Exposing Query Identification for Search Transparency

Ruohan Li
Carnegie Mellon University, Microsoft
United States
ruohanli@microsoft.com

Jianxiang Li
Carnegie Mellon University, Microsoft
United States
jianxiangli@microsoft.com

Bhaskar Mitra
Microsoft, University College London
Canada
bmitra@microsoft.com

Fernando Diaz
Mila Quebec
Artificial Intelligence Institute
Canada
diazf@acm.org

Asia J. Biega*
Max Planck Institute
for Security and Privacy
Germany
asia.biega@mpi-sp.org

ABSTRACT

Search systems control the exposure of ranked content to searchers. In many cases, creators value not only the exposure of their content but, moreover, an understanding of the specific searches where the content is surfaced. The problem of identifying which queries expose a given piece of content in the ranked results is an important and relatively underexplored search transparency challenge. Exposing queries are useful for quantifying various issues of search bias, privacy, data protection, security, and search engine optimization. Exact identification of exposing queries in a given system is computationally expensive, especially in dynamic contexts such as web search. We explore the feasibility of approximate exposing query identification (EQI) as a retrieval task by reversing the role of queries and documents in two classes of search systems: dense dual-encoder models and traditional BM25. We then improve upon this approach through metric learning over the retrieval embedding space. We further derive an evaluation metric to measure the quality of a ranking of exposing queries, as well as conducting an empirical analysis of various practical aspects of approximate EQI. Overall, our work contributes a novel conception of transparency in search systems and computational means of achieving it.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Information systems** → **Specialized information retrieval**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Search exposure, Exposing queries, Transparency, Privacy

ACM Reference Format:
Ruohan Li, Jianxiang Li, Bhaskar Mitra, Fernando Diaz, and Asia J. Biega. 2022. Exposing Query Identification for Search Transparency. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3485447.3512262

*Work partly done while at Microsoft Research.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9096-5/22/04.
https://doi.org/10.1145/3485447.3512262

1 INTRODUCTION

Given a large repository of searchable content, information retrieval (IR) systems control the exposure of the content to searchers. In many cases, creators value not only the exposure of their content but also an understanding of the search contexts in which their content is surfaced. To enable transparency about search contexts facilitated by a given system, search engines should identify which *search queries expose each of the corpus documents in the ranking results* [4]. Yet, search systems currently do not provide such transparency to content producers, as few non-brute-force methods exist that can be used for this task. Identification of such exposing queries is our focus in this paper.

Exposing Query Use Cases. Evidence that exposure transparency is important permeates multiple lines of work, including fairness and bias, privacy, search engine optimization, and even security. Azzopardi and Vinay [1] defined the *retrievability bias* as the disparity in the weighted sum of ranks different documents are retrieved at over all possible queries. Finding which queries expose which documents is therefore necessary to quantify and audit this bias. Fair rankings typically focus on particular queries [11], while there is virtually no way for an individual user to get an overview of which rankings overall expose their content. This is especially crucial in settings where exposure in rankings is highly monetizable, like hiring [13] or online marketplaces [22]. Biega et al. [4] argued that exposing queries are an important *privacy* tool in search systems, demonstrating a range of scenarios where *exposure to sensitive queries* is problematic [27]. The *right to be forgotten* [3] is one of the key data protection rights: Knowledge of exposing queries might enable users to execute this right more precisely than removing themselves from search results completely. Content creators could furthermore use the exposure transparency to aid in improving the performance of their documents in search results [30, 38]. Beyond content creators, service providers may also care about understanding how different content is exposed [26, 41] to better understand their systems. Finally, controlling which queries expose certain types of content might prevent unintended leaking of information through search engines [40].

Approaches. An intuitive strategy for identifying exposing queries is to analyze past search logs to find which documents were actually returned in response to specific queries. Yet, such logs quickly lose their relevance as a source of exposing queries as the search ecosystem is constantly changing and both the collection to be searched and the search models are updated. Furthermore, in certain

sensitive search contexts, the potential for exposure should ideally be captured before exposure occurs and is logged. Thus, exposing queries need to be computed based on a given search ranker rather than from logs. Exact brute-force computation assuming we have a universe of feasible queries—for instance, based on existing query logs and/or generated from the corpus documents—is practically inefficient as it involves issuing all these queries to the retrieval system and re-aggregating the ranking results. The question then becomes of whether we can do better.

Contributions. In this paper, we explore the feasibility of approximate exposing query identification (EQI) as a *retrieval* task in which the function of queries and documents is *reversed*. We study this idea in the context of two classes of search systems: dense embedding-based nearest neighbor search—*e.g.*, [18, 24, 42] and traditional BM25-based [32] retrieval using inverted-index [48]. In case of embedding-based search systems, we show how retrieval-based EQI can be improved if we treat the problem as a form of metric learning in a joint query-document embedding space, transforming an embedding space of a document retrieval model such that a nearest neighbor search in the transformed embedding space approximately corresponds to a reverse nearest neighbor search over the original embeddings. In search systems, whether a query exposes a document is a function of both the document relevance as well as how many other documents in the corpus are relevant. Our intuition, therefore, is that we might be able to reorganize queries and documents in the embedding space by adequately selecting training data that captures the corpus characteristics of dense document regions.

We furthermore derive an evaluation metric to measure the quality of a ranking of exposing queries similar to two-sided metrics previously used for assessing the effectiveness of query suggestion algorithms [33]. The metric reveals that the exposure significance of a query is dependent on the behavior of both the document retrieval users and the EQI system user. Overall, our work contributes a novel conception of transparency in search systems and explores a potential computational approach to achieving it.

2 RELATED WORK

Exposing query identification. The exposing query identification problem was defined by in the context of content creator privacy Biega et al. [4]. The authors proposed an EQI approach that identifies unigram and bigram exposing queries in the specific case of a ranker that uses unigram language models with Dirichlet smoothing. The solution generates synthetic candidate queries from the corpus and then prunes this queryset using a ranker-specific heuristic. The output exposing query set is exact, although computed over a limited set (unigram and bigram) of not always realistic synthetic queries. Instead, formulating EQI as a retrieval task over a corpus of queries allows us to: (i) model content exposure more adequately by incorporating user browsing models; and (ii) expand the space of solutions to include query retrieval methods that can handle arbitrary-length exposing queries.

EQI is closely related to the concept of *retrievability* developed by Azzopardi and Vinay [1]. Retrievability is a dot product score assigned to a document quantifying whether (1) a document is exposed within a rank a user is patient enough to inspect and (2) how likely a query is to be issued to the search engine. Our goal is to discover specific queries which expose a document to a searcher and could thus be used to estimate component (1) of retrievability. Further related concepts include *keyqueries* [14], which are the minimal queries retrieving a document in a given document retrieval system. By definition, keyqueries are thus a subset of all exposing queries.

