Using Keyqueries to Reduce Misinformation in Health-Related Search Results

Maik Fröbe, Sebastian Günther, Alexander Bondarenko, Johannes Huck and Matthias Hagen

Martin-Luther-Universität Halle-Wittenberg

Abstract

In the scenario of health-related searches, we investigate whether explicit relevance feedback by experts can guide query expansion methods to formulate queries that return fewer misleading or wrong results. In contrast to standard query expansion methods that pay no attention to the ranks of the feedback documents in the results of the expanded query, we experiment with a keyquery-based approach to identify expanded queries for which the feedback documents are ranked as high as possible. Experiments on the TREC 2019–2021 Decision and Health Misinformation tracks show that our keyquery-based method substantially reduces the portion of harmful results and improves the overall retrieval effectiveness.

Keywords

Health misinformation, Keyqueries, Query expansion, TREC evaluation

1. Introduction

Health-related web search results often contain wrong or misleading information that can be harmful to searchers who simply trust the presented information returned at the top ranks [1, 2, 3]. Since many people nowadays use search engines to look for health information online [4], research on how to return helpful instead of harmful health-related search results has gained attention [5, 6, 7]—with the particular challenge that scientific knowledge changes rapidly.[1]

In the context of general ad-hoc search, query expansion through relevance feedback can improve the ranking effectiveness [8]—motivating us to study the effect for health-related searches. In our approach, we examine different amounts of explicit relevance feedback by medical experts who identify relevant, up-to-date, and scientifically grounded information for health-related searches (i.e., the feedback documents should be topically relevant and should not promote potentially harmful actions according to the current scientific knowledge).

Effective query expansion approaches like RM3 [9] add new terms to a query by exploiting information from feedback documents labeled as relevant to the initial query. However, RM3 does not consider the ranks of the feedback documents in the result list of the expanded query and also does not check whether all expansion terms are actually needed. In the scenario

ROMCIR 2022: The 2nd Workshop on Reducing Online Misinformation through Credible Information Retrieval, held as part of ECIR 2022: the 44th European Conference on Information Retrieval, April 10–14, 2022, Stavanger, Norway

maik.froebe@informatik.uni-halle.de (M. Fröbe)

0000-0002-1003-981X (M. Fröbe)

[1] www.nytimes.com/2017/01/16/upshot/how-to-prevent-whiplash-from-ever-changing-medical-advice.html

of a health-related search, this behavior might be inefficient when employing the possibly costly feedback from some medical experts. We thus experiment with an RM3-based expansion approach that accounts for the ranks of the feedback documents and tries to use as little expansion terms as possible. To this end, we combine RM3 with the idea of keyqueries [10, 11]. A keyquery for some document set D is a query that returns (many of) the documents of D at high result ranks (effectiveness) while returning at least a specified number of results (generality) and using as few terms as possible (minimality). Our idea is to use the original query and the RM3 expansion terms to formulate a keyquery for the feedback documents. In this way, the effectiveness of the RM3-expanded query is somewhat controlled (i.e., at least the feedback documents are ranked high), and the query is as general as possible (i.e., overfitting is minimized due to the keyquery's minimality and generality constraints). The underlying hypothesis is that the other results of a keyquery returned "around" the feedback documents then are also relevant and do not contain harmful information.

We compare the effectiveness of RM3 and our proposed keyquery-enhanced RM3 variant for health-related searches from the TREC 2019–2021 Decision and Health Misinformation tracks [5, 6, 7] (131 topics with manual relevance and helpful/harmful judgments). Using a subset of the annotations to "simulate" explicit relevance feedback from medical experts, our experimental results indicate that the keyquery enhancement improves upon RM3 in most cases, and that both expansion approaches substantially improve upon a BM25 baseline. Our code, feedback seeds, and results are publicly available.[2]

2. Related Work

In this section, we review studies about health-related searches and misinformation, as well as existing query expansion approaches.

