ExDocS: Evidence based Explainable Document Search

Sayantan Polley*[1], Atin Janki*[1], Marcus Thiel[1], Juliane Hoebel-Mueller[1] and Andreas Nuernberger[1]

[1]Otto von Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany – first authors with * have equal contribution

Abstract

We present an explainable document search system (ExDocS), based on a re-ranking approach, that uses textual and visual explanations to explain document rankings to non-expert users. ExDocS attempts to answer questions such as "Why is document X ranked at Y for a given query?", "How do we compare multiple documents to understand their relative rankings?". The contribution of this work is on re-ranking methods based on various interpretable facets of evidence such as term statistics, contextual words, and citation-based popularity. Contribution from the user interface perspective consists of providing intuitive accessible explanations such as: "document X is at rank Y because of matches found like Z" along with visual elements designed to compare the evidence and thereby explain the rankings. The quality of our re-ranking approach is evaluated on benchmark data sets in an ad-hoc retrieval setting. Due to the absence of ground truth of explanations, we evaluate the aspects of interpretability and completeness of explanations in a user study. ExDocS is compared with a recent baseline - explainable search system (EXS), that uses a popular posthoc explanation method called LIME. In line with the "no free lunch" theorem, we find statistically significant results showing that ExDocS provides an explanation for rankings that are understandable and complete but the explanation comes at the cost of a drop in ranking quality.

Keywords

Explainable Rankings, XIR, XAI, Re-ranking

1. Introduction

Explainability in Artificial intelligence (XAI) is currently a vibrant research topic that attempts to make AI systems transparent and trustworthy to the concerned stakeholders. The research in XAI domain is interdisciplinary but is primarily led by the development of methods from the machine learning (ML) community. From the classification perspective, e.g., in a diagnostic setting a doctor may be interested to know that how prediction for a disease is made by the AI-driven solution. XAI methods in ML are typically based on exploiting features associated with a class label, development of add-on model specific methods like LRP [2], model agnostic ways such as LIME [3] or causality driven methods [4]. The explainability problem in IR is inherently different from a classification setting. In IR, the user may be interested to know how a certain document is ranked for the given query or why a certain document is ranked higher than others [5]. Often an explanation is an answer to a why question [6].

In this work, Explainable Document Search (ExDocS), we focus on a non-web ad-hoc text retrieval setting and aim to answer the following research questions:

1. Why is a document X ranked at Y for a given query?
2. How do we compare multiple documents to understand their relative rankings?
3. Are the explanations provided interpretable and complete?

There have been works [5], [7] in the recent past that attempted to address related questions such as "Why is a document relevant to the query?" by adapting XAI methods such as LIME [3] primarily for neural rankers. We argue that the idea of relevance has deeper connotations related to the semantic and syntactic notion of similarity in text. Hence, we try to tackle the XAI problem from a ranking perspective. Based on interpretable facets we provide a simple re-ranking method that is agnostic of the retrieval model. ExDocS provides local textual explanations for each document (Part D in Fig. 1). The re-ranking approach enables us to display the "math behind the rank" for each of the retrieved documents (Part E in Fig. 1). Besides, we also provide a global explanation in form of a comparative view of multiple retrieved documents (Fig. 4).

We discuss relevant work for explainable rankings in section two. We describe our contribution to the re-ranking approach and methods to generate explanation in section three. Next in section four, we discuss the quantitative evaluation of rankings on benchmark data sets and a comparative qualitative evaluation with an explainable search baseline in a user study. To our knowledge, this is one of the first works comparing two explainable search systems in a user study. In section five, we conclude that ExDocS provides explanations that are interpretable and complete. The results are statistically significant in Wilcoxon signed-rank test. However, the explanations

The 1st International Workshop on Causality in Search and Recommendation (CSR'21), July 15, 2021, Online
sayantan.polley@ovgu.de (S. Polley*); atin.janki@ovgu.de (A. Janki*); marcus.thiel@ovgu.de (M. Thiel); juliane.hoebel@ovgu.de (J. Hoebel-Mueller); andreas.nuernberger@ovgu.de (A. Nuernberger)

