The Curious Case of IR Explainability: Explaining Document Scores within and across Ranking Models

Procheta Sen
Dublin City University, Dublin, Ireland
procheta.sen2@mail.dcu.ie

Debasis Ganguly
IBM Research, Dublin, Ireland
debasis.ganguly1@ie.ibm.com

Manisha Verma
Verizon Media, New York, USA
manishav@verizonmedia.com

Gareth J. F. Jones
Dublin City University, Dublin, Ireland
gareth.jones@dcu.ie

ABSTRACT

It can often be useful for an information retrieval (IR) practitioner to analyze the behaviour of the similarity function of an IR model in terms of the three fundamental aspects: a) *frequency of a term in a document*, b) *frequency of a term in a collection* and c) *the length of a document*, in order to optimize the relative importance of each component for a particular document collection and type of queries. The importance of the *model explanations* in terms of the fundamental components is potentially more useful for neural models, where the overall similarity function is not a closed-form functional typical of traditional IR models. We propose a general methodology for approximating an IR model as the coefficients of a linear function of these three fundamental aspects (and an additional aspect of semantic similarity between terms for neural models), which can potentially assist with optimization of the relative importance of each aspect for a specific task. Our analysis shows that these coefficients are useful to compare a model's different parametric instantiations btween alternative models.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**.

ACM Reference Format:
Procheta Sen, Debasis Ganguly, Manisha Verma, and Gareth J. F. Jones. 2020. The Curious Case of IR Explainability: Explaining Document Scores within and across Ranking Models . In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401286

1 INTRODUCTION

Both traditional statistical and neural models of information retrieval (IR) involve aggregation of non-linear functions of individual per-document term matching weights along with global term statistics (e.g. logarithm of term frequency, reciprocal of collection frequency of a term etc.) to compute an overall similarity score, $s(D, Q)$, between each document D, with respect to a query Q. Existing research has investigated whether standard ranking functions (e.g. BM25 or LM) satisfy a set of expected behavioural contracts or *axioms* [3]. An example of such an axiom is that the rate of increase in the document score for a query, $s(Q, D)$, should decrease with an increase in term frequency of a matched query term $q \in Q$. Such axiomatic analysis can be useful in extending this set of axioms, e.g. to account for non-exact or semantic term matches [4], and may eventually help to construct novel similarity functions. In contrast to this axiomatic thread, previous research has also represented a similarity function as a parameterized linear function of features, where each one maps to a concept (a term or a phrase, or a class of terms). The parameter values of such linear functions are tuned by coordinate ascent for a given collection and a set of queries [6].

Different from the axiomatic approach [3, 4], which involves manually analyzing if a model respects a set of axiomatic constraints, we propose an automated method of representing any given similarity function with a set of parameters. These parameters, instead of corresponding to concepts [6], rather correspond to the fundamental building blocks of a standard IR similarity function, namely: i) the *term frequency* of a term in a document, ii) the *document or collection frequency* of a term (an estimate of its global informativeness measure), and iii) the *length of a document*. This enables us to represent a retrieval model as a point in this 3-dimensional function space.

We hypothesize that representing an IR model as a point in a function space, comprised of a set of three fundamental functional components, helps to provide an *explanation* of the working principle of the model. The questions these explanations could potentially address are: a) *differences within a ranking model:* what are the fundamental factors that cause a particular document D to be ranked at position r whereas another document D' at $r'(> r)$?, and b) *differences across ranking models:* what are the relative effects of the fundamental factors that cause a document D to be retrieved at substantially different positions - r_M and r'_M, by two models M and M', respectively. Our analysis shows that the proposed coefficients, which represent the relative importance of the three fundamental aspects, are useful to compare between a model's different parametric instantiations, or even to compare between different models.

2 RELATED WORK

Standard approaches of instance-wise explanations for classification include those of employing linear regression to learn a simplified decision boundary by sampling points around a data instance [8],

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401286

applying a deep convolutional network to estimate instance-wise feature importance [2] etc. Specific to the domain of images, it is a standard practice to conduct importance analysis of different (abstract data representation) layers of a network [7, 12]. Explanations in IR mainly involve estimating the relative importance of matches in term weights [10, 11]. In contrast, the explanations provided by our model are in a different space (specifically the *function space* instead of the *term space*). We argue that such explanations can potentially be useful to explain results across different models, which is difficult to achieve with only term-level importance.

3 APPROXIMATING RANKING MODELS

Before describing a general approach of parametric approximation of a given similarity function $s(D, Q)$, we review the similarity functions of the standard ranking models used in our investigations.

