On the Explainability of Exposing Query Identification

Amin Abolghasemi
Leiden University
Leiden, The Netherlands
m.a.abolghasemi@liacs.leidenuniv.nl

Suzan Verberne
Leiden University
Leiden, The Netherlands
s.verberne@liacs.leidenuniv.nl

Leif Azzopardi
University of Strathclyde
Glasgow, UK
leifos@acm.org

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

Exposing query identification (EQI) is the task of identifying queries for which it is likely that a retrieval system shows a given document in its ranked list of retrieved documents. The purpose of EQI mainly fits into the context of search transparency: to make the user aware of queries for which a document they are about to publish, e.g., a tweet, would be retrieved in the top-ranked results of a search engine. In this paper, we propose a method to make EQI results explainable. The motivation for this is to enable users to better understand why a query in the set of exposing queries (i.e., the output of the EQI system) exposes their content. As the first work addressing explanations for EQI, we use query expansion as a form of explanation, following prior work on explainability in document retrieval. In addition, we propose an evaluation framework for measuring the fidelity of explanations in expansion-based methods. Our evaluation using three retrieval models and two query expansion methods shows that these expansion methods fall short of achieving an ideal performance when being evaluated based on our proposed fidelity measures. Moreover, we find that the fidelity of explanations varies across different retrieval models. Our study contributes to search transparency and facilitates future work on explainability of EQI systems. Furthermore, our work offers insights into challenges regarding quantitatively assessing the quality of explanations in information retrieval systems.

KEYWORDS

Explainability, Transparency, Exposing Queries, Evaluation

1 INTRODUCTION

Exposing query identification (EQI) is the task of retrieving a list of queries that will expose a given document [3, 4, 13]. Showing these queries to a content creator can warn them about potentially sensitive content being exposed by a search engine. An *exposing query* for a document is a query for which the document will be shown to a user in the top-k of a ranked list of documents retrieved by a document retrieval system. See Figure 1 for an illustration.

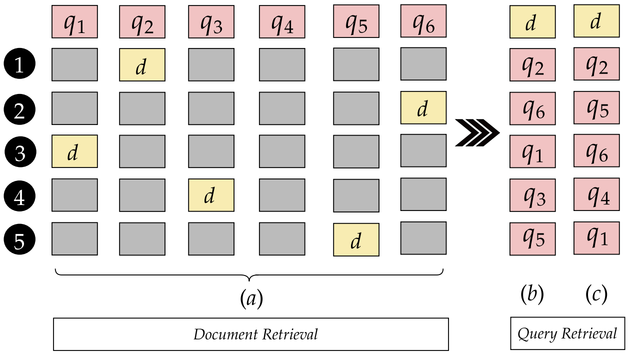

Figure 1: Document retrieval (a) and exposing query identification (b, c). The document in yellow is the same across all ranked lists in both systems. The ranked list of queries (b) is the ground truth list of queries retrieved for the document in yellow. The ranked list of queries (c) shows an example output of an EQI system.

Prior work [4, 13] defines this task in the context of the search transparency: it can help users gain a better understanding of how search engines map queries to documents, resulting in higher transparency of search engines to the user. Transparency is naturally related to explainability [9, 10, 18, 28] where the goal is for the user to understand how and why a retrieval or recommender system returns a specific result [13].

Motivation. Prior work on EQI specifically focuses on formulating the EQI task as a query retrieval problem: for a given document, retrieve, from a query log, the queries that will show the document in the ranked list. In this work, we adopt the same approach, and contribute to search transparency by studying the explainability of exposing query identification. The task we address is:

> *Given a document and a ranked list of exposing queries (retrieved by the EQI system), explain why these queries might expose the given document in the retrieval system.*

In this work, we define explainability of EQI as explaining why a query in the list retrieved by the EQI system would expose a given document in the document retrieval system. We focus on post-hoc text-based explanations. We hypothesize that such explanations could increase the usage of EQI systems, by providing transparency to users who create content [13]. Explanations could be used in a number of use-cases: (i) Content creator risk awareness: a user might not be aware of the consequences of the content they are

FAccTRec '23, September 18–22, 2023, Singapore

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Document:
For these hot days, I decided to jump with these penguins! ...