Health-related web searches. Since the early days of web search, people look for diseases, symptoms, and treatments. For instance, studies by Spink et al. [12] and Jansen and Spink [13] found that 4.5–11.5% of the queries submitted to AltaVista, Ask Jeeves, or Excite in the late 1990s and early 2000s were health-related. Later, Purcell et al. [14] interviewed Americans and found that 66% used the Web to read about or search for health-related information (only 'weather' with 81% and 'national events' with 73% were more popular); a similar number was also found for Saudi Arabians by AlGhamdi and Moussa [15] (58% for health-related information). The Web thus is a primary source to gather information about diseases, symptoms, and treatments [16]. Many people even use the Web as a diagnostic "tool" [12, 17, 4] and, instead of visiting a medical professional, start their endeavor using a web search engine [4, 18]. Studies of a year-long log of 1.5 billion questions submitted to Yandex showed that the overall share of health-related questions is rather stable over the year [19]—even though some information needs are seasonal (e.g., influenza-related ones [20])—and that about 5% of the questions focus on the helpfulness of treatments for medical conditions [3].

Health-related misinformation. The Web is full of wrongful information of any kind, including the health domain [2] as recently indicated by COVID-related misinformation [21]. Statements like "ginger is more effective at killing cancer than chemo" may cause severe harm to

[2] https://github.com/webis-de/ROMCIR-22

people who simply believe them and ground their decisions and actions on the misinformation [1, 2]. Some years ago, several studies showed that more than half of the top web search results to medical yes/no questions return incorrect answers [22, 23]. Recently, Bondarenko et al. [3] analyzed the web search result snippets for questions asking about treatments and still found that in at least 44% of the cases the answers are misleading (e.g., suggesting treatment that are not helpful or even harmful according to the current scientific knowledge). Searchers are often influenced by such wrong online information and will believe that ineffective treatments are effective [24]. Given these alarming findings, we take one step back from the fully automatic approaches often employed to reduce misinformation in health-related searches [5, 6, 7, 25]. We analyze to what extent query expansion methods can leverage explicit feedback by experts to reduce the harmfulness while increasing the helpfulness of search results.

Reducing harmful misinformation. Some suggestions to address the issue of health-related misinformation are health cards [26], nutrition labels and fact boxes [27], or nudging [28, 29] and boosting [30]. Focusing on the retrieval phase, the TREC 2019–2021 Decision and Health Misinformation tracks [5, 6, 7] ask to develop systems that reduce harmful misinformation in health-related search results.[3] The state of the art at these tracks is the Vera system [25] that linearly combines the relevance score of MonoT5 (for low-ranked documents) or DuoT5 (for the top 50 documents only, since it is computationally expensive) with a T5 prediction that a given document aligns with the current scientific knowledge. Vera also employs expert feedback (reformulating queries based on the topic description and the answer field that indicates the scientific answer to the information need) and outperforms manual runs [31]. However, Vera has not been compared to RM3 query expansion with explicit relevance feedback so far—a gap that we close in our experiments.

Query expansion. Query expansion methods extend an original query with additional terms to retrieve relevant documents with a higher probability [32]. The additional terms often are derived from a set of feedback documents that is either explicitly created from user feedback (e.g., clicks or judgments) or implicitly created (i.e., pseudo-relevance feedback) from the original query's top-ranked results. The RM3 query expansion method [9] can use both, explicit or pseudo-relevance feedback, and is a strong baseline [33]. In our experiments on the task of reducing misinformation in health-related search results, we compare a "classic" RM3-based query expansion approach with explicit expert feedback to modern transformer-based retrieval approaches that are said to have caused a paradigm shift in the recent years [34, 35].