Search transparency and explainability. EQI is a form of search *transparency*: the goal is to help users understand comprehensively what outcomes the search system yields for them. Transparency is closely related to interpretability and explainability, where the goal is to understand how and why a ranking algorithm returns certain results. Explainability approaches include generation of post-hoc explanations for black-box models, both local (explaining a single ranking result) [36, 39] and global (explaining a ranking model) [35]. Model-oriented techniques include modifying models to be more interpretable and transparent [47] or explaining through visualizations [8, 9], while user-oriented techniques include inferring the intent behind a search query that a model has assumed [34] or tailoring explanations to user mental models of search [37], evaluating explanations through the lens of trust [29].

Reversing the role of queries and documents. Our approach to EQI is inspired by several IR problems in which the role of documents and queries is reversed. Yang et al. [45] employ a query-by-document technique issuing the most informative phrases from a document as queries to identify related documents. In the context of EQI, the setup has a few differences: We retrieve exposing queries rather than other related documents, and are interested in an extensive coverage of exposing queries thus using the whole document as a query rather than just the most informative phrases. Pickens et al. [28] index query result sets, where each retrievable item is a query and each query is described by a list of pseudo-terms that correspond to document IDs that the query retrieves using a chosen retrieval system. This approach, referred to as *reverted indexing*, allows for query-by-multiple-documents. In our context, where we are interested in retrieving queries that expose a single document, reverted indexing is akin to the exact brute-force method.

Santos et al. [33] develop a framework for suggesting queries which might retrieve more relevant results for a given query. Nogueira et al. [25] develop a generative model which, for a given document, produces queries that could expand the document for more effective retrieval. Both approaches produce a set of queries to which a given document may be *relevant*, but not those that would expose the document when issued specifically to a given document retrieval system. We expect that the set of queries generated by these approaches to intersect with the ideal set of exposing queries, but as the behavior of the particular document retrieval system diverges from the labeled data, the intersection will be smaller; and the effectiveness may also vary between labeled-relevant, labeled-nonrelevant, and unlabeled documents. It is important in these cases for EQI to be effective for documents that are actually exposed, irrespective of their relevance.

3 EXPOSING QUERY IDENTIFICATION

We are given a *document retrieval* system producing a ranked list of documents retrieved from a collection $\mathcal{D}$ in response to a query.

Table 1: List of notations.

Notation	Description
$\mathcal{D}$	A collection of documents
$\mathcal{Q}$	A corpus of queries
d	An individual document
q	An individual query
Document retrieval	
σ_q	A ranked list of documents retrieved in response to q
σ_q^*	The ideal set of documents for q
$n_{q\to d}$	Number of documents retrieved per query
$\mu_{q\to d}$	A user browsing model for inspecting σ_q
$\gamma_{q\to d}$	Persistence parameter for inspecting σ_q
$\rho(d, \sigma_q)$	0-based rank of d in σ_q
Exposing query retrieval	
ψ_d	A ranked list of queries retrieved in response to d
ψ_d^*	The ideal ranked list of queries for d
$n_{d\to q}$	Number of queries retrieved per document
$\mu_{d\to q}$	A user browsing model for inspecting ψ_d
$\gamma_{d\to q}$	Persistence parameter for inspecting ψ_d
$\rho(q, \psi_d)$	0-based rank of q in ψ_d

We define EQI as a complementary retrieval task where, given a document d, the system is responsible for retrieving a ranked list of queries from a query log $\mathcal{Q}$, ranked according to their probability of exposing the document in the document retrieval system. Thus, EQI technically means reversing document retrieval. We use subscripts $\square_{q\to d}$ and $\square_{d\to q}$ to denote variables corresponding to the document retrieval and the EQI task, respectively. Table 1 presents the notations.

3.1 Deriving an Evaluation Metric

Retrieval metrics often take the form of a cumulative quality score summed over all items in the top-ranking positions weighted by the probability that a searcher inspects the document at a given rank: $\mathcal{M}(\sigma) = \sum_{d\in\sigma} \mu(d,\sigma)\cdot g_d$. Here, σ is the result list, the quality score g_d of a document d is typically a function of its relevance to the query and $\mu(d,\sigma)$ is the probability of a user inspecting d under the user browsing model μ. Metrics—such as, NDCG [19]—further normalize this value by the ideal value of the metric: Norm-$\mathcal{M}(\sigma) = \frac{\sum_{d_i\in\sigma}\mu(d_i,\sigma)\cdot g_{d_i}}{\sum_{d_j\in\sigma^*}\mu(d_j,\sigma^*)\cdot g_{d_j}}$ Where σ^* is the ideal ranked list of results. Similarly, in the exposing query retrieval scenario, we want to measure the quality of the ranked list of retrieved queries, which we refer to as the exposure list, in response to a document. Towards that goal, we propose the Ranked Exposure List Quality (RELQ), which assumes a similar form as Norm-$\mathcal{M}$:

$$\text{RELQ}_{\mu_{d\to q}}(\psi_d) = \frac{\sum_{q_i\in\psi_d}\mu_{d\to q}(q_i,\psi_d)\cdot g_{q_i}}{\sum_{q_j\in\psi_d^*}\mu_{d\to q}(q_j,\psi_d^*)\cdot g_{q_j}} \tag{1}$$

Where ψ_d and ψ_d^* are the retrieved and ideal exposure list for d.

The *key observation* now is that in case of EQI, the quality value g_q corresponds to the probability that q exposes d in the original retrieval system—which is given by $\mu_{q\to d}$:

$$\text{RELQ}_{\mu_{d\to q},\mu_{q\to d}}(\psi_d) = \frac{\sum_{q_i\in\psi_d}\mu_{d\to q}(q_i,\psi_d)\cdot\mu_{q\to d}(d,\sigma_{q_i})}{\sum_{q_j\in\psi_d^*}\mu_{d\to q}(q_j,\psi_d^*)\cdot\mu_{q\to d}(d,\sigma_{q_j})} \tag{2}$$

This derivation leads us to a metric that jointly accounts for the behavior of two distinct users: (1) the searcher using the document retrieval system and to whom the content is exposed, and (2) the user of the EQI system. Intuitively, the proposed metric has a higher value when the EQI user is exposed to queries that in turn prominently expose the target document in their corresponding rank lists.[1]

Metric instantiations. To compute the metric, we can plug in standard user behavior models for $\mu_{d\to q}$ and $\mu_{q\to d}$ that make different assumptions about how users interact with retrieved results under the document retrieval and the exposing query retrieval settings. For example, if we assume that both $\mu_{d\to q}$ and $\mu_{q\to d}$ are based on the same user model as the Rank-Biased Precision (RBP) metric [23] with persistence parameters $\gamma_{d\to q}\in(0,1]$ and $\gamma_{d\to q}\in(0,1]$, we get the following variant:

$$\text{RELQ}_{\text{RBP, RBP}}(\psi_d) = \frac{\sum_{q_i\in\psi_d}\gamma_{d\to q}^{\rho(q_i,\psi_d)}\cdot\gamma_{q\to d}^{\rho(d,\sigma_{q_i})}}{\sum_{q_j\in\psi_d^*}\gamma_{d\to q}^{\rho(q_j,\psi_d^*)}\cdot\gamma_{q\to d}^{\rho(d,\sigma_{q_j})}} \tag{3}$$

Where, $\rho(x,\Omega)$ is the 0-based rank of x in the ranked list Ω.