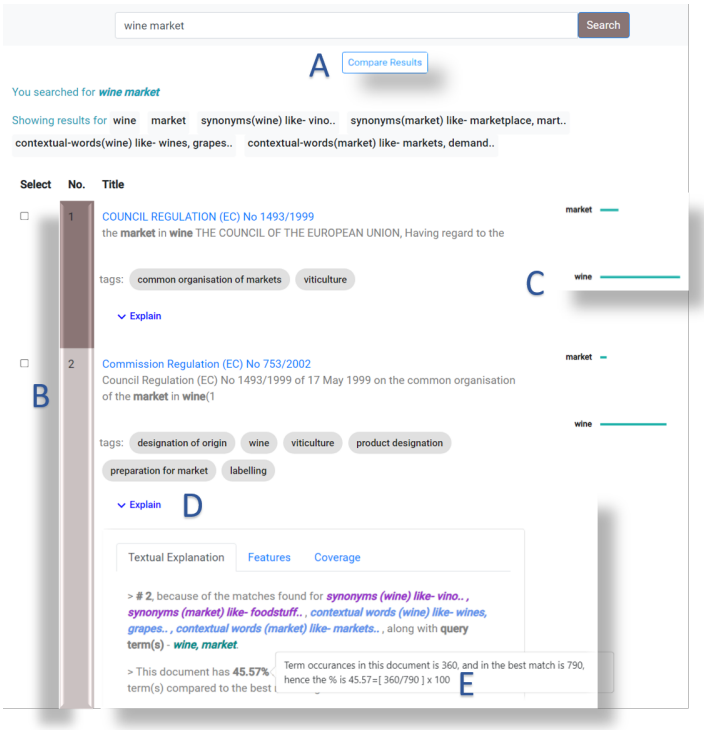

Figure 1: The ExDocS Search Interface. Local Textual explanation, marked (D), explains the rank of a document with a simplified mathematical score (E) used for re-ranking. A query-term bar, marked (C), for each document signifies the contribution of each query term. Other facets of Local explanation can be seen in Fig 2 & 3. A running column in the left marked (B) shows a gradual fading of color shade with decreasing rank. Global explanation via document comparison marked here as (A), is shown in Fig 4. Showing search results for a sample query - 'wine market' on EUR-Lex [1] dataset.

come at a cost of reduced ranking performance paving way for future work. The ExDocS system is online[1] and the source code is available on-request for reproducible research.

2. Related Work

The earliest attempts on making search results explainable can be seen through the visualization paradigms [8, 9, 10] that aimed at explaining term distribution and statistics. Mi and Jiang [11] noted that IR systems were one of the earliest among other research fields to offer interpretations of system decisions and outputs, through search result summaries. The areas of product search [12] and personalized professional search [13], have explored explanations for search results by creating knowledge-graphs based on user's logs. In [14] Melucci made a preliminary study and suggested that structural equation models from the causal perspective can be used to generate explanations for search systems. Related to explainability, the perspective of ethics and fairness [15, 16] is also often encountered in IR whereby the retrieved data may be related to disadvantaged people or groups. In [17] a categorization of fairness in rankings is devised based on the use of pre-processing, in-processing, or

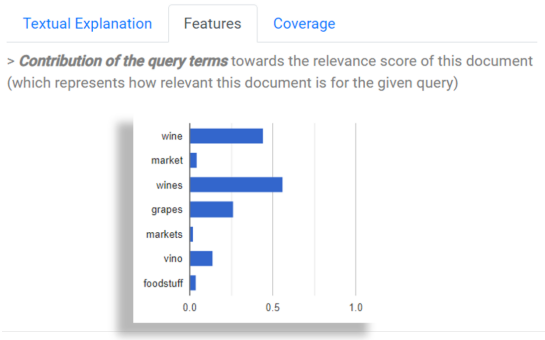

Figure 2: Contribution of Query Terms for relevance

[1] https://tinyurl.com/ExDocSearch

Figure 3: Coverage of matched terms in a document

post-processing strategies.

Recently there has been a rise in the study of interpretability of neural rankers [5, 7, 18]. While [5] uses LIME, [7] uses DeepSHAP for generating explanations and both of them differ considerably. Neural ranking can be thought of as an ordinal classification problem, thereby making it easier to leverage the XAI concepts from the ML community to generate explanations. Moreover, [18] generates explanations through visualization using term statistics and highlighting important passages within the documents retrieved. Apart from this, [19] offers a tool built upon Lucene to explain the internal workings of the Vector Space Model, BM25, and Language Model, but it is aimed at assisting researchers and is still far from an end user's understanding. ExDocS also focuses on explaining the internal operations of the search system similar to [19], however, it uses a custom ranking approach.