Query: *penguins hockey*
Expansion terms: *ice, club, league, pittsburgh*

Query: *arcademics*
Expansion terms: *game, penguin, jump*

Figure 2: Examples of exposing queries with their corresponding expansion terms for a given document (a tweet in this case). The expansion terms clarify to the user that the word 'penguin' has multiple meanings and the document might be exposed for queries related to the other meanings.

sharing on social media. Providing an explanation of the retrieved list of exposing queries may help to provide increased awareness to the user with regard to the potential exposure of publishing information online. For instance, an exposing query could be semantically ambiguous or have multiple intents [28]; an epxlanation should help the user understand why the exposing query would retrieve their content. (ii) EQI as recommendation: EQI can be viewed as a task with a recommendation goal in which a content creating user not only becomes aware of potential risks of their document, but also receives recommendation of possible modifications that could be applied to their document to prevent it from being exposed by a specific query, or to make it exposed to more relevant queries.

Expansion-based explanation. To provide a text-based explanation solution, we follow prior work [21] in using expansion terms for the purpose of explaining a ranked list of documents. In this approach, a set of expansion terms for a given query is shown to the user. The idea behind this is that to a user, the connection between a query and a document might not always be clear, e.g., because there is no lexical overlap between the query and the document and a search engine can still identify the document as relevant to a query because of semantic similarity. The query expansion model serves as an interpretable term-based ranking model to mimic the complex document ranking model to be interpreted [2, 14]. In other words, the expanded query terms act as an explanation for the intent perceived by the complex ranking model [2]. In contrast to prior work, we do not use expansion to explain a ranked document list but to explain a ranked query list in an EQI system. Figure 2 shows a document and two exposing queries of this document together with their expansion terms: the terms can indicate to the user that words in the document have multiple meanings and the search engine might expose the document to users who have a completely different intent in mind.

Evaluation of explanation. While there are different dimensions to evaluate the quality of explanations, we focus on *fidelity-based evaluation.* Fidelity is used to measure how well explanations reconstruct the original outcomes [2, 24]. Prior work on explainability for document ranking models has also defined the fidelity as the degree to which the explanations can replicate the ranked list for the original query. For instance, Llordes et al. [14], Singh and Anand [21] employ ranked list similarity metrics, e.g., Rank Biased Overlap (RBO), to measure the divergence between the ranked lists with and without the explanation [2].

EQI, however, is a two-sided retrieval setting [13] in the sense that the EQI system mirrors the outcome of a document ranking model, i.e., whether the query is exposing or non-exposing for a document depends on the rank at which the document retrieval system retrieves the document. For instance, if we assume that users will only examine the top k documents, then a query that retrieves the document at position k or less, is said to be exposing, otherwise it is not considered to be exposing. Consequently, the quality of a retrieved list of queries for an input document is also determined by the document retrieval system (see Section 2). This "two-sidedness"[1] imposes a new challenge on the evaluation of explanations for EQI systems: how to evaluate the fidelity of the explained query in this two-sided setting? To the best of our knowledge there is no prior work on explainability of two-sided retrieval settings. In this work, we define and provide two fidelity-based evaluation metrics with which one can quantify how well the explained queries reflect their corresponding original queries.

In summary, our main contributions are as follows:

(1) We propose explainability as an additional goal for exposing query identification (EQI) in order to provide more transparency of search systems by sketching the relation between a document and the list of exposing queries.
(2) We address the explainability of EQI using query expansion-based methods.
(3) We propose an evaluation framework for measuring the fidelity of expansion-based explanations for EQI. Our work is the first to study the evaluation of fidelity of explanations in a two-sided retrieval setting.

2 PRELIMINARIES

We provide preliminaries for two lines of prior work for this paper, namely evaluation of EQI and explainability of search engines using query expansion. Table 1 lists the notation we use.

Evaluation of exposing query identification. The performance of an exposing query identification system is evaluated using Ranked Exposure List Quality (RELQ) as a two-sided metric [13], identifying to what extent the retrieved exposing query list is the same as the ground truth exposing query list:

$$\mathrm{RELQ}(\mathcal{L}_d) = \frac{\sum_{q_i \in \mathcal{L}_d} \mu_{d \to q}(q_i, \mathcal{L}_d) \cdot g_{q_i}}{\sum_{q_j \in \mathcal{L}^*_d} \mu_{d \to q}(q_j, \mathcal{L}^*_d) \cdot g_{q_j}}. \tag{1}$$

RELQ is a multiplicative metric accounting for (i) the EQI system as $\mu_{d \to q}$ determines how a user inspects the queries in the list of retrieved exposing queries $\mathcal{L}_d$, and (ii) the document retrieval system as g_q (the gain for query q) is determined by the level of exposure the document retrieval system gives to document d when query q is being issued (see [13] for more details).