The first usage of relevance feedback through a relevance model [36] (referred to as RM1 [37]) assigns weights to documents by their retrieval score for the original query and derives expansion term weights as a weighted average of relative occurrence frequencies in the feedback documents. Given a set R of relevance feedback documents for the query q, the RM1 score of term t is:

$$\mathrm{RM1}(t, R) = \sum_{d \in R} P(t|d) \frac{score(q, d)}{\sum\limits_{d' \in R} score(q, d')} ,$$

where $P(t|d)$ is the probability that term t occurs in document d, and $score(q, d)$ is the retrieval score of d for the original query q (e.g., using BM25). A typical estimation for $P(t|d)$ (e.g.,

[3] https://trec-health-misinfo.github.io/

implemented in Anserini [38]) is to divide the frequency of t in d by the number of terms in d, i.e., $tf(t,d)/|d|$. The effectiveness of RM1 is improved by RM3 by linearly combining the RM1 weight with a query term weight as:

$$\text{RM3}(t,R) = \alpha \cdot \text{RM1}(t,R) + (1-\alpha) \cdot P(t|q) \ ,$$

where $P(t|q)$ is the probability that t occurs in the query (e.g., $tf(t,q)/|q|$)) and the $[0,1]$-valued α controls the feedback impact. Note that RM3 will assign non-zero weights to many terms. Still, implementations like the one in Anserini only use the m highest-weighting expansion terms to avoid retrieval efficiency issues for overlong queries.

Interestingly, RM3 pays no attention to the actual position of the feedback documents in the final ranking. In pseudo-relevance setups this might be a good decision since otherwise the pseudo-relevant top results of the original query might just stay on top. However, in scenarios with costly explicit expert feedback, not ranking the feedback documents high might "miss" some potential. Hence, we combine the idea of keyqueries [10, 11] (i.e., formulating queries that rank specific documents as high as possible) with RM3 in our experiments.

3. Keyquery-based Query Expansion

Based on the assumption that explicit relevance feedback in form of a few annotated relevant and helpful documents for a health-related topic is available from medical experts,[4] we combine RM3 query expansion [9] with the concept of keyqueries [10, 11, 39]. A query q is a *keyquery* for a set D of documents against some search engine S, iff q fulfills the following three conditions [11]: (1) every $d \in D$ is in the top-k results returned by S for q, (2) q has at least l results, and (3) no $q' \subset q$ fulfills the first two conditions. The first two conditions (i.e., the parameters k and l) determine the desired specificity and the generality of a keyquery. Following previous work [39], we set $k = 10$ and $l = 100$ to ensure that a keyquery retrieves each of the few feedback documents in the top-10 results while still being "general" enough to return at least 100 results. The third condition is a minimality constraint to avoid adding further terms to a query that already retrieves the target documents at high ranks.

Given a vocabulary V (in our case, the original query terms and the m expansion terms with the highest RM3 weights; i.e., m as a further parameter), the set $\mathcal{Q} = 2^V \setminus \{\emptyset\}$ represents the meaningful queries that can be formulated with terms from V. Note that $\mathcal{Q}$ might not contain any query that returns all documents from the relevance feedback set R in the top-k results (even for large k). In such cases, we iteratively relax the first keyquery condition by requiring that a keyquery retrieves $|R|-1$ feedback documents within the top-k results of the search engine S, if not possible then $|R|-2$, etc., until the condition is relaxed enough so that some keyqueries are found at some level. When more than one keyquery is found at some level, we select the one with the highest nDCG@k with respect to the feedback documents (i.e., all documents in R have a relevance of 1 and all other documents are irrelevant).

Since our focus is on effectiveness, we employ a simple brute-force method for keyquery computation and try every candidate query at each level. Our experiments on the TREC 2019–2021

[4] Being costly in practice, this setting is inspired by the TREC Relevance Feedback track [8]. Future work might try to replace explicit expert relevance feedback by similar schema.org annotations like https://schema.org/ClaimReview.