Singh and Anand's EXS [5] comes closest to ExDocS in terms of the questions they aim to answer through explanations, such as - "Why is a document relevant to the query?" and "Why is a document ranked higher than the other?". EXS uses DRMM (Deep Relevance Matching Model), a pointwise neural-ranking model that uses a deep architecture at the query term level for relevance matching. For generating explanations it employs LIME [3]. We consider the explanations from EXS as a fair baseline and compare with ExDocS in a user-study.

3. Concept: Re-ranking via Interpretable facets

The concept behind ExDocS is based on the re-ranking of interpretable facets of evidence such as term statistics, contextual words, and citation-based popularity. Each of these facets is also a selectable search criterion in the search interface. We have a motivation to provide a simple intuitive mathematical explanation of each rank with reproducible results. Hence, we start with a common TF-IDF based vector space model (VSM as OOTB Apache Solr) with cosine similarity (ClassicSimilarity). VSM helped us to separate the contributions of query terms enabling us to analytically explain the ranks. BM25 was not deemed suitable for explaining the rankings to a user, since it could not be interpreted completely analytically. On receiving a user query, we expand the query and search the index. The top hundred results are passed to the re-ranker (refer to Algo. 1) to get the final results. Term-count is taken as the first facet of evidence since we assumed that it is relatively easy to analytically explain to a non-expert end-user as: "document X has P % relative occurrences.. compared to the best matching document" (refer to Part E in 1). The assumption on term-count is also in line with a recent work [18] on explainable rankings.

Skip-gram word-embeddings are used to determine contextual words. About two to three nearest neighbor words are used to expand the query. Additionally, the WordNet thesaurus is used to detect synonyms. The optimal combination of the ratio of word-embeddings versus synonyms is empirically determined by ranking performance. Re-ranking is performed based on the proportion of co-occurring words. This enables us to provide local explanations such as "document X is ranked at position Y because of matches found for synonyms like A and contextual words like B". Citation analysis is performed by making multiple combinations of weighted in-links, Page Rank, and HITS score for each document. Citation analysis was selected and deemed as an interpretable facet that we named "document popularity". We argue that this could be used to generate understandable explanations such as: "document X is ranked at Y because of the presence of popularity". Finally, we re-rank using the following facets as shown below:

- `Keyword Search`: 'term statistics' (term-count)
- `Contextual Search`: 'context-words' (term-count of query words + expanded contextual words by word-embeddings).
- `Synonym Search`: 'contextual words' (term-count of query words + expanded contextual words). Contextual words are synonyms, in this case, using Word-Net.
- `Contextual and Synonym Search`: 'contextual words' (term-count of query words + expanded contextual words). Contextual words are word-embeddings+synonyms in this case.
- `Keyword Search with Popularity score`: 'citation-based popularity' (popularity score of a document)

Based on benchmark ranking performance, we empirically determine a weighted combination of these facets

which is also available as a search criteria choice in the interface. Additionally, we provide local and global visual explanations. Local ones in form of visualizing the contribution of features (expanded query terms) for each document as well as comparing them globally for multiple documents (refer the Evidence Graph in the lower part of Fig. 4).

input : q = {w1,w2,...,wn}, D = {d1,d2,...,dm}, facet
output : A re-ranked doc list

Select top-k docs from D using cosine similarity, such as

$$\{d'1, d'2, ..., d'k\} \in Dk$$

for $i \leftarrow 1$ **to** k **do**
  if *facet == 'term statistics' or 'contextual words'* **then**
    evidence(di)$\leftarrow \Sigma_{w \in q} count(w, di)$
    `// count(w, di) is count of term w in di`
  end
  if *facet == 'citation-based popularity'* **then**
    evidence(di)$\leftarrow popularityScore(di)$
    `// popularityScore(di) could be inLinks count, PageRank or HITS score of di`
  end
end
Rerank all docs in Dk using evidence
return Dk

Algorithm 1: Re-ranking algorithm

4. Evaluation

We have two specific focus areas in evaluation. The first one is related to the quality of the rankings and the second one is related to the explainability aspect. We leave out evaluation of the popularity score model for future work.