Explainability with query expansion. Query expansion has proved to be a successful method for resolving term discrepancies between a provided query and the relevant documents within a collection [7, 16, 17]. There are different approaches to expand queries such as language modeling-based approaches [12], and thesaurus-based expansion [25]. When presenting potential expansions to

[1] We follow prior work [13] in the use of the term "two-sided". It can be found in other retrieval settings, including query suggestion [13, 20], and differs from what "two-sided" commonly refers to in the context of recommender systems.

Table 1: List of notation.

Notation	Description
$\mathcal{D}$	A collection of documents
$\mathcal{Q}$	A collection of queries
d	An individual document
q	An individual query
q^e	An expanded version of query q
λ_q^d	Exposure status of document d for query q
Document retrieval	
$\mathcal{L}_q$	A ranked list of documents for query q
$\mathcal{L}_q^*$	The ideal ranked list of documents for query q
$n_{q\to d}$	Number of documents retrieved per query
$\mu_{q\to d}$	A user browsing model for inspecting $\mathcal{L}_q$
$\rho(d, \mathcal{L}_q)$	Rank of document d in $\mathcal{L}_q$
Exposing query retrieval	
$\mathcal{L}_d$	A ranked list of queries for document d
$\mathcal{L}_d^*$	The ideal ranked list of queries for document q
$\mathcal{L}_d^e$	A ranked list of explained queries for document d
$n_{d\to q}$	Number of documents retrieved per query
$\mu_{d\to q}$	A user browsing model for inspecting $\mathcal{L}_q$
$\rho(q, \mathcal{L}_d)$	Rank of query q in $\mathcal{L}_d$

users, query expansion additionally serves as an interpretable technique for rephrasing queries [2, 21].

In terms of the explanation method, our work is related to work by Singh and Anand [21] and Llordes et al. [14]. The former authors propose a model-agnostic approach to interpret the intended meaning of a query as perceived by a black-box ranker. To this aim, they expand the user query and employ a simple lexical ranker that can faithfully and accurately mimic the original ranker. Their objective is to discover a group of query expansion terms that maintain the majority of the pairwise orderings in the resulting ranked list of documents [2]. The expanded query terms are regarded as an explanation for the intent understood by the original ranking model [2]. Llordes et al. [14] introduce a similar approach; however, instead of adding expansion terms, they generate an equivalent query to faithfully approximate the performance of a neural ranker (operating on the original user query) with a sparse lexical ranker (operating on the equivalent query).

3 METHODOLOGY

In this section we describe our method for explaining the output of an EQI system. We also formulate our fidelity-based evaluation framework for explainablity of EQI.

3.1 Explanation method

We approach the explanation of a ranked list of exposing queries produced by an EQI system using expansion methods following the prior work [14, 21], as described in Section 2. Our method differs in the sense that we do not us a simple lexical ranker to mimic the original ranker. Instead, we employ the same ranking model to perform the retrieval for the expanded queries. We use terms from two query expansion-based methods: RM3 [12] and Bose-Einstein 1 (Bo1) [1].

3.2 Evaluation method

While different measures can be employed to evaluate explanations, we focus on fidelity-based evaluation. We focus on formalizing the evaluation of fidelity for the explainability methods and leave other measurements (including human evaluation) to future work.

The goal is to explain query q in $\mathcal{L}_d$ in such a way that the relevance (and consequently, the exposing status of a query for document d) does not change fundamentally when expansion terms are added. In other words, we require that the results of a document retrieval systemon the expanded queries reflect the results of the system on the original queries without explanation. We present two metrics for measuring this fidelity.

Metric 1: Exposing status preservation (ESP). For the first metric to measure the fidelity of explanations to the original document ranking results, we define a fidelity constraint based on the exposing status of a query with regard to a document. To formulate this constraint, we first define an indicator function $\lambda_q^d(k)$ that determines the exposing status of a query q for the document d, given a ranking cut-off k for the ranked lists of documents:

$$\lambda_q^d(k) = \begin{cases} 1 & 1 \leq \rho(d, \mathcal{L}_q) \leq k \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In other words, $\lambda_q^d(k)$ is equal to 1 if d is ranked in the top-k ranked list of documents retrieved for query q. Using $\lambda_q^d(k)$, we define a variable for capturing the change C in the exposing status of query q as follows:

$$C(d, q, q^e, n_{q\to d}) = \mathbb{1}[|\lambda_q^d(n_{q\to d}) - \lambda_{q^e}^d(n_{q\to d})| = 0]. \tag{3}$$

