT-KIS as an enrichment of the training data provided by the organizers of the upcoming TREC 2023 ToT track.

References

[1] A. Z. Broder, A taxonomy of web search, SIGIR Forum 36 (2002) 3–10. URL: https://doi.org/10.1145/792550.792552.

[2] M. Hagen, D. Wägner, B. Stein, A corpus of realistic known-item topics with associated web pages in the ClueWeb09, in: Proceedings of ECIR 2015, 2015, pp. 513–525. URL: https://doi.org/10.1007/978-3-319-16354-3_57.

[3] J. Arguello, A. Ferguson, E. Fine, B. Mitra, H. Zamani, F. Diaz, Tip of the tongue known-item retrieval: A case study in movie identification, in: Proceedings of CHIIR 2021, 2021, pp. 5–14. URL: https://doi.org/10.1145/3406522.3446021.

[4] F. Meier, T. Bogers, M. Gäde, L. E. Thomsen, Towards understanding complex known-item requests on Reddit, in: Proceedings of HT 2021, 2021, pp. 143–154. URL: https://doi.org/10.1145/3465336.3475096.

[5] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, sMARE: A new paradigm to evaluate and understand query performance prediction methods, Information Retrieval Journal 25 (2022) 94–122. URL: https://doi.org/10.1007/s10791-022-09407-w.

[6] Data and code freely available under a permissive license: https://github.com/webis-de/QPP-23

[6] D. Carmel, E. Yom-Tov, Estimating the Query Difficulty for Information Retrieval, Morgan & Claypool Publishers, 2010. URL: https://doi.org/10.2200/S00235ED1V01Y201004ICR015.

[7] C. Hauff, D. Hiemstra, F. de Jong, A survey of pre-retrieval query performance predictors, in: Proceedings of CIKM 2008, 2008, pp. 1419–1420. URL: https://doi.org/10.1145/1458082.1458311.

[8] S. Bhargav, G. Sidiropoulos, E. Kanoulas, 'It's on the tip of my tongue': A new dataset for known-item retrieval, in: Proceedings of WSDM 2022, 2022, pp. 48–56. URL: https://doi.org/10.1145/3488560.3498421.

[9] I. K. H. Jørgensen, T. Bogers, "Kinda like The Sims … But with ghosts?": A qualitative analysis of video game re-finding requests on Reddit, in: Proceedings of FDG 2020, 2020, pp. 40:1–40:4. URL: https://doi.org/10.1145/3402942.3402971.

[10] C. Hauff, M. Hagen, A. Beyer, B. Stein, Towards realistic known-item topics for the ClueWeb, in: Proceedings of IIiX 2012, 2012, pp. 274–277. URL: https://doi.org/10.1145/2362724.2362773.

[11] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, An enhanced evaluation framework for query performance prediction, in: Proceedings of ECIR 2021, 2021, pp. 115–129. URL: https://doi.org/10.1007/978-3-030-72113-8_8.

[12] R. Clancy, N. Ferro, C. Hauff, J. Lin, T. Sakai, Z. Z. Wu (Eds.), Proceedings of OSIRRC@SIGIR 2019, 2019. URL: http://ceur-ws.org/Vol-2409.