4.1. Evaluation of re-ranking algorithm

We experimented the re-ranking algorithm on the TREC Disk 4 & 5 (-CR) dataset. The evaluations were carried out by using the trec_eval[20] package. We used TREC-6 adhoc queries (topics 301-350) and used only 'Title' of the topics as the query. We noticed that `Keyword Search`, `Contextual Search`, `Synonym Search`, and `Contextual Synonym Search` systems were unable to beat the 'Baseline ExDocS' (OOTB Apache Solr) on metrics such as MAP, R-Precision, and NDCG (refer to Table 1). We benchmark our retrieval performance by comparing with [21] and confirm that our ranking approach needs improvement to at least match the baseline performance metrics.

4.2. Evaluation of explanations

We performed a user study to qualitatively evaluate the explanations. Also, to compare ExDocS's explanations with that of EXS; we integrated EXS's explanation model into our interface. Therefore, keeping the look and feel of both systems alike, we tried to reduce user's bias towards any system.

4.2.1. User study setup

A total of 32 users participated in a lab controlled user study. 30 users were from a computer science background while 26 users had a fair knowledge of information retrieval systems. Each user was asked to test out both the systems and the questionnaire was formatted in a Latin-block design. The name of the systems was masked as System-A (EXS) and System-B (ExDocS).

4.2.2. Metrics for evaluation

We use the existing definitions ([6] and [22]) of *Interpretability*, *Completeness* and *Transparency* in the community with respect to evaluation in XAI. The following factors are used for evaluating the quality and effectiveness of explanations:

- *Interpretability*: describing the internals of a system in human-understandable terms [6].
- *Completeness*: describing the operation of a system accurately and allowing the system's behavior to be anticipated in future [6].
- *Transparency*: an IR system should be able to demonstrate to its users and other interested parties, why and how the proposed outcomes were achieved [22].

4.3. Results and Discussion

We discuss the results of our experiments and draw conclusions to answer the research questions.

RQ1. Why is a document X ranked at Y for a given query?
We answer this question by providing the individual textual explanation for every document (refer to Part D of Fig. 1) on the ExDocS interface. The "math behind the rank" (refer to Part E of Fig. 1) of a document is explained as a percentage of the evidence with respect to the best matching document.

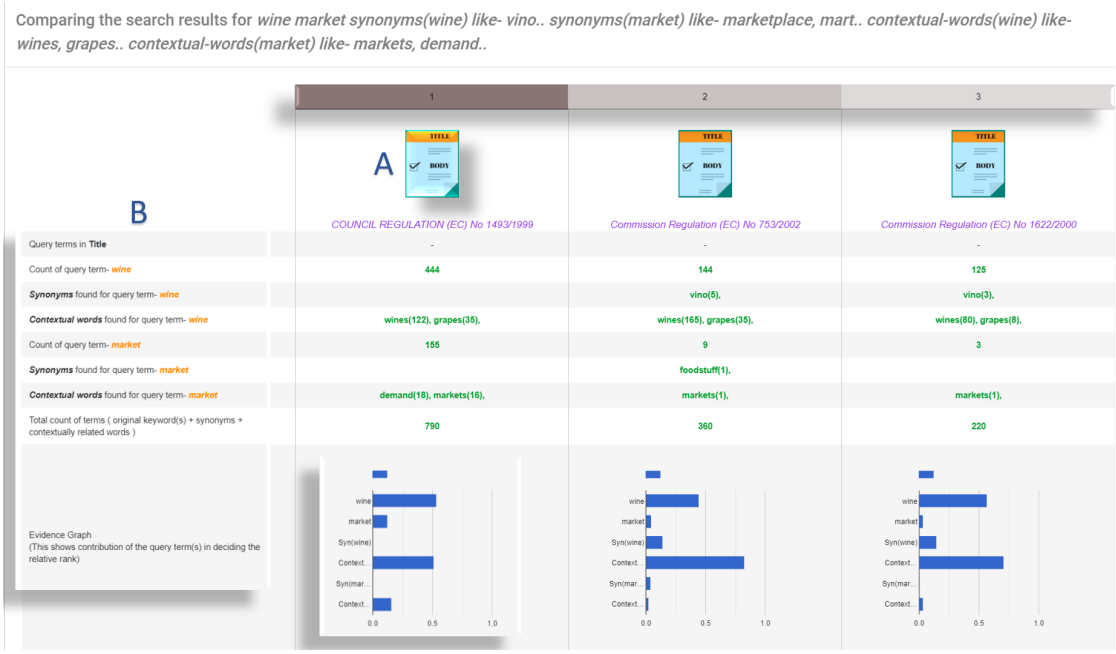

Figure 4: Global Explanation by comparison of evidence for multiple documents (increasing ranks from left to right). A title-body image is provided, marked (A), to indicate whether the query term was found in title and/or body. The column marked (B), represents the attributes for comparison.