Now, the fidelity of the explained list of queries $\mathcal{L}_d^e$ to the original list $\mathcal{L}_d$ in terms of *exposing status preservation* (ESP) can be formulated as:

$$\mathrm{ESP}(\mathcal{L}_d, \mathcal{L}_d^e) = \sum_{i=1}^{n_{d\to q}} \mu_{d\to q}(q_i, \mathcal{L}_d) \cdot g_{q_i}. \tag{4}$$

Here, $\mu_{d\to q}$ is the user browsing model and g_{q_i} is the fidelity gain for a query retrieved for d in $\mathcal{L}_d$ with $\rho(d, \mathcal{L}_q)$. We use DCG as the user browsing model $\mu_{d\to q}$ in Eq. 4 to discount the absolute fidelity gain (Eq. 4) and weigh the higher ranked queries:

$$\mu_{d\to q}^{DCG}(q_i, \mathcal{L}_d) = \frac{1}{\log(i+1)}. \tag{5}$$

We estimate the fidelity of an explained query q using the preservation of its exposing status with regard to the document d:

$$g_{q_i} = C(d, q_i, q_i^e, n_{q\to d}). \tag{6}$$

Accordingly, the *fidelity* of the explained list of queries $\mathcal{L}_d^e$ to the original list $\mathcal{L}_d$, i.e., fidelity corresponding to document d, can be formulated as:

$$\mathrm{ESP}(\mathcal{L}_d, \mathcal{L}_d^e) = \sum_{i=1}^{n_{d\to q}} \frac{C(d, q_i, q_i^e, n_{q\to d})}{\log(i+1)}. \tag{7}$$

Finally, a normalization is applied to the values achieved in Eq. 7. The final ESP is estimated using an average over $\mathrm{ESP}(\mathcal{L}_d, \mathcal{L}_d^e)$ of all documents in $\mathcal{D}$.

To provide some intuitions, Figure 1 (c) shows an example of ranked query list of 5 queries retrieved by the EQI system. Out of

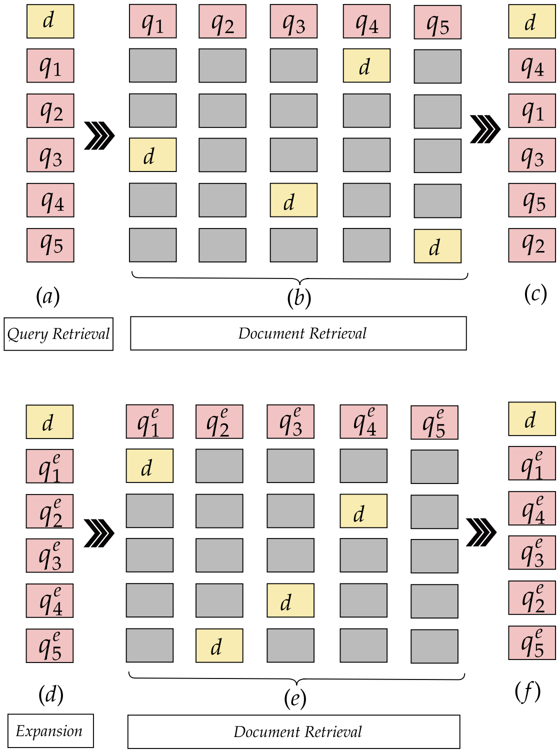

Figure 3: Retrieved query list $\mathcal{L}_d$ (a) and its corresponding re-ranked list $\mathcal{U}_d$ (c), which is achieved using document retrieval system (b); (f) shows the re-ranked query list $\mathcal{U}_d^e$ when performing document retrieval using explained queries. EOP is estimated based on the correlation between the query lists of $\mathcal{U}_d$ (c) and $\mathcal{U}_d^e$ (f).

these five queries, q_4 is non-exposing (NE) and the rest are exposing (E). We argue that the expansion terms added to each query should not change the exposing status of that query for the document to correctly mirror the performance of the corresponding document retrieval system which is shown in Figure 1 (b). In ESP (Eq. 4), the exact rank of the document is not relevant, as long as the exposing status of the query does not change. At best,[2] such explanation could result in the ranked list where the NE queries (here, q_4) are being ranked at the bottom of the list.