Instead of RBP, we can plug in different combinations of user models, including NDCG or exhaustive (where a user inspects all ranking results). For example, if we assume an exhaustive model for EQI, while adopting NDCG for the document retrieval:

$$\text{RELQ}_{\text{EXH, NDCG}}(\psi_d) = \frac{\sum_{q_i\in\psi_d}\mathrm{q}\{\mathrm{d}\in\sigma_{\mathrm{q_i}}\}\cdot\frac{\rho(\mathrm{d},\sigma_{\mathrm{q_i}})}{\log_2(\mathrm{i}+1)}}{\sum_{q_j\in\psi_d^*}\mathrm{q}\{\mathrm{d}\in\sigma_{\mathrm{q_j}}\}\cdot\frac{\rho(d,\sigma_{q_j})}{\log_2(j+1)}} \tag{4}$$

Where, $\mathrm{q}\{.\}$ is the indicator function.

3.2 Practical considerations

3.2.1 Prototypical users. $RELQ_{\text{RBP, RBP}}$ is parametrized by two user patience parameters: $\gamma_{q\to d}$ and $\gamma_{d\to q}$. Their values should be chosen to best reflect the behavior of users in a given underlying task. Absent behavioral data, we can instead represent certain prototypical users. Patient document retrieval users might be curious or malicious (in case of positive or negative query contexts, respectively). Patient EQI system users might be worried about their privacy or interested in monetization of their content.

3.2.2 Query collections. EQI assumes the existence of an exhaustive query collection. There are a number of approaches to creating such a collection. First, we can use an existing log from the document retrieval system that contains all the queries ever issued by searchers. The advantage of this approach is that the queries will be realistic. However, the model will not be able to capture

[1] Similar two-sided metrics have previously been used in the task of query suggestion: the success of a query prediction system depends on the quality of the query ranking but also on the quality of the document ranking each query retrieves [33]. EQI requires a similar multiplicative metric; unlike the query prediction metric, however, an exposure set quality metric needs to: (i) quantify whether a query *exposes* a document rather than whether the document is relevant; and (ii) incorporate two different models of user behavior (EQI and document retrieval system users).

instances of problematic exposure [4, 5, 40] by unseen queries. Synthetically generating a collection of queries is an alternative. To limit the collection size when this approach is adopted, queries can be truncated at a certain length and further pruned according to various frequency heuristics [1, 4, 7]. The approach might allow the system to detect problematic exposure before it occurs, but it will prevent the system from surfacing exposure to many viable queries. A practitioner should choose an approach depending on whether it is important to report the common or worst-case exposure contexts in a given application.

4 EQI FOR EMBEDDING-BASED SEARCH SYSTEMS

The first family of retrieval systems that we consider in this work are embedding-based search systems, also referred to as dual-encoder systems, that learn independent vector representations of query and document such that their relevance can be estimated by applying a similarity function over the corresponding vectors. Let $f_Q : \mathcal{Q} \rightarrow \mathbb{R}^n$ and $f_D : \mathcal{D} \rightarrow \mathbb{R}^n$ be the query and document encoders, respectively, and $\otimes : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be the vector similarity operator. The score $s_{q,d}$ for a query-document pair is then given by,

$$s_{q,d} = f_Q(q) \otimes f_D(d) \tag{5}$$

Ideally, $s_{q,d_+} > s_{q,d_-}$, if d_+ is more relevant to the query q than d_-. While $\otimes$ is typically a simple function, such as dot-product or cosine similarity, the encoding functions f_Q and f_D are often learned by minimizing an objective of the following form:

$$\mathcal{L}_{q \rightarrow d} = \mathbb{E}_{q,d_+,d_- \sim \theta}[\ell_{q \rightarrow d}(s_{q,d_+} - s_{q,d_-})] \tag{6}$$

Commonly used functions for the instance loss $\ell_{q \rightarrow d}$ include hinge [17] or RanknNet [6] loss. For example, in case of RankNet:

$$\ell_{q \rightarrow d}(\Delta) = \log(1 + e^{-\Delta}) \tag{7}$$

Metric learning for EQI. For EQI, we want to learn a similarity metric such that a query q^+ should be more similar to document d than q^-, if d has a higher probability of exposure to the user on the search result page σ_{q^+} compared to σ_{q^-}. Similar to the embedding-based model for document retrieval, we now define two new encoders $h_Q : \mathcal{Q} \rightarrow \mathbb{R}^n$ and $h_D : \mathcal{D} \rightarrow \mathbb{R}^n$ for query and document, respectively, and write our new optimization objective as follows:

$$\mathcal{L}_{d \rightarrow q} = \mathbb{E}_{d,q_+,q_- \sim \theta}[\ell_{d \rightarrow q}(u_{d,q_+} - u_{d,q_-})] \tag{8}$$

Where, $u_{d,q} = h_Q(q) \otimes h_D(d)$. If we assume that exposure of a document depends only on the rank position at which it is displayed on a result page, then q_+ and q_- should be sampled such that $\rho(d, \sigma_{q_+}) < \rho(d, \sigma_{q_-})$—*i.e.*, d is displayed at a higher rank position for q_+ than q_-. Let, $X_{q \rightarrow d} = \{f_Q(q) | q \in \mathcal{Q}\} \bigcup \{f_D(d) | d \in \mathcal{D}\}$ and $X_{d \rightarrow q} = \{h_Q(q) | q \in \mathcal{Q}\} \bigcup \{h_D(d) | d \in \mathcal{D}\}$. Then, $(X_{q \rightarrow d}, \otimes)$ and $(X_{d \rightarrow q}, \otimes)$ define two distinct metric spaces corresponding to the embedding-based document retrieval system and exposing query retrieval system, respectively. To retrieve exposing queries for a document d, we need to perform a reverse nearest neighbor (rNN) search in the $(X_{q \rightarrow d}, \otimes)$ metric space. However, rNN solutions are inefficient for high-dimensional and dynamic models. We instead propose to learn a new metric space $(X_{d \rightarrow q}, \otimes)$ where a NN search would approximate the rNN search in the $(X_{q \rightarrow d}, \otimes)$ space.