Decision and Health Misinformation tracks showed that the run time varies widely for different parameter settings and corpora. On a single thread of an Intel Xeon E5-2670 with 2.50 GHz, the "slowest" parameter setting resulted in 11:02 minutes per topic on the C4 dataset (Health Misinformation track 2021), 6:52 minutes on the Common Crawl News crawl (Health Misinformation track 2020), and 3:34 minutes on the ClueWeb12 category B (Decision track 2019). The median times varied between 1:40 minutes for the C4 dataset and 35 seconds for the ClueWeb12 category B. If the possible effectiveness benefit over plain RM3 expansion is substantial (i.e., returning the relevance feedback documents high in the rankings helps), speeding up the keyquery computation thus is an interesting direction for future research. Possible ideas might be to use a more efficient enumeration scheme [39] or a reverted index [40].

4. Evaluation

We compare the effectiveness of keyquery-enhanced RM3 expansion to "traditional" and neural retrieval systems on the TREC 2019–2021 Decision and Health Misinformation tracks.

4.1. Experimental Setup

We describe the corpora and health-related topics used in our experiments, how we "simulate" explicit expert feedback for the query expansion, and how the retrieval models were configured.

Topics and corpora. We use the 131 topics with relevance judgments from the TREC 2019–2021 Decision and Health Misinformation tracks and the corpora of the tracks. For each topic, documents were judged as relevant or irrelevant to the information need, and relevant documents were further annotated with helpful/harmful labels indicating their medical correctness. In the tracks' setup, relevant documents with harmful information are deemed worse than irrelevant documents. We use the evaluation scheme employed in the tracks: one qrel file with helpful relevant documents and one qrel file with harmful relevant documents against which the effectiveness (e.g., nDCG) should be maximized (help) or minimized (harm). Both scores can be combined by subtracting the harmful from the helpful effectiveness (help–harm).

The TREC 2019 Decision track [5] (HMI 19, for short, as the track was later renamed) used the ClueWeb12 category B subset[5] as the document corpus (52 million English web pages, crawled in 2012). We split the 50 topics with judgments into 3 folds (topics 1—17, 18—34, and 35—50) to run 3-fold cross-validation experiments. We use 3-fold cross-validation since we want to have the same number of folds for all tracks but 5-fold or 10-fold would yield rather small folds for the 2021 Health Misinformation track with judgments for only 35 topics.

The TREC 2020 Health Misinformation track [6] (HMI 20) used the Common Crawl News crawl as the document corpus (65 million news articles, crawled from January to April 2020). We split the 46 topics with judgments into 3 folds (topics 1—15, 16—32, and 33—50).

The TREC 2021 Health Misinformation track [7] (HMI 21) used the noclean version of the C4 dataset [41] as the document corpus (1 billion English web pages). We split the 35 topics with judgments into 3 folds (topics 101—113, 114—129, and 130—150).

[5] https://lemurproject.org/clueweb12/

Simulated explicit relevance feedback. Inspired by the setup of the TREC 2010 Relevance Feedback track [8], for each topic, the explicit relevance feedback are the highest ranked k documents from a BM25 ranking (Anserini implementation with default parameters) that are judged as relevant and helpful. We show results for $k = 1, \ldots, 5$ and for k being a hyperparameter tuned in the cross-validation.

Retrieval models and training. We compare six retrieval systems (using Anserini [38] and PyGaggle [42] implementations) and also include the three best submissions from the corresponding tracks. As our four baselines, we use BM25, MonoBERT, MonoT5, and a naïve re-ranker that simply moves the feedback documents to the top ranks of the BM25 ranking. The two query expansion approaches (RM3 with and without keyquery-enhancement) are implemented as an extension to Anserini's RM3 query expansion. We preprocess queries and the indexed texts via Porter stemming and stopword removal using Lucene's default stopwords for English. Score ties within a ranking are resolved via alphanumeric ordering by document ID as implemented in Anserini (given random document IDs, this leads to a random distribution with respect to other document properties such as text length [43]).