Metric 2: Exposing order preservation (EOP). The first proposed metric ESP captures the fidelity to the exposing status of each query (whether it is exposing or non-exposing). However, it does not assess the degree to which the relative exposure level for document d with respect to different queries is preserved. In other words, ESP only accounts for point-wise fidelity of queries in $\mathcal{L}_d$ and not for fidelity of the explanations to the pairwise and listwise [15] order of exposing queries in the ranked list $\mathcal{L}_d$. For instance, imagine q_1 ranks a document at rank 1, and q_2 ranks the same document at rank 20. Now, if q_1^e ranks the document at rank 20, and q_2^e ranks the document at rank 1, still both are exposing (with $n_{q\to d} = 20$); however, q_1^e and q_2^e are downgrading and upgrading the level of relevance respectively, which is equivalent to not reflecting the performance of the system on the original queries.

Algorithm 1 EOP

```
for $d$ in $\mathcal{D}$ do
    Retrieve $n_{d\to q}$ queries from $Q \to \mathcal{L}_d$
    for $q$ in $\mathcal{L}_d$ do
        Retrieve $n_{q\to d}$ documents for query $q \to \mathcal{L}_q$
    end for
    Re-rank queries of $\mathcal{L}_d$ based on $\mathcal{L}_q$ of all $q$ in $\mathcal{L}_d \to \mathcal{U}_d$
    for $q$ in $\mathcal{L}_d$ do
        Expand $q \to q^e$
        Retrieve $n_{q\to d}$ documents for query $q^e \to \mathcal{L}_{q^e}$
    end for
    Re-rank queries of $\mathcal{L}_d$ based on $\mathcal{L}_{q^e}$ of all $q$ in $\mathcal{L}_d \to \mathcal{U}_d^e$
    EOP$(d)$ = Kendall $(\mathcal{U}_d, \mathcal{U}_d^e)$
end for
```

To account for such pairwise and list-wise fidelity of explanations, we propose another metric which we refer to as *exposing order preservation* (EOP). To measure the preservation of the relative exposure level of queries for a document, we leverage the similarity between a re-ranked version of $\mathcal{L}_d$ (the original retrieved list of queries), which we refer to as $\mathcal{U}_d$ (obtained based on the actual exposing level of queries in $\mathcal{L}_d$), and a re-ranked version of the explained retrieved list of queries $\mathcal{L}_d^e$, which we refer to as $\mathcal{U}_d^e$ (obtained based on the actual exposing level of queries in $\mathcal{L}_d^e$). We use the Kendall τ correlation coefficient to measure the similarity between the two ranked list of queries $\mathcal{U}_d$ and $\mathcal{U}_d^e$. The Kendall τ correlation coefficient returns a value of −1 to 1, where −1 indicates perfect negative correlation and 1 shows a perfect match between the two ranked lists. Algorithm 1 describes the computation of the EOP metric for the test document set $\mathcal{D}$. Figure 3 illustrates the procedure to obtain the two query lists $\mathcal{U}_d$ and $\mathcal{U}_d^e$, based on which we estimate the EOP. The final EOP is estimated using an average over EOP(d) of all documents in $\mathcal{D}$:

$$\mathrm{EOP}(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{d\in\mathcal{D}} \mathrm{Kendall}(\mathcal{U}_d, \mathcal{U}_d^e). \tag{8}$$

4 EXPERIMENTAL SETTINGS

As this work is the first to study the explainability for EQI, we need to establish the first experimentation for this task. We hope that our proposal will facilitate further development of expansion methods for explainability. To this aim, we use a benchmark and evaluate three retrieval models as follows.

4.1 Benchmark

EQI experimentation requires a large query collection [4, 13]. We use the FEVER dataset [23] from the BEIR benchmark [22]. This dataset consists of 110K training queries and two sets of 6.7K queries as development and test sets, respectively.

4.2 Ranking models

In our experiments, we leverage the following rankers, which are two widely-used dense retrieval models:

[2] Assuming that there are only 5 queries to be ranked by EQI system.

(1) ANCE [27]: a dense retrieval model based on BERT-base [8] which mines hard negative documents using the model itself during training.
(2) TCT-ColBERT: a dense retrieval model trained with knowledge distillation using ColBERT [11] as the teacher model.