4.1 Training data generation

Generating training data is a key challenge in this learning task. To train the model, we need a dataset of documents and their exposing queries. We thus need to devise a training data generation strategy which ensures that: (1) we only issue a limited number of queries to the document retrieval model to keep the approach computationally efficient, (2) we have at least one exposing query for each training document, and (3) the training data represents the distribution of queries around each training document. The pseudocode for the proposed procedure is outlined in Algorithm 1. Training data is ensured to be representative and high-quality under strict efficiency constraints as we only conduct document retrieval for the training queries (line 5). By selecting training documents from the candidate set $\mathcal{D}_{\text{candidate}}$ (line 9), we guarantee that each training document has at least one exposing query that retrieves it in top ranks. We then do a KNN search from Q_{train} of the queries closest to each training document in the document retrieval embedding space (line 12). This gives us a sample distribution of queries around each document and the positive and negative document-query pairs will match the current errors in the document retrieval embedding space (where similarity between document and query does not necessarily reflect the exposure).

Note that from Ψ_{train}, as constructed in Algorithm 1, we can sample two types of training instances of form $< d, q^+, q^- >$, corresponding to: Case 1: $\rho(d, \sigma_{q^+}) < \rho(d, \sigma_{q^-}) < n_{q \rightarrow d}$, and Case 2: $\rho(d, \sigma_{q^+}) < n_{q \rightarrow d} \leq \rho(d, \sigma_{q^-})$. During training, we randomly sample training instances for Case 1 and Case 2 with probability α and $1 - \alpha$, respectively. Intuitively, Case 1 helps the model better rank the exposing queries already retrieved by the document retrieval embedding space, while Case 2 helps the model learn to retrieve exposing queries that are far away from the document in the starting document embedding space.

4.2 Model architecture

We adopt the publicly available[2] BERT-based dual-encoder architecture by Xiong et al. [42] as our dual-encoder search system. The model is trained in an active metric learning setting [12, 21, 43], optimizing towards the so-called approximate nearest neighbor negative contrastive estimation (ANCE) objective. As far as we are aware, the model trained on the ANCE loss has demonstrated a state-of-the-art performance by a dual-encoder system. We therefore adopt this model as our base architecture for both the document and the exposing query retrieval stages. Unless specifically mentioned, we use the same hyperparameters as in the original paper.

We anticipate that both models—for the document retrieval and the exposing query retrieval tasks—need to learn a fundamental notion of topical relevance between queries and passages that are largely similar in both cases. Therefore, we pre-initialize the exposing query retrieval model with the learned parameters of the document retrieval model before training. Furthermore, we add a few additional layers on top of the query and document encoders. However, efficiency constraints limit the amount of data we can use for training the extra layers. To avoid overfitting, we design the network architecture to only slightly transform the original embedding space. In particular, we explore following two settings:

[2] https://github.com/microsoft/ANCE

Algorithm 1 Training data generation

Randomly sample Q_{train} from Q
Initialize $\mathcal{D}_{\text{candidate}} = \emptyset$
Initialize cache results $C = \{\}$
for all $q_{\text{train}} \in Q_{\text{train}}$ **do**
  Retrieve top $n_{q \to d}$ ranked list of documents $\sigma_{q_{\text{train}}}$ from $\mathcal{D}$
  Cache retrieval results $C[q_{\text{train}}] = \sigma_{q_{\text{train}}}$
  $\mathcal{D}_{\text{candidate}} \leftarrow \mathcal{D}_{\text{candidate}} \bigcup n_{q \to d}$ documents in $\sigma_{q_{\text{train}}}$
end for
Randomly sample $\mathcal{D}_{\text{train}}$ from $\mathcal{D}_{\text{candidate}}$
Initialize training dataset $\Psi_{\text{train}} = \{\}$
for all $d_{\text{train}} \in \mathcal{D}_{\text{train}}$ **do**
  Retrieve top $n_{d \to q}$ ranked list of queries $\psi_{d_{\text{train}}}$ from Q_{train}
  Initialize $\beta_{d_{\text{train}}} = \{\}$
  for all $q \in \psi_{d_{\text{train}}}$ **do**
    if $q \in C[q]$ **then**
      Get document rank $\rho(d_{\text{train}}, \sigma_q)$ from $C[q] = \sigma_q$
    else
      Label $\rho(d_{\text{train}}, \sigma_q)$ as ∞
    end if
      $\beta_{d_{\text{train}}}[q] = \rho(d_{\text{train}}, \sigma_q)$
  end for
  Add to the training dataset $\Psi_{\text{train}}[d_{\text{train}}] = \beta_{d_{\text{train}}}$
end for
return Ψ_{train}

Append: Let $\vec{x}$ be the vector output of the base encoder, which for our base model has a size of 768. Then, under the Append settings, the final embedding output is $\vec{y} = (\vec{x}, \text{FF}(\vec{x}))$, where $(,)$ denotes a concatenation of two vectors, and FF is a feedforward network. Specifically, in our experiments we employ a four-layer feedforward neural network, with hidden layer sizes of 64 and the last output dimension as 32. Consequently, after concatenation the final embedding vector dimensions is 800. For each hidden layer of the sub-network, we apply a ReLU nonlinearity, a dropout layer, and a layer normalization.

Residual: Using a similar notation, the output of the residual layer [16] can be written as $\vec{y} = \vec{x} + \text{FF}(\vec{x})$. The feedforward network FF in this case comprises a single hidden layer of size 384 and an output vector size of 768. We add ReLU nonlinearity, dropout, and layer normalization similar to the Append setting.

4.3 Training and retrieval

We train all models, including baselines, for 1K iterations, where each iteration consists of 100 batches with 1000 samples per batch. We use dot product as the similarity measurement as in the ANCE model. We use a pairwise optimization objective in the form of the cross entropy loss. We set α, as introduced in Section 4, to 0.5 and the dropout rate to 0.1. We use Adam optimizer with a fixed learning rate of 1e-4—and set the coefficients beta1 to 0.9, beta2 to 0.999, and eps to 1e-08. Unless otherwise specified, all results reported in Section 7 are obtained from models trained on 50K queries and 200K passages.

For experimentation agility under limited GPU budget, we freeze the base encoders and only optimize the parameters of the FF layers on top—under both Append and Residual settings—when training for the exposing query retrieval task. We expect that additional improvements may be possible from fine-tuning the encoders end-to-end but we do not explore that direction in our current work.

We use the Faiss IndexFlatIP Index [20], which guarantees efficient computation of exact nearest neighbors, for nearest neighbor search in the context of both our baseline and proposed models, as well as for the ground truth results generation.

5 EQI FOR TRADITIONAL SEARCH SYSTEMS

We also study EQI in the context of BM25 [32], an important traditional retrieval model. BM25 estimates the query-document relevance score as a function of how often the term occurs in the text (term frequency or TF) and some measure of how discriminative the term is—*e.g.*, inverse-document frequency (IDF) [31].

$$s_{q,d} = \sum_{t \in q} IDF(t) \cdot \frac{TF(t,d) \cdot (k_1 + 1)}{TF(t,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{\text{AvgDL}})} \quad (9)$$

Where, k_1 and b are parameters of the model that can be tuned and AvgDL is the average document length in the collection. A simple approach to retrieving exposing queries in the context of BM25 may therefore involve simply exchanging the role of query and documents under this setting. Instead of precomputing term-document scores and indexing the document collection, we can instead precompute the term-query scores and index the query log. Given the weighted term vector corresponding to a document, we can retrieve queries from the prepared index. Real life IR systems often make practical assumptions about the queries being typically short and documents being longer in their design to achieve more efficient retrieval. While we may need to work around some of these system constraints, the general framework for reversing a BM25-based system seems straightforward. The term weighting functions can be further tuned, if necessary, for the exposing query retrieval setting independently of the scoring function employed for document retrieval.