For each topic, we first retrieve the top-1000 BM25 results (Anserini implementation) and then apply 3-fold cross-validation as implemented in PyTerrier [44] to optimize help–harm for nDCG@10. During cross-validation, for BM25, we tune $k_1 \in \{0.7, 0.8, 0.9, 1.0, 1.1\}$ and $b \in \{0.3, 0.35, 0.4, 0.45, 0.5\}$. For MonoBERT and MonoT5, we re-rank the top-100 results of BM25 (default configuration) and leave all hyperparameters at their defaults (the models are pre-trained on MS MARCO). For RM3, we tune the number of feedback terms between 5 and 10, and $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$. For our keyquery-enhanced RM3 approach, we tune the size $|V|$ of the keyquery vocabulary between 8 and 13 terms and $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$ but ensuring that the final expanded query has the same length as the "plain" RM3 expansion. For expansions with variable amount of feedback (cf. 'var' in Table 1), we tune the number of feedback documents between 1 and 5.

4.2. Experimental Results

Table 1 shows the 3-fold cross-validated experimental results on the TREC 2019–2021 Decision and Health Misinformation tracks (column groups HMI 19, HMI 20, and HMI 21). We report the nDCG@10 using the official qrels that assign positive gain scores only to relevant and helpful documents (column 'Help') and the official qrels that assign "positive" gain scores only to relevant and harmful documents (column 'Harm' that indicates the "effectiveness" of retrieving documents that may be harmful for a searcher). The goal of an effective system is to maximize the help score while minimizing the harm score. We follow the track organizers' suggestion [5, 6, 7] and report the help–harm difference as a single value to compare systems (column 'Diff.'). The baselines (BM25, and the MonoBERT and MonoT5 re-rankers, as well as the 'Top' re-ranker that moves the feedback documents to the top of the BM25 ranking) are contrasted by the three best runs submitted to the corresponding track and the RM3 query expansion with and without keyqueries for 1 to 5 feedback documents and a cross-validation-tuned number of feedback documents (row group 'var').

Note that the results we report for the best runs submitted to TREC may differ from the tracks' original results, since we have re-evaluated the runs in our cross-validation setup. In particular,

Table 1
The effectiveness of retrieval systems on the TREC 2019–2021 Decision and Health Misinformation tracks (column groups HMI 19, HMI 20, and HMI 21) measured as nDCG@10 in retrieving helpful (column 'Help', higher scores are better) or harmful ('Harm', lower scores are better) documents; score difference in column 'Diff.' (higher is better). The column 'Feedback' indicates the number of provided feedback documents. Statistically significant differences ($p = 0.05$ with Bonferroni correction) are indicated by † (compared to 1st@TREC) or ‡ (comparing RM3 to keyquery-enhanced RM3 (KQ-RM3)).