We perform EQI using models that are already trained on the MS MARCO Passage Ranking collection. This reflects our application scenario where the document retrieval system is given as-is and not optimized for the task. In addition, we evaluate BM25 [19] as a widely-used traditional lexical ranker. For both the document retrieval system and the EQI system we use these models. The use of retrieval models for EQI corresponds to the model-reversed setting in prior work [13].

4.3 Test document set

To construct the test document set, we first retrieve documents for all queries using the three ranking models listed in Section 4.2. Then we select documents that are ranked at least one time (for one query) in the top-k of the retrieved list of a query. Next, we aggregate the set of these documents and select those at the intersection of all three ranking models. This results in 25,454 documents. By doing so, we have a single test collection across all models.

5 RESULTS

In this section, we address the following research questions:

(RQ1) What is the performance of the EQI models in terms of RELQ (Eq. 1) when performing the query retrieval using expanded queries?
(RQ2) What is the fidelity of our explainability method in terms of exposing status preservation (ESP)?
(RQ3) What is the fidelity of our explainability method in terms of exposing order preservation (EOP)?

5.1 RQ1: Results of exposing query identification

The results of EQI in terms of RELQ [13] on the original and expanded queries are shown in Table 2. The RELQ scores are around 0.5, indicating that retrieved queries are not completely the same as the truly exposing queries. As we can see, EQI on queries expanded with both RM3 and Bo1 expansion terms results in a small improvement in the performance of EQI models in terms of RELQ for ANCE. In contrast, EQI using expanded queries for BM25 and TCT-ColBERT results in comparable performance to when the original queries are used. Furthermore, we see that EQI on expanded queries with either RM3 and Bo1 is comparable. To better analyze this, we measure the similarity between the retrieved query list of the two methods for each ranking model. We use Rank Biased Overlap (RBO) [26]. RBO measures the similarity of two ranked lists [5, 6] by considering the overlap and order of items at different depths. It assigns higher weights to items at the top of the rankings and gradually reduces the weight as it moves down the lists. The RBO value ranges from 0 to 1, where 0 indicates no overlap between the rankings, and 1 represents identical ranked lists [26]. Table 3 shows the RBO of the ranked query lists expanded with the terms from two methods of RM3 and Bo1. The parameter p in this table determines the top-weightedness of the metric. While RM3 and Bo1 show comparable EQI performance in terms of RELQ (Table 2), their retrieved lists do not show high similarities (Table 3) across all values of p. This suggests that the EQI method is not sensitive to differences in expansion terms from the two methods.

Table 2: EQI results of query expansion on the FEVER test set in terms of exposure list quality (RELQ). The document retrieval for each category of models is its corresponding base model, i.e., BM25, ANCE, TCT-ColBERT.

EQI-Model	RELQ
BM25	0.4937
BM25 w/ RM3	0.4967
BM25 w/ Bo1	0.4964
ANCE	0.5097
ANCE w/ RM3	0.5262
ANCE w/ Bo1	0.5245
TCT-ColBERT	0.5357
TCT-ColBERT w/ RM3	0.5314
TCT-ColBERT w/ Bo1	0.5307
Brute force	1.000

Table 3: Rank Biased Overlap (RBO) between the ranked query lists of EQI models using the expansion terms from RM3 and Bo1 methods.

	RBO		
EQI Model	$p = 0.9$	$p = 0.95$	$p = 0.97$
BM25	0.6070	0.6211	0.6245
ANCE	0.5797	0.5865	0.5817
TCT-ColBERT	0.6338	0.6512	0.6550

5.2 RQ2: Results of fidelity based on ESP

Table 4 shows the ESP fidelity scores (Eq. 7) with two settings of evaluation: (i) using both exposing (E) and non-exposing (NE) queries, and (ii) using only exposing queries. As we can see from the results in the (E/NE) column, ANCE shows higher fidelity scores than BM25 and TCT-ColBERT across all ranking cut-offs, and with the expansion terms from both RM3 and Bo1 methods. This could mean that, on average, for all documents in $\mathcal{D}$, ANCE is more robust to retrieval with the terms added to the queries by the expansion methods.

Effect of non-exposing queries. The number of non-exposing queries at each rank cut-off of the EQI output (averaged over all documents in $\mathcal{D}$) is also shown in Table 4. As can be seen, at higher ranking cut-off values ($n_{d\rightarrow q}$), there exist more proportion of non-exposing queries. In addition, intuitively, non-exposing queries are more robust in terms of changing the exposure status of a given document as they are less relevant to the query.[3] This could justify the increase in the fidelity scores based on exposing status preservation at higher cut-offs values in Table 4 under the (E/NE) column. To further explore the effect of non-exposing queries (NE), the ESP fidelity scores based on only exposing queries (E) are also reported

[3] Only a small proportion of queries would expose a document at their top ranked lists of documents.