5.1 Indexing and retrieval

In our work, we adopt the Anserini toolkit [44] to index the query log and issue documents as queries. We tune the model parameters, observing that the impact is limited under the reversed setting as the term-saturation is applied on the short indexed queries. We foresee future work that applies the term-saturation on the input to the retrieval system instead, but that consequently requires the retrieval system to support explicit term-weighting.

6 EXPERIMENTS

6.1 Data

For EQI experimentation, we need a large-scale dataset accompanied by a large query collection. We thus use the MS MARCO passage retrieval benchmark [2], extensively used to compare neural methods for retrieval, including in the TREC Deep Learning track [10]. The MS MARCO dataset contains 8,841,823 passages and 532,761 relevant query-passage pairs for training, as labeled by human assessors.

6.2 Methods

Brute-Force (Exact EQI) Baseline. An obvious baseline for our work corresponds to the brute-force approach of running every

Table 2: Model performance on the EQI task benchmarked using the MS MARCO dataset. We report the RELQ metric corresponding to four different user model combinations.

Model	$RELQ_{RBP,RBP}$ with $(\gamma_{q\to d}, \gamma_{d\to q})$ (0.5, 0.5)	(0.5, 0.9)	(1, 1)	$RELQ_{EXH,NDCG}$
BM25-reverse	0.441	0.624	0.840	0.645
BM25-tuned	0.442	0.626	0.845	0.648
Brute force	1.000	1.000	1.000	1.000

(a) EQI for BM25

Model	$RELQ_{RBP,RBP}$ with $(\gamma_{q\to d}, \gamma_{d\to q})$ (0.5, 0.5)	(0.5, 0.9)	(1, 1)	$RELQ_{EXH,NDCG}$
ANCE-reverse	0.493	0.685	0.887	0.698
ANCE-append	0.624	0.825	0.982	0.837
ANCE-residual	0.633	0.834	0.984	0.845
Brute force	1.000	1.000	1.000	1.000

(b) EQI for ANCE

query in our log through the document retrieval system and recording all the retrieved results. Then for a given passage, we can iterate over all the document retrieval results to produce an exact solution to the exposing query retrieval problem. While the brute force method achieves perfect recall of all exposing queries, it also incurs both a heavy computation cost for the document retrieval step.

Original Retrieval Space Methods. For traditional search systems, an intuitive baseline model can index the query log as a document collection and use documents as queries. We refer to this method as BM25-reverse. We also tune the term weighting parameters k_1 and b, which we refer to as BM25-tuned. In the context of dual-encoder search systems, we expect that, even in the document retrieval embedding space, exposing query and document pairs would be relatively close to each other, and a nearest neighbor search around a passage in this space would identify some of its exposing queries. We employ this approach as a baseline, which we refer to as the ANCE-reverse in the remainder of this paper.

Metric learning. We also implement the metric learning approaches described in Sec. 4, referring to the two architectures as ANCE-append and ANCE-residual.

6.3 Ground truth end experimental protocol

In our experiments, we set $n_{q\to d} = 100$ and $n_{d\to q} = 100$. To obtain the ground truth exposing queries we obtained a ranked list σ_q^* of the top 100 passages for each query q in the collection. To evaluate the EQI models, for each model and each passage d in the test set of size 20,000, we retrieve the top 100 queries as the exposing ranked list ψ_d.

7 RESULTS

7.1 Overall performance

Table 2 summarizes the EQI performance results on the MS MARCO benchmark, using different parametrizations of the RELQ metric. The brute force baseline, by definition finding all exposing queries, achieves a perfect value of the RELQ metric. In the context of traditional search models, BM25-tuned consistently shows a slightly better performance than BM25-reverse according to all metrics. In

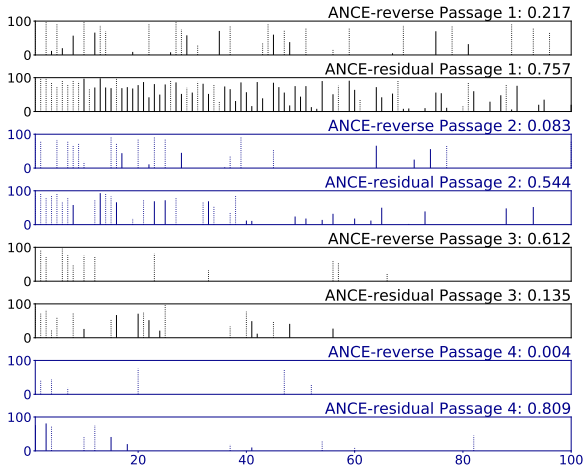


Figure 1: Visualization of retrieved exposing query lists. The figure presents retrieved ranked lists of exposing queries for four test passages. Dotted lines represent the exposing queries retrieved by both models. The number reported is the $RELQ_{RBP,RBP}$ with $\gamma_{q\to d} = 0.5, \gamma_{d\to q} = 0.9$. The x-axis represents the ranking position in the exposing query retrieval model $\rho(q, \psi_d)+1$ and the y-axis represents the rank at which the query retrieves the test passage in the document retrieval system $n_{q\to d} - \rho(d, \sigma_q)$ (the higher, the more exposing).

the context of dual-encoder models, both the proposed metric learning approaches (ANCE-append and ANCE-residual) significantly outperform the ANCE-reverse under different parametrizations of RELQ. The improvements of both models against ANCE-reverse are statistically significant according to a one-tailed paired t-test ($p < 0.01$). Of the two models, Residual performs better than Append, and the difference in their performance are again statistically significant based on one-tailed paired t-test ($p < 0.01$). Therefore, we only use the Residual model in the remaining analyses.

Illustration. Beyond the mean metric values, we illustrate the behavior of ANCE-reverse and ANCE-residual on four selected test passages in Figure 1. These passages were selected to reflect different performance levels. For each passage, we compare the ranked exposing query lists retrieved by the ANCE-Reverse and the Residual models. Recall that $n_{d\to q} = 100$. Vertical bars denote exposing queries, with the dotted vertical bars showing queries retrieved by both models. The x-axis represents the ranking position of the query in the EQI model, while the y-axis represents the ranking position at which the document retrieval system retrieves the passage for a given query. More precisely, we plot $n_{q\to d} - \rho(d, \sigma_q)$, such that higher bar indicates that the document was retrieved at a higher rank for that query.