Retrieval system		**HMI 19**			**HMI 20**			**HMI 21**		
Feedback	Model	Help	Harm	Diff.	Help	Harm	Diff.	Help	Harm	Diff.
—	BM25	0.19	0.35†	-0.16†	0.29†	0.04	0.25†	0.29†	0.18†	0.11†
	+MonoBERT	0.21	0.33†	-0.12	0.16†	**0.03**	0.14†	0.21†	0.12†	0.09†
	+MonoT5	0.22†	0.34†	-0.11	0.30†	0.06	0.24†	0.20†	0.14†	0.07†
	1st @TREC	**0.26**	0.28	-0.02	**0.66**	0.05	**0.62**	0.52	**0.08**	0.44
	2nd@TREC	0.21	0.27	-0.06	0.46	0.05	0.41	**0.57**	**0.08**	**0.49**
	3rd @TREC	0.17	**0.15**	**0.02**	0.43	0.09	0.35	0.53	**0.08**	0.46
1	+Top	0.38†	0.25	0.13†	0.37†	**0.03**	0.34†	0.27†	**0.08**	0.19†
	+RM3	0.42†	**0.19**†	0.22†	**0.48**†	0.05	**0.43**†	0.37†	0.09	0.28†
	+KQ-RM3	**0.44**†	0.20†	**0.24**†	0.47†	0.06	0.41†	**0.43**‡	**0.08**	**0.36**‡
2	+Top	0.47†	0.21	0.26†	0.46†	**0.04**	0.43†	0.30	**0.07**	0.24
	+RM3	0.48†	0.20†	0.27†	0.55†	0.06	0.48†	0.36†	0.14	0.22†
	+KQ-RM3	**0.54**†‡	**0.18**†	**0.36**†‡	**0.56**†	0.06	**0.51**†	**0.41**†‡	0.10‡	**0.32**‡
3	+Top	**0.53**†	**0.19**	**0.34**†	0.52†	**0.02**	**0.50**	0.33	**0.05**	0.28
	+RM3	0.45†	0.22	0.23†	**0.55**†	0.06	0.48†	0.36†	0.11	0.25†
	+KQ-RM3	0.49†	0.21	0.29†	0.52†	0.06	0.46†	**0.41**†	0.11	**0.29**†
4	+Top	**0.57**†	**0.19**	**0.38**†	**0.56**	**0.02**	**0.54**	0.37	**0.03**	0.34
	+RM3	0.46†	0.22	0.24†	0.53†	0.06	0.47†	0.39†	0.12	0.27†
	+KQ-RM3	0.50†	0.22	0.27†	0.54†	0.05	0.49†	**0.46**‡	0.09	**0.37**‡
5	+Top	**0.60**†	**0.17**	**0.43**†	**0.60**	**0.01**†	**0.59**	0.39	**0.02**	**0.37**
	+RM3	0.45†	0.21	0.23†	0.54†	0.05	0.49†	0.37†	0.10	0.27†
	+KQ-RM3	0.52†‡	0.23	0.29†	0.56†	0.04	0.51†	**0.43**†‡	0.10	0.33‡
var	+Top	**0.60**†	**0.17**	**0.43**†	**0.60**	**0.01**†	**0.59**	0.39	**0.02**	**0.37**
	+RM3	0.45†	0.21	0.24†	0.52†	0.06	0.46†	0.36†	0.12	0.24†
	+KQ-RM3	0.54†‡	0.18†	0.36†‡	0.55†	0.06	0.49†	**0.46**‡	0.09	**0.37**‡

for HMI 19 and HMI 21, the cross-validated help–harm differences would have resulted in another ranking of the best three TREC runs.

From the results in Table 1, it can be observed that both query expansion approaches substantially improve the help–harm difference compared to BM25, MonoBERT, and MonoT5 (often by reducing the harmfulness while increasing the helpfulness). The keyquery-enhanced RM3 approach achieves better effectiveness than the plain RM3 expansion in almost all setups (i.e., higher helpfulness at lower harmfulness). Interestingly, more relevance feedback documents are not necessarily better; often one or two feedback documents yield the highest keyquery-

enhanced RM3 effectiveness. Starting from three feedback documents, the simple 'Top' baseline that moves the feedback documents to the top of the BM25 ranking often achieves better results than the RM3 variants. This indicates that the expanded queries of both RM3 approaches then cannot retrieve many of the feedback documents at the absolute top ranks.

On HMI 19, a single feedback document suffices to improve upon the most effective runs submitted to the TREC track. For HMI 20 and HMI 21, Vera, the most effective run submitted to these TREC tracks, is more effective than the query expansion variants—though sometimes the difference is not statistically significant.

Overall, our experiments indicate the usefulness of explicit relevance feedback for health-related searches—the RM3 variants are always more effective than the BM25-based baselines. In most cases, the keyquery-enhanced RM3 variant (i.e., taking the result ranks of the explicit feedback documents into account when expanding a query) improves upon plain RM3. Even Vera could be viewed to incorporate explicit feedback since the actual correct medical answer for a topic is used to formulate a better query.

5. Conclusion and Future Work

In the scenario of returning fewer misleading or wrong search results for health-related information needs, we have studied whether enhancing RM3 query expansions with the concept of keyqueries leads to more effective BM25 queries. Our experiments show that the effectiveness of standard RM3 is improved by our new keyquery-enhanced variant.