Ranking Model Components. Standard statistical models in IR involve a closed-form similarity function, $s(Q, D)$ of the form

$$s(Q, D) = \sum_{q \in Q \cap D} s(q, D), \quad s(q, D) = \Phi(\phi_x(q, D), \phi_y(D), \phi_z(q)), \quad (1)$$

involving aggregation over a set of per-term similarity scores, $s(q, D)$, each of which is, in turn, a function of:

(1) $\phi_x(q, D) : \mathbb{Z} \mapsto \mathbb{R}$ is a function that transforms an integer raw frequency of a term q in a document D, $x = f(q, D) \in \mathbb{Z}$, to a real number, e.g. $\phi_x(x) = \{\sqrt{x}, \log(x) \ldots\}$.
(2) $\phi_y(D) : \mathbb{Z} \mapsto \mathbb{R}$ is a function that transforms the integer value of the length of a document $y = |D|$, to a real number, e.g. $\phi_y(y) = y^{-1}$ etc.
(3) $\phi_z(q) : \mathbb{Z} \mapsto \mathbb{R}$ is a function that transforms the number of documents across the whole collection in which the term t occurs , $z = c(q)$ (an integer), to a real, e.g. $\phi_z(z) = z^{-1}$ etc.

Different ranking models prescribe alternative ways of defining the functions ϕ_x, ϕ_y and ϕ_z, e.g., the BM25 similarity [9] is defined as,

$$s(q, D) = \frac{N}{log(c(q))} \frac{f(q, D)(k_1 + 1)}{f(q, D) + k_1(1 - b + b\frac{|D|}{|\hat{D}|})}, \quad (2)$$

where $\phi_x(x) = x$, $\phi_z(z) \propto z^{-1}$, and $\phi_y(y) \propto y^{-1}$ (N is the number of documents, and $|\hat{D}|$ is the average document length in the collection). The parameters k_1 and b in BM25 control the relative effects of term frequency and document lengths.

Lack of Transparency in Neural Models. Generally speaking, deep neural models such as DRMM [5] employ pairwise learning to rank, where a triplet loss function of the form

$$\mathcal{L}(q, D^+, D^-; \Theta) = \max(0, 1 - s(q, D^+) + s(q, D^-)), \quad (3)$$

which seeks to learn to predict a higher score for a relevant document D^+ with respect to a non-relevant one (D^-). The features on which DRMM applies a convolution-based feed-forward network are binned histograms of term-match score distributions (weighted by inverse document frequencies). However, is is difficult to see what effects term frequency, ϕ_x, and document length functions, ϕ_y play in the overall predicted score of DRMM.

Consider a question of the form - *why does a model M rank a document D at position r > 1 (say 5) instead of retrieving it at rank* 1. If M is a statistical model (e.g. BM25 or LM), it is possible for an IR practitioner (with a working knowledge of the model) to compute the values of term frequencies and collection statistics of matched terms in the document D, along with the length of D, and to compare these values with the top-ranked document retrieved by another model (say D_{top}). We hypothesize that aligning these compared values can be helpful in answering the question, e.g., if b is set to a high value in BM25, then it could be seen from Equation 2 that BM25 should favour shorter documents. If $|D| < |D_{top}|$, an IR practitioner could see why D is ranked at position 5, as compared to D_{top}. However, it is not possible to conduct an analysis of this manner for the case of neural models, such as DRMM [5].

Functional Representation of an IR Model. We represent the similarity score of an IR model, M, as a 3-dimensional vector of coefficients of the components, $\phi = (\phi_x, \phi_y, \phi_z)$ of a similarity scoring function $s(q, D)$ of Equation 1. The input to our regression model is a 3-dimensional vector of term weights of the form

$$(x, y, z)_{w,D} = (\phi_x(w, D) \stackrel{\text{def}}{=} f(w, D), \phi_y(D) \stackrel{\text{def}}{=} |D|, \phi_z(w) \stackrel{\text{def}}{=} c(w)) \quad (4)$$

The regressed (output) values correspond to the score assigned by an IR model M to a term w in relation to the event of its match in a document D. Additionally, for a neural model we extend this 3-dimensional functional space to 4 dimensions, where the additional dimension, denoted as

$$\phi_\omega(w, t, D) = \mathbf{w} \cdot \mathbf{t}, \ \ w, t \in D, \quad (5)$$

corresponds to the similarity between the embedded vector representations of terms w and t in a document D. The rationale of including this additional feature is to account for the non-exact term matches contributing to the aggregated similarity score [5].