An ideal ranked list should arrange the vertical bars from the highest (queries exposing the passage at top ranking positions) to the lowest with no gap in between. Except for Passage 3, the Residual model retrieves more exposing queries and ranks them higher than the baseline Reverse model. This shift in exposing queries in the ranked lists indicates that the proposed learning

Table 3: Impact of the training sample size. Residual model performs better when either the sample size of training queries or training passages increases.

		Passages		
		200, 000	400, 000	600, 000
Queries	50, 000	0.834	0.854	0.859
	250, 000	0.872	0.877	0.880
	500, 000	0.879	0.882	0.883

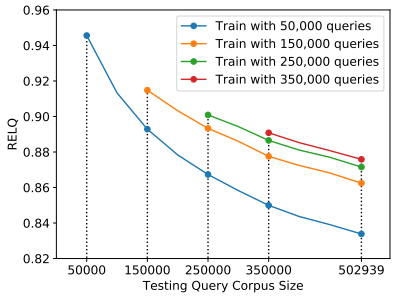


Figure 2: Each line is generated from a Residual model trained on a particular size of training queries. The starting point of each line is to train and test on the same query log. Along the line, query log is gradually expanding.

algorithm can learn a new metric space that can recover more exposing queries.

For Passage 1 and Passage 2, the Residual model furthermore retrieves many exposing queries which are originally not within the 200 nearest neighbors under the reverse baseline model. This example illustrates clearly that the nearest neighbor problem and the reverse nearest neighbor problem are not symmetric.

Visualization of Passages 3 and 4 illustrates that, when a passage has fewer exposing queries overall, the RELQ metric tends to be more sensitive to the ranks $\rho(q, \psi_d)$ of highly exposing queries. This observations may, however, change under different user behavior assumptions for the RELQ metric. We analyze this aspect in the Appendix A.1.

7.2 Impact of the training data size

Approximate EQI trades exactness for performance; in the context of dual-encoder models, performance depends on the ability to train with little data. We examine the impact of training data size on model performance. Table 3 shows the results for the Residual model trained for 1000 epochs. We report $RELQ_{RBP,RBP}$ with $\gamma_{q\to d} = 0.5, \gamma_{d\to q} = 0.9$. We find that increasing the number of both queries and training passages leads to diminishing returns. One-tailed paired t-tests are conducted along each column and each row. Except for the case, where the number of training queries is fixed at 500K and the number of training passages is increased from 400K to 600K, all increases in performance are statistically significant ($p < 0.01$) after Bonferroni correction. The Residual model trained on even a relatively small training data—50K queries and 200K passages—shows good performance.

7.3 Impact of expanding query collections

The search ecosystem is dynamic. Searchers may issue previously unknown queries, while content producers may create new documents. Note that the evaluation in Section 7.1 already covers the scenario in which exposure needs to be computed for new documents, as none of the test passages were seen or present in the corpus during training. We additionally examine the effects of a growing query collection.

Our goal is to see how long the EQI model can sustain good retrieval performance without retraining as the query collection grows. This process is simulated by first sampling an initial set of queries available during training, and then test using a gradually expanding collection by embedding new queries without model retraining. The number of training passages is fixed as 200,000. Four Residual models are then trained for 1000 epochs on different sizes of a starting query collections: 50,000, 150,000, 250,000 and 350,000. Note that this experimental protocol implies that new queries are drawn from the same distribution as existing queries. Simulation results are illustrated in Figure 2. We report $RELQ_{RBP,RBP}$ with $\gamma_{q\to d} = 0.5, \gamma_{d\to q} = 0.9$.

Decreasing concave trend along each line: The decrease in model performance generally signals the need for model retraining. In different real-world applications, service providers may have different tolerance levels to performance loss vs the retraining costs. A visualization like the one in Figure 2 can be used as a monitoring tool for deciding when to retrain as the query collection expands.

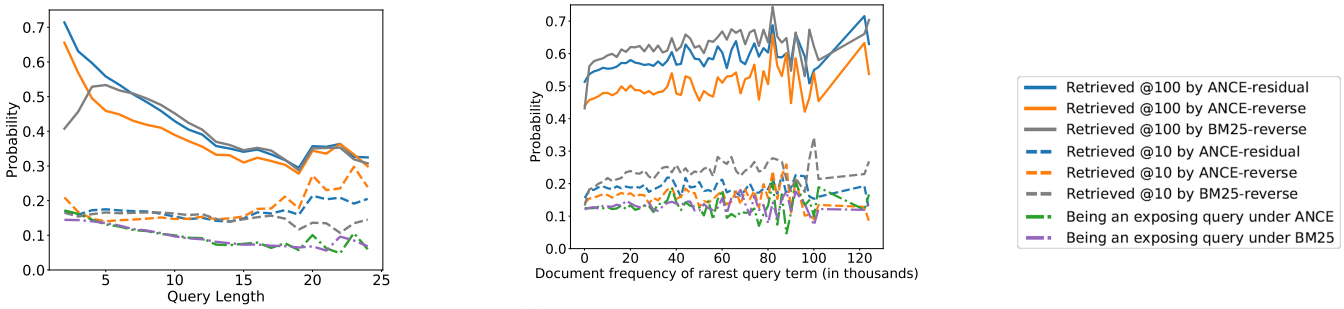
The concave shape indicates that the model performance stabilizes. Recall that ANCE-Reverse achieves a RELQ of only 0.685 on the full query collection. We reckon that the Residual model might eventually converge to a performance level still outperforming the Baseline model should the query collection increase further.

Decreasing rate drop with larger training sample size: The performance curves are flatter for models initially trained with more queries. These models capture the characteristics of a fully expanded query collection better.

Increasing task difficulty: Both the decreasing trend along each curve and the decreasing starting value for each curve might suggest that the difficulty of the EQI task increases with the size of the query collection. Interestingly, an opposite observation has been made about document retrieval systems by Hawking and Robertson [15] who showed that the document retrieval precision increases with increasing collection size due to the sparsity of relevant documents in sub samples. Reconciling this difference requires further analysis. We hypothesize that the difference may be due to the fact that Hawking and Robertson [15] performed their analysis on term-matching-based retrieval models, whereas recent evidence shows that in the context of neural models an increase in the collection size may lead to a decrease in retrieval effectiveness [46].

7.4 Impact of query characteristics

Probability of retrieval vs. query length: We investigate whether the EQI models are more likely to retrieve longer queries. As shown in Figure 3a, under both BM25 and ANCE, the probability of being an exposing query for one of the test passages is slightly higher in case of shorter queries, *i.e.*, queries with fewer than 10 words.



(a) By query length

(b) By document frequency of rarest query term

Figure 3: Probability of being an exposing query for one of the test passages vs probability of retrieval.