In future work, we plan to expand our study to other relevance feedback approaches and retrieval models implemented in Anserini and to incorporate more efficient enumeration schemes for the keyquery computation. Furthermore, we will also experiment with replacing the costly explicit expert relevance feedback by trusted information available on the Web (e.g., by exploiting https://schema.org/ClaimReview annotation).

References

[1] F. A. Pogacar, A. Ghenai, M. D. Smucker, C. L. A. Clarke, The positive and negative influence of search results on people's decisions about the efficacy of medical treatments, in: Proceedings of ICTIR, 2017, pp. 209–216.

[2] E. Dai, Y. Sun, S. Wang, Ginger cannot cure cancer: Battling fake health news with a comprehensive data repositor, in: Proceedings of ICWSM, 2020, pp. 853–862.

[3] A. Bondarenko, E. Shirshakova, M. Driker, M. Hagen, P. Braslavski, Misbeliefs and biases in health-related searches, in: Proceedings of CIKM, 2021, pp. 2894–2899.

[4] S. Fox, M. Duggan, Health online 2013, Pew Internet Report, 2013.

[5] M. Abualsaud, M. D. Smucker, C. Lioma, M. Maistro, G. Zuccon, Overview of the TREC 2019 Decision track, in: Proceedings of TREC, 2019.

[6] C. L. A. Clarke, S. Rizvi, M. D. Smucker, M. Maistro, G. Zuccon, Overview of the TREC 2020 Health Misinformation track, in: Proceedings of TREC, 2020.

[7] C. L. A. Clarke, M. Maistro, M. D. Smucker, Overview of the TREC 2021 Health Misinformation track, in: Proceedings of TREC, 2021.

[8] C. Buckley, M. Lease, M. Smucker, Overview of the TREC 2010 Relevance Feedback track, in: Proceedings of TREC, 2010.
[9] N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, C. Wade, UMass at TREC 2004: Novelty and HARD, in: Proceedings of TREC, 2004.
[10] T. Gollub, M. Hagen, M. Michel, B. Stein, From keywords to keyqueries: Content descriptors for the Web, in: Proceedings of SIGIR, 2013, pp. 981–984.
[11] M. Hagen, A. Beyer, T. Gollub, K. Komlossy, B. Stein, Supporting scholarly search with keyqueries, in: Proceedings of ECIR, 2016, pp. 507–520.
[12] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D. P. Lorence, S. Ozmutlu, H. C. Ozmutlu, A study of medical and health queries to web search engines, Health Information & Libraries Journal 21 (2004) 44–51.
[13] B. J. Jansen, A. Spink, How are we searching the World Wide Web? A comparison of nine search engine transaction logs, Inf. Process. Manag. 42 (2006) 248–263.
[14] K. Purcell, L. Rainie, A. Mitchell, T. Rosenstiel, K. Olmstead, Understanding the participatory news consumer, Pew Internet Report, 2010.
[15] K. M. AlGhamdi, N. A. Moussa, Internet use by the public to search for health-related information, Int. J. Medical Informatics 81 (2012) 363–373.
[16] M. Cartright, R. W. White, E. Horvitz, Intentions and attention in exploratory health search, in: Proceeding of SIGIR, 2011, pp. 65–74.
[17] R. W. White, E. Horvitz, Cyberchondria: Studies of the escalation of medical concerns in web search, ACM Trans. Inf. Syst. 27 (2009) 23:1–23:37.
[18] L. J. Finney Rutten, K. D. Blake, A. J. Greenberg-Worisek, S. V. Allen, R. P. Moser, B. W. Hesse, Online health information seeking among US adults: Measuring progress toward a healthy people 2020 objective, Public Health Reports 134 (2019) 617–625.
[19] M. Völske, P. Braslavski, M. Hagen, G. Lezina, B. Stein, What users ask a search engine: Analyzing one billion Russian question queries, in: Proceedings of CIKM, 2015, pp. 1571–1580.
[20] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data, Nature 457 (2009) 1012–1014.
[21] A. C. Windfeld, F. M. Meier, Does vinegar kill coronavirus? – Using search log analysis to estimate the extent of COVID-19-related misinformation searching behaviour in the United States, in: Proceedings of iConference, 2021.
[22] R. White, Beliefs and biases in web search, in: Proceedings of SIGIR, 2013, pp. 3–12.
[23] R. W. White, A. H. Awadallah, Content bias in online health search, ACM Trans. Web 8 (2014) 25:1–25:33.
[24] A. Ghenai, M. D. Smucker, C. L. A. Clarke, A think-aloud study to understand factors affecting online health search, in: Proceedings of CHIIR, 2020, pp. 273–282.
[25] R. Pradeep, X. Ma, R. Nogueira, J. Lin, Vera: Prediction techniques for reducing harmful misinformation in consumer health search, in: Proceedings of SIGIR, 2021, pp. 2066–2070.
[26] Jimmy, G. Zuccon, B. Koopman, G. Demartini, Health cards for consumer health search, in: Proceedings of SIGIR, 2019, pp. 35–44.
[27] S. Zimmerman, S. Herzog, J. Chamberlain, D. Elsweiler, U. Kruschwitz, Towards a framework for harm prevention in web search, in: Proceedings of BIRDS@SIGIR, 2020, pp. 30–46.