Table 4: ESP w/ (E/NE) shows two-sided fidelity scores using the change in the Exposure Status over all exposing (E) and non-exposing (NE) queries. ESP w/ (E) shows two-sided fidelity scores using the change in the Exposure Status over only exposing (E) queries. Avg #NE represents the average number of non-exposing queries at a rank cut-off of EQI system.

	ESP w/ (E/NE)			ESP w/ (E)			Avg #NE		
Models	@3	@5	@10	@3	@5	@10	@3	@5	@10
BM25 w/ RM3	0.7965	0.8122	0.8383	0.6693	0.6881	0.7103	1.089	2.122	5.215
BM25 w/ Bo1	0.7885	0.8047	0.8320	0.6604	0.6788	0.7013	1.098	2.127	5.197
ANCE with RM3	0.8384	0.8490	0.8691	0.6392	0.6640	0.6922	1.250	2.383	5.756
ANCE w/ Bo1	0.8444	0.8542	0.8731	0.6459	0.6697	0.6982	1.206	2.308	5.615
TCT-ColBERT w/ RM3	0.7824	0.7957	0.8210	0.6685	0.6840	0.7014	1.093	2.127	5.267
TCT-ColBERT w/ Bo1	0.7776	0.7922	0.8189	0.6622	0.6701	0.6913	1.078	2.096	5.202

Table 5: Two-sided fidelity scores using exposure order preservation (EOP) metric.

	EOP		
Models	@3	@5	@10
BM25 w/ RM3	0.4910	0.4072	0.3497
BM25 w/ Bo1	0.5037	0.4154	0.3549
ANCE w/ RM3	0.5106	0.4422	0.4110
ANCE w/ Bo1	0.5294	0.4633	0.4278
TCT-ColBERT w/ RM3	0.4348	0.3621	0.3140
TCT-ColBERT w/ Bo1	0.4489	0.3739	0.3261

in Table 4. As can be seen, taking only exposing queries into account results in lower fidelity scores than when both exposing and non-exposing queries are counted towards the computation of ESP fidelity. This indicates the role of non-exposing queries in higher fidelity scores.

Moreover, in contrast to the results of ESP with (E/NE), ANCE shows lower fidelity scores than BM25 and TCT-ColBERT in terms of ESP with only (E). This could mean that the higher values of ESP w/ (E/NE) for ANCE in comparison to BM25 and TCT-ColBERT are rooted in the exposing status preservation of ANCE on non-exposing queries, i.e., ANCE preserves non-exposing queries as non-exposing ones more effectively than the two other models.

5.3 RQ3: Results of fidelity based on EOP

Table 5 shows the results of the three ranking models with expansion terms from RM3 and Bo1 in terms of EOP. As it can be seen, similar to ESP, in terms of EOP, ANCE also shows higher performance than BM25 and TCT-ColBERT when performing EQI for their corresponding document retrieval system. This further indicates the robustness of ANCE to retrieval with the terms added to the queries by the expansion methods, as explored in RQ2. Figure 4 visualizes the ESP and EOP results at different EQI ranking cut-off values for all ranking models using expansion terms from RM3.[4] As it can be seen, in contrast to ESP where the performance of the models increases at higher EQI ranking cut-off values, in terms of EOP, the performance of models decreases in deeper ranks of EQI. This indicates the difficulty of preserving the relative exposing level of queries in comparison to preserving their exposing status as we increase the number of queries in $\mathcal{L}_d$.

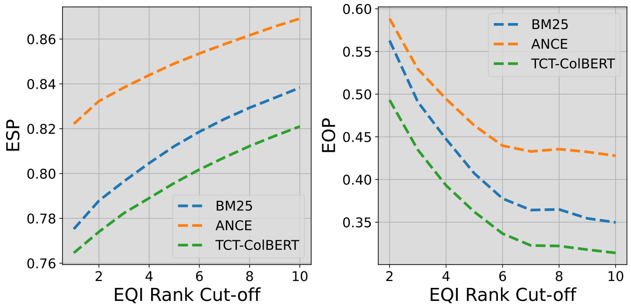

Figure 4: ESP and EOP at different EQI ranking cut-off values.