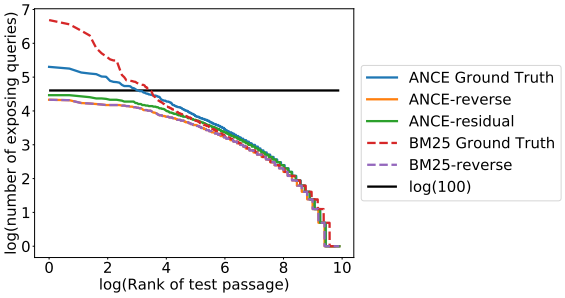


Figure 4: Rank-size distribution of the number of ground truth and retrieved exposing queries per passage. A few passages have a large number of exposing queries but there is also a long tail of passages with few or no exposing queries.

As for the top-10 retrieval probability, we find that both ANCE-reverse and ANCE-residual tend to slightly under-retrieve shorter queries and over-retrieve longer queries, while BM25-reverse retrieves equally regardless of query length. However, for the top-100 retrieval probability, we observe that all models are likely to significantly over-retrieve queries, in particular shorter ones.

Probability of retrieval vs. query term frequency: We investigate whether the models are able to retrieve both rare and common queries. We measure the rarity of a query by computing the minimum document frequency of the query terms. This analysis sheds a light on the usefulness of the approach in providing privacy-awareness, as rare queries can contain personally identifying information. Figure 3b demonstrates that ANCE-reverse, ANCE-residual and BM25-reverse tend to retrieve such queries at a rate roughly proportional to their probability of being exposing for one of the test passages.

Distributions of exposing queries per passage: Figure 4 shows a log-log plot of the distribution of the number of exposing queries (y-axis) per passage. For all the methods, we find that there is a long tail of passages with very few or no exposing queries, while some passages may have tens, or even hundreds, of exposing queries. This result suggests the strong retrievability bias [1], and is in line with prior work on search exposure [4].

7.5 Discussion

The results reported earlier in this section indicate that reversing the role of queries and documents can be a reasonable lightweight strategy for EQI, in particular when the model parameters are further optimized for the query retrieval task in the form of metric learning. However, deploying an EQI system in the wild requires several other practical considerations. While Section 7.2 tells us that we can achieve reasonable effectiveness in exposing query retrieval using moderately sized training datasets (50K queries), the required amount of training will depend on the size of the query corpus from which we are retrieving, as seen in Section 7.3. As we can produce almost infinitely large training data for the query retrieval model automatically using the document retrieval model, the size of required training dataset itself may not be a big concern but may still have implications for training time of the query retrieval model and thus the attractiveness of EQI as a lightweight alternative to brute force aggregation.

Another important concern may be around potential systemic biases in retrieved queries. In Section 7.4, we notice that the methods we study retrieve queries roughly in proportion to their probability of being exposing queries independent of query term rarity which is promising in terms of our ability to retrieve tail queries, and thus downstream privacy applications. However, these models disproportionately over-retrieve shorter queries which may have implications for downstream applications such as quantifying the retrievability bias, or enforcing the right to be forgotten. Studying these effects is an interesting direction for future work.

8 CONCLUSIONS

This paper focused on the search transparency problem of exposing query identification, that is, finding queries that expose documents in the top-ranking results of a search system. We studied the feasibility of approximate EQI as a retrieval problem for two classes of search systems (dual-encoder and traditional BM25 systems). We further demonstrated the potential of improving baseline EQI approaches through metric learning. We also derived an evaluation metric called Ranked Exposure List Quality (RELQ) that jointly models the behavior of two users who influence the EQI outcomes: a user of the EQI system and the document retrieval user.

Many promising future directions remain to be explored. Examples include developing better network architectures, algorithms for EQI in black-box search systems or for document retrieval models that use external click signals for ranking, developing generative models for query collections better grounded in downstream exposure tasks, as well as domain-specific quantitative and qualitative application and user studies. Our work provides a conceptual and empirical foundation for further study of EQI, which can contribute to advances in various areas of search and ranking: bias and fairness, privacy, data protection, security, and SEO.

REFERENCES

[1] Leif Azzopardi and Vishwa Vinay. 2008. Retrievability: an evaluation measure for higher order information access tasks. In *Proc. CIKM*. ACM, 561–570.

[2] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268* (2016).

[3] Theo Bertram, Elie Bursztein, Stephanie Caro, Hubert Chao, Rutledge Chin Feman, Peter Fleischer, Albin Gustafsson, Jess Hemerly, Chris Hibbert, Luca Invernizzi, et al. 2019. Five Years of the Right to be Forgotten. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 959–972.

[4] Asia J Biega, Azin Ghazimatin, Hakan Ferhatosmanoglu, Krishna P Gummadi, and Gerhard Weikum. 2017. Learning to Un-Rank: quantifying search exposure for users in online communities. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 267–276.

[5] Joanna Asia Biega, Krishna P Gummadi, Ida Mele, Dragan Milchevski, Christos Tryfonopoulos, and Gerhard Weikum. 2016. R-susceptibility: An IR-centric approach to assessing privacy risks for users in online communities. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 365–374.

[6] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*. ACM, New York, NY, USA, 89–96.

[7] Jamie Callan and Margaret Connell. 2001. Query-based sampling of text databases. *ACM Transactions on Information Systems (TOIS)* 19, 2 (2001), 97–130.

[8] Ioannis Chios and Suzan Verberne. 2021. Helping results assessment by adding explainable elements to the deep relevance matching model. *arXiv preprint arXiv:2106.05147* (2021).

[9] Jaekeol Choi, Jungin Choi, and Wonjong Rhee. 2020. Interpreting neural ranking models using grad-cam. *arXiv preprint arXiv:2005.05768* (2020).

[10] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2020. Overview of the TREC 2019 deep learning track. In *TREC*.

[11] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating Stochastic Rankings with Expected Exposure. *arXiv preprint arXiv:2004.13157* (2020).

[12] Sandra Ebert, Mario Fritz, and Bernt Schiele. 2012. Active metric learning for object recognition. In *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*. Springer, 327–336.

[13] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2221–2231.

[14] Tim Gollub, Matthias Hagen, Maximilian Michel, and Benno Stein. 2013. From keywords to keyqueries: content descriptors for the web. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 981–984.

[15] David Hawking and Stephen Robertson. 2003. On collection size and retrieval effectiveness. *Information retrieval* 6, 1 (2003), 99–105.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[17] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 2000. Large margin rank boundaries for ordinal regression. (2000).

[18] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proc. CIKM*. ACM, 2333–2338.

[19] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* (2019).

[21] Krishnan Kumaran, Dimitri Papageorgiou, Yutong Chang, Minhan Li, and Martin Takáč. 2018. Active metric learning for supervised classification. *arXiv preprint arXiv:1803.10647* (2018).

[22] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*. 2243–2251.

[23] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.* 27, Article 2 (December 2008), 27 pages. Issue 1. https://doi.org/10.1145/1416950.1416952

[24] Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. 2016. Improving Document Ranking with Dual Word Embeddings. In *Proc. WWW*.