Table 1
MAP, R-Precision, and NDCG values for ExDocS search systems against TREC-6 benchmark values*[21]

IR Systems	MAP	R-Precision	NDCG
csiro97a3*	0.126	0.1481	NA
DCU97vs*	0.194	0.2282	NA
mds603*	0.157	0.1877	NA
glair61*	0.177	0.2094	NA
Baseline ExDocS	0.186	0.2106	0.554
Keyword Search	0.107	0.1081	0.462
Contextual Search	0.080	0.0955	0.457
Synonym Search	0.078	0.0791	0.411
Contextual and Synonym Search	0.046	0.0526	0.405

RQ2. How do we compare multiple documents to understand their relative rankings?
We provide an option to compare multiple documents through visual and textual paradigms (refer to Fig. 4). The evidence can be compared and contrasted and thereby understand the reasons for a document's rank being higher or lower than others.

RQ3. Are the generated explanations interpretable and complete?
We evaluate the quality of the explanations in terms of their interpretability and completeness. Empirical evidence from the user study on Interpretability:

1. 96.88% of the users understood the textual explanations of ExDocS
2. 71.88% of the users understood the relation between the query term and features (synonyms or contextual words) shown in the explanation
3. Users gave a mean rating of 4 out of 5 (standard deviation = 1.11) to ExDocS on the understandability of the percentage calculation for rankings, shown as part of the explanations

When users were explicitly asked - whether they could "gather an understanding of how the system functions based on the given explanations", users gave a positive

response with a mean rating of 3.84 out of 5 (standard deviation = 0.72). The above-mentioned empirical evidence indicates that the ranking explanations provided by ExDocS can be deemed as interpretable.

Empirical evidence from the user study on Completeness:

1. All users found the features shown in the explanation of ExDocS to be reasonable (i.e. sensible or fairly good)
2. 90.63% of the users understood through comparative explanations of ExDocS that- why a particular document was ranked higher or lower than other documents

Moreover, 78.13% of total users claimed that they could anticipate ExDocS behavior in the future based on the understanding gathered through explanations (individual and comparative). Based on the above empirical evidence we argue that the ranking explanations generated by ExDocS can be assumed to be complete.

Transparency: We investigate if the explanations make ExDocS more transparent [22] to the user. Users gave ExDocS a mean rating of 3.97 out of 5 (standard deviation = 0.86) on 'Transparency' based on the individual (local) explanations. In addition to that, 90.63% of the total users indicated that ExDocS became more transparent after reading the comparative (global) explanations. This indicates that explanations make ExDocS more transparent to the user.

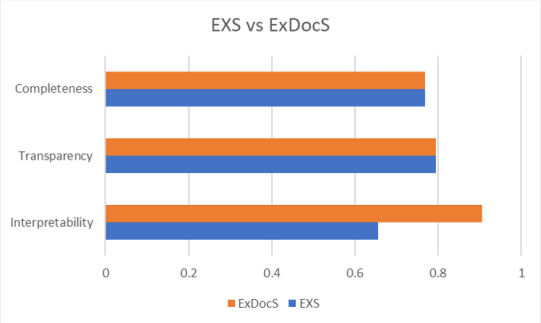

Figure 5: Comparison of explanations from EXS and ExDocS on different XAI metrics. All the values shown here are scaled between [0-1] for simplicity.

Comparison of explanations between ExDocS and EXS:
Both the systems performed similarly in terms of $Transparency$ and $Completeness$. However, users found ExDocS explanations to be more interpretable compared to that of EXS (refer to Fig. 5), and this comparison was statistically significant in WSR test ($|W| < W_{critical(\alpha=0.05, N_r=10)}$ = 10, where $|W| = 5.5$).