Furthermore, as we can see in Figure 4, ESP shows less variability to the EQI ranking cut-offs in contrast to EOP which decreases with a higher steepness in lower ranks. Besides, the performance gap between ANCE and the two other models BM25 and TCT-ColBERT in terms of EOP is further magnified at deeper ranking cut-offs.

6 CONCLUSIONS

EQI systems provide transparency of search engines by identifying exposing queries for a given document, especially in the context of content creators. Providing explanation for these exposing queries is an additional goal introduced in this work to offer more insights and transparency on how a document retrieval system maps queries to documents. We specifically explored query expansion-based methods to explain the list of queries retrieved by an EQI system. Moreover, as there is no grounded work on evaluating the quality of explanation in a EQI system, we proposed two metrics to evaluate the fidelity of the explanations.

Our results with three retrieval models show that ANCE has higher fidelity than BM25 and TCT-ColBERT in terms of exposing status preservation as well as exposing order preservation. Moreover, the discrepancy between the results based on ESP and EOP shows the importance of utilizing proper viewpoint as to measuring the quality of explanations.

Our work is the starting point for follow-up studies on transparency of search engines beyond document ranking, helping users in their role as content creators to understand how search engines might expose their content. In future work, we plan to conduct human evaluations of the explanations for EQI methods to qualitatively assess their comprehensibility, i.e., how well do users understand these explanations?

[4] Our findings were the same for Bo1.

7 ACKNOWLEDGMENTS

This work is funded by the DoSSIER project under European Union's Horizon 2020 research and innovation program, Marie Skłodowska-Curie grant agreement No. 860721.

REFERENCES

[1] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 357–389.

[2] Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable Information Retrieval: A Survey. *arXiv preprint arXiv:2211.02405* (2022).

[3] Leif Azzopardi, Rosanne English, Colin Wilkie, and David Maxwell. 2014. Page Retrievability Calculator. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416* (Amsterdam, The Netherlands) *(ECIR 2014)*. 737–741.

[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] Charles L.A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. 2020. Offline Evaluation by Maximum Similarity to an Ideal Ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 225–234.

[6] Charles L.A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2020. Offline Evaluation Without Gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 185–192.

[7] Jeffrey Dalton, Shahrzad Naseri, Laura Dietz, and James Allan. 2019. Local and Global Query Expansion for Hierarchical Complex Topics. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*. Springer, 290–303.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.

[10] Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards Explainable Conversational Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2786–2795. https://doi.org/10.1145/3539618.3591884

[11] Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 39–48.

[12] Victor Lavrenko and W Bruce Croft. 2017. Relevance-based Language Models. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 260–267.

[13] 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*. 3662–3672.

[14] Michael Llordes, Debasis Ganguly, Sumit Bhatia, and Chirag Agarwal. 2023. Explain Like I am BM25: Interpreting a Dense Model's Ranked-List with a Sparse Approximation. (2023).

[15] Lijun Lyu and Avishek Anand. 2023. Listwise Explanations for Ranking Models Using Multiple Explainers. In *European Conference on Information Retrieval*. Springer, 653–668.

[16] Shahrzad Naseri, Jeffrey Dalton, Andrew Yates, and James Allan. 2021. CEQE: Contextualized Embeddings for Query Expansion. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, Cham, 467–482.

[17] Shahrzad Naseri, Jeffrey Dalton, Andrew Yates, and James Allan. 2022. CEQE to SQET: A Study of Contextualized Embeddings for Query Expansion. *Information Retrieval Journal* 25, 2 (2022), 184–208.

[18] Razieh Rahimi, Youngwoo Kim, Hamed Zamani, and James Allan. 2021. Explaining Documents' Relevance to Search Queries. *arXiv preprint arXiv:2111.01314* (2021).

[19] Stephen E. Robertson and Steve Walker. 1994. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *SIGIR'94*. Springer, 232–241.

[20] Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. 2013. Learning to Rank Query Suggestions for Adhoc and Diversity Search. *Information Retrieval* 16 (2013), 429–451.

[21] Jaspreet Singh and Avishek Anand. 2020. Model Agnostic Interpretability of Rankers Via Intent Modelling. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 618–628.

[22] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[23] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 809–819.

[24] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 13–22.

[25] Ellen M. Voorhees. 1994. Query Expansion Using Lexical-semantic Relations. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer, 61–69.

[26] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity Measure for Indefinite Rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (2010), 1–38.

[27] 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).

[28] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Query Understanding Via Intent Description Generation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1823–1832.