[25] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375* (2019).

[26] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 6620–6631.

[27] Sai Teja Peddinti, Aleksandra Korolova, Elie Bursztein, and Geetanjali Sampemane. 2014. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In *2014 IEEE Symposium on Security and Privacy*. IEEE, 493–508.

[28] Jeremy Pickens, Matthew Cooper, and Gene Golovchinsky. 2010. Reverted indexing for feedback and expansion. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1049–1058.

[29] Sayantan Polley, Rashmi Raju Koparde, Akshaya Bindu Gowri, Maneendra Perera, and Andreas Nuernberger. 2021. Towards Trustworthiness in the context of Explainable Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2580–2584.

[30] Nimrod Raifer, Fiana Raiber, Moshe Tennenholtz, and Oren Kurland. 2017. Information retrieval meets game theory: The ranking competition between documents' authors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 465–474.

[31] Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation* 60, 5 (2004), 503–520.

[32] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[33] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2013. Learning to rank query suggestions for adhoc and diversity search. *Information Retrieval* 16, 4 (2013), 429–451.

[34] Jaspreet Singh and Avishek Anand. 2018. Interpreting search result rankings through intent modeling. *arXiv preprint arXiv:1809.05190* (2018).

[35] Jaspreet Singh and Avishek Anand. 2018. Posthoc interpretability of learning to rank models using secondary training data. *arXiv preprint arXiv:1806.11330* (2018).

[36] Jaspreet Singh and Avishek Anand. 2019. Exs: Explainable search using local model agnostic interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 770–773.

[37] Paul Thomas, Bodo Billerbeck, Nick Craswell, and Ryen W White. 2019. Investigating searchers' mental models to inform search explanations. *ACM Transactions on Information Systems (TOIS)* 38, 1 (2019), 1–25.

[38] Ziv Vasilisky, Moshe Tennenholtz, and Oren Kurland. 2020. Studying Ranking-Incentivized Web Dynamics. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2093–2096.

[39] Manisha Verma and Debasis Ganguly. 2019. LIRME: locally interpretable ranking model explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1281–1284.

[40] Washington Post. 2018. Facebook: 'Malicious actors' used its tools to discover identities and collect data on a massive global scale. https://www.washingtonpost.com/news/the-switch/wp/2018/04/04/facebook-said-the-personal-data-of-most-its-2-billion-users-has-been-collected-and-shared-with-outsiders/. Online; accessed 3 February 2021.

[41] Colin Wilkie and Leif Azzopardi. 2017. Algorithmic bias: do good systems make relevant documents more retrievable?. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2375–2378.

[42] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[43] Liu Yang, Rong Jin, and Rahul Sukthankar. 2012. Bayesian active distance metric learning. *arXiv preprint arXiv:1206.5283* (2012).

[44] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of Lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1253–1256.

[45] Yin Yang, Nilesh Bansal, Wisam Dakka, Panagiotis Ipeirotis, Nick Koudas, and Dimitris Papadias. 2009. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. 34–43.

[46] Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From Neural Re-Ranking to Neural Ranking: Learning a Sparse Representation for Inverted Indexing. In *Proc. CIKM*. ACM, 497–506.
[47] Honglei Zhuang, Xuanhui Wang, Michael Bendersky, Alexander Grushetsky, Yonghui Wu, Petr Mitrichev, Ethan Sterling, Nathan Bell, Walker Ravina, and Hai Qian. 2020. Interpretable Learning-to-Rank with Generalized Additive Models. *arXiv preprint arXiv:2005.02553* (2020).
[48] Justin Zobel and Alistair Moffat. 2006. Inverted files for text search engines. *ACM computing surveys (CSUR)* 38, 2 (2006), 6.

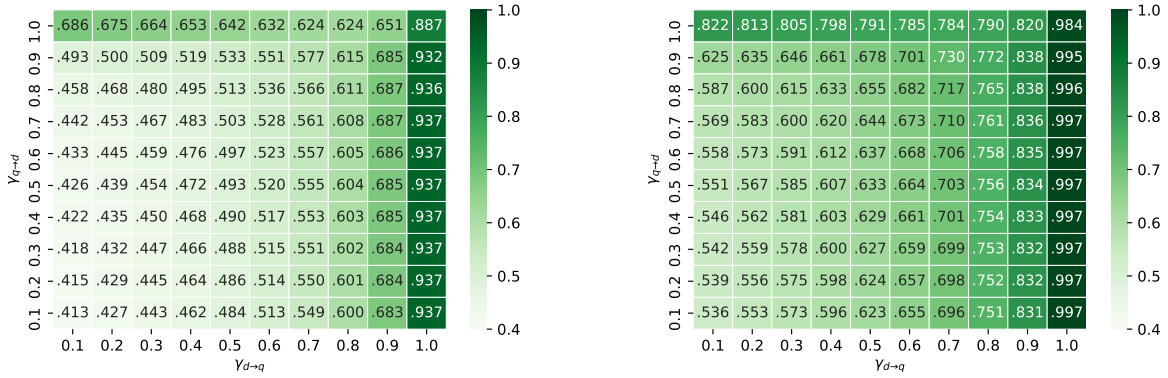
A APPENDIX

A.1 Impact of metric parametrization

Under different user behavior assumptions, the same exposing query list will achieve different absolute value of RELQ. We explore the relationship between user behavior and model performance for different combinations of user patience parameters $\gamma_{q\to d}$ (document retrieval user) and $\gamma_{d\to q}$ (EQI user). Figure 5 presents the resulting performance heatmaps for ANCE-reverse and ANCE-residual. Both heatmaps exhibit similar patterns. Higher $\gamma_{q\to d}$ generally correlates with a high value of RELQ, except for very high values of $\gamma_{d\to q}$, such as 0.9 or 1.0, where a higher value of RELQ is observed for lower values of $\gamma_{q\to d}$.

These patterns are expected. Having a more patient user of the document retrieval system (high $\gamma_{q\to d}$) will mean more exposing queries per document: as the user goes deeper down the rank list more documents will get exposed to the query. Intuitively then, when the user of the EQI system is impatient, the presence of many ground truth exposing queries makes the task easier. However, when the EQI user is very patient and inspects the exposure set exhaustively, fewer ground truth exposing queries (impatient document retrieval user) do not lead to a higher task difficulty, and in this case, we indeed observe the highest absolute RELQ values.

ANCE-residual outperforms the ANCE-reverse model under all metric parametrizations. Notably, however, ANCE-reverse performs very well for patient EQI users ($\gamma_{d\to q} = 1.0$). This result suggests that even though both models are able to retrieve many exposing queries, the ANCE-residual model tends to rank the exposing queries higher.



(a) ANCE-reverse **(b) ANCE-residual**

Figure 5: Model performance (RELQ values) for different user behavior models (the EQI system user on the x-axis, and the document retrieval user on the y-axis). The darker the color, the better the performance.