5. Conclusion and Future Work

In this work, we present an Explainable Document Search (ExDocS) system that attempts to explain document rankings using a combination of textual and visual elements to a non-expert user. We make use of word embeddings and WordNet thesaurus to expand the user query. We use various interpretable facets such as term statistics, contextual words, and citation-based popularity. Re-ranking results from a simple vector space model with such interpretable facets help us to explain the "math behind the rank" to an end-user. We evaluate the explanations by comparing ExDocS with another explainable search baseline in a user study. We find statistically significant results that ExDocs provides interpretable and complete explanations. Although, it was difficult to find a clear winner between both systems in all aspects. In line with the "no free lunch" theorem, the results show a drop in ranking quality on benchmark data sets at the cost of getting comprehensible explanations. This paves way for ongoing research to include user feedback to adapt the rankings and explanations. ExDocS is currently being evaluated in domain-specific search settings like law search where explainability is a key factor to gain user trust.

References

[1] E. L. Mencia, J. Fürnkranz, Efficient multilabel classification algorithms for large-scale problems in the legal domain, in: Semantic Processing of Legal Texts, Springer, 2010, pp. 192–215.

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[3] M. T. Ribeiro, S. Singh, C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144.

[4] J. Pearl, et al., Causal inference in statistics: An overview, Statistics surveys 3 (2009) 96–146.

[5] J. Singh, A. Anand, EXS: Explainable Search Using Local Model Agnostic Interpretability, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 770–773.

[6] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE

5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2018, pp. 80–89.

[7] Z. T. Fernando, J. Singh, A. Anand, A Study on the Interpretability of Neural Retrieval Models Using DeepSHAP, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1005–1008.

[8] M. A. Hearst, TileBars: Visualization of Term Distribution Information in Full Text Information Access, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '95, ACM Press/Addison-Wesley Publishing Co., USA, 1995, p. 59–66.

[9] O. Hoeber, M. Brooks, D. Schroeder, X. D. Yang, TheHotMap.Com: Enabling Flexible Interaction in Next-Generation Web Search Interfaces, in: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, WI-IAT '08, IEEE Computer Society, USA, 2008, p. 730–734.

[10] M. A. Soliman, I. F. Ilyas, K. C.-C. Chang, URank: Formulation and Efficient Evaluation of Top-k Queries in Uncertain Databases, in: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, SIGMOD '07, Association for Computing Machinery, New York, NY, USA, 2007, p. 1082–1084.

[11] S. Mi, J. Jiang, Understanding the Interpretability of Search Result Summaries, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019, p. 989–992.

[12] Q. Ai, Y. Zhang, K. Bi, W. B. Croft, Explainable Product Search with a Dynamic Relation Embedding Model, ACM Trans. Inf. Syst. 38 (2019).

[13] S. Verberne, Explainable IR for personalizing professional search, in: ProfS/KG4IR/Data: Search@ SIGIR, 2018.

[14] M. Melucci, Can Structural Equation Models Interpret Search Systems?, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, Association for Computing Machinery, New York, NY, USA, 2019. URL: https://ears2019.github.io/Melucci-EARS2019.pdf.

[15] A. J. Biega, K. P. Gummadi, G. Weikum, Equity of attention: Amortizing individual fairness in rankings, in: The 41st International ACM SIGIR conference on Research & Development in Information Retrieval, 2018, pp. 405–414.

[16] S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-Aware Ranking in Search and Recommendation Systems with Application to LinkedIn Talent Search, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2221–2231. URL: https://doi.org/10.1145/3292500.3330691. doi:10.1145/3292500.3330691.

[17] C. Castillo, Fairness and Transparency in Ranking, SIGIR Forum 52 (2019) 64–71.

[18] V. Chios, Helping results assessment by adding explainable elements to the deep relevance matching model, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY, USA, 2020. URL: https://ears2020.github.io/accept_papers/2.pdf.

[19] D. Roy, S. Saha, M. Mitra, B. Sen, D. Ganguly, I-REX: A Lucene Plugin for EXplainable IR, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2949–2952.

[20] C. Buckley, et al., The trec_eval evaluation package, 2004.

[21] D. K. Harman, E. Voorhees, The Sixth Text REtrieval Conference (TREC-6), US Department of Commerce, Technology Administration, National Institute of Standards and Technology (NIST), 1998.

[22] A. Olteanu, J. Garcia-Gathright, M. de Rijke, M. D. Ekstrand, Workshop on Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval (FACTS-IR), in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1423–1425.