[28] R. Thaler, C. Sunstein, Nudge: Improving decisions about health, wealth, and happiness, Penguin, 2009.
[29] S. Zimmerman, A. Thorpe, C. Fox, U. Kruschwitz, Privacy nudging in search: Investigating potential impacts, in: Proceedings of CHIIR, 2019, pp. 283–287.
[30] S. Zimmerman, A. Thorpe, J. Chamberlain, U. Kruschwitz, Towards search strategies for better privacy and information, in: Proceedings of CHIIR, 2020, pp. 124–134.
[31] J. Bevendorff, A. Bondarenko, M. Fröbe, S. Günther, M. Völske, B. Stein, M. Hagen, Webis at TREC 2020: Health Misinformation track, in: Proceedings of TREC, 2020.
[32] C. Carpineto, G. Romano, A survey of automatic query expansion in information retrieval, ACM Comput. Surv. 44 (2012) 1:1–1:50.
[33] J. Lin, The neural hype and comparisons against weak baselines, SIGIR Forum 52 (2018) 40–51.
[34] J. Lin, The neural hype, justified!: A recantation, SIGIR Forum 53 (2019) 88–93.
[35] J. Lin, R. Nogueira, A. Yates, Pretrained transformers for text ranking: BERT and beyond, Morgan & Claypool Publishers, 2021.
[36] V. Lavrenko, W. B. Croft, Relevance-based language models, in: Proceedings of SIGIR, 2001, pp. 120–127.
[37] Y. Lv, C. Zhai, A comparative study of methods for estimating query language models with pseudo feedback, in: Proceedings of CIKM, 2009, pp. 1895–1898.
[38] P. Yang, H. Fang, J. Lin, Anserini: Enabling the use of Lucene for information retrieval research, in: Proceedings of SIGIR, 2017, pp. 1253–1256.
[39] M. Fröbe, E. O. Schmidt, M. Hagen, Efficient query obfuscation with keyqueries, in: Proceedings of WI-IAT, 2021.
[40] M. Völske, T. Gollub, M. Hagen, B. Stein, A keyquery-based classification system for CORE, D Lib Mag. 20 (2014).
[41] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67.
[42] J. Lin, X. Ma, S. Lin, J. Yang, R. Pradeep, R. Nogueira, Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations, in: Proceedings of SIGIR, 2021, pp. 2356–2362.
[43] J. Lin, P. Yang, The impact of score ties on repeatability in document ranking, in: Proceedings of SIGIR, 2019, pp. 1125–1128.
[44] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, PyTerrier: Declarative experimentation in Python from BM25 to dense retrieval, in: Proceedings of CIKM, 2021, pp. 4526–4533.