To capture the relative differences in the term weights across different documents that are ranked by M, it is required to compute the regression model over a *set of documents*, which in this case corresponds to the top-K retrieved documents by the model M, $\mathcal{D}_k(M)$. Specifically, the training data comprises input-output value mappings of the form

$$\theta_x x + \theta_y y + \theta_z z + \theta_0 = \hat{s}(w, D), \ \ \forall w \in D, \forall D \in \mathcal{D}_k(M), \quad (6)$$

where $\hat{s}(w, D)$ denotes the predicted score of a term w in document D. The 4-dimensional parameter vector $\theta = (\theta_0, \theta_x, \theta_y, \theta_z) \in \mathbb{R}^4$ (including the additional bias term) is learned by minimizing the L_2-regularized square loss function with stochastic gradient descent.

$$\mathcal{L}(\theta) = (s(w, D) - \hat{s}(w, D; \theta))^2 + |\theta|_2 \quad (7)$$

The number of training data instances and their values thus depend on the parameter K (the number of top ranked documents to consider) and the set of unique terms found in the set $\mathcal{D}_k(M)$.

4 INTERPRETABILITY OF MODELS

To compare the generic characteristics of two retrieval models M_1 and M_2 (note that M_1 and M_2 could be the same function with different parameters, e.g. BM25 with two different instances for (k_1, b)), we first obtain a number of estimated versions of the same model M using each query from the set $\mathcal{Q}$, and then compute the average of these per-query parameter vectors. Formally,

$$\theta(M, \mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} \theta(Q, \mathcal{D}_k(M)), \quad (8)$$

where $\theta(Q, \mathcal{D}_k(M))$ denotes the parameters learned with linear regression (Equation 7) using query Q, and the top documents, $\mathcal{D}_k(M)$, retrieved with model M. Since each ranking model M is approximated using the same set of queries and the same set of features (term-frequency, document length etc.), each belongs to a common vector space equipped with the usual linear operations and inner product between the vectors (in this case, functions). This means that the magnitudes and the directions (signs) of these estimated vectors can be interpreted in the following ways.

(1) **Magnitude:** $|\theta_i(M_1, Q)| > |\theta_i(M_2, Q)|$, indicates that M_1 puts more emphasis on the value of the i^{th} feature component than M_2 in computing the overall similarity score. For example, if the i^{th} component of the parameter vector, $\theta_i(M_1, Q)$, corresponds to the term-frequency feature, i.e. $i = 1$ in Equation 4 indicating ϕ_x, we could argue that a match in term-frequency plays a higher role in computing the overall score in M_1 than M_2.
(2) **Direction:** $\text{sgn}(\theta_i(M_1, Q))$, i.e. the direction of the slope of the i^{th} component of the parameter vector indicates whether the overall similarity function Φ (Equation 1) increases (for positive slope values) or decreases (otherwise).

Explanation within a Ranking Model. After presenting a general interpretation on the models themselves, we now delve into understanding the relative characteristics of documents and term importance measures that a model M considers for obtaining the scores. The question that we seek to address is

Q1: Why does a model M retrieve a document D_1 at rank r_1 and D_2 at r_2 ($r_2 > r_1$ without loss of generality) for a query Q?

This question is more relevant in cases when D_2 is a known relevant document, and an IR practitioner needs to know what characteristics of D_2 could be exploited (possibly by applying a different model that is better able to leverage those characteristics) to improve its rank. To understand document characteristics at a relative level, we propose to use a *reference document*, which in our case is the top-ranked document retrieved by M for query Q, $\mathcal{D}_{top}(M; Q)$. To see the relative differences in the values of the functional components for a document D, we first define an aggregate function (averaging operation) over the values of these features (Equation 4), e.g. for the term-frequency feature this is defined as

$$\bar{\phi_x}(D) = \frac{1}{|Q \cap D|} \sum_{w \in |Q \cap D|} \phi_x(w, D), \tag{9}$$

and so on for feature components y and z. The average relative differences between the feature values between a document D_1 retrieved at a position r_1 (> 1) and the top ranked document D_{top} are then computed using Equation 9, e.g., the relative difference between D and D_{top} for the term-frequency component is given by

$$\Delta_{x,D,D_{top}} = \frac{\bar{\phi_x}(D_{top}) - \bar{\phi_x}(D)}{\bar{\phi_x}(D_{top})}, \tag{10}$$

and so on for components y, z etc. We then define an intrinsic fidelity measure as the dot-product (similarity) between the parameter vector θ and the average vector of the feature values computed as per Equation 10, i.e.,

$$\xi(D, D_{top}) = (\Delta_{x,D,D_{top}}, \Delta_{y,D,D_{top}}, \Delta_{z,D,D_{top}}) \cdot \theta. \tag{11}$$

To see why such agreements between the feature values and the estimated coefficients may indicate an intrinsic fidelity measure, consider the example when $\Delta_{x,D,D_{top}} > 0$, which indicates that the aggregate term frequency averaged over the matched query terms of D_{top} is higher than that of D. An estimated value of the coefficient $\theta_x > 0$, in turn, indicates that a document score should increase with increasing values of the matched term frequencies. Since D is retrieved at a higher position than D_{top}, this is an example of a *consistent* observation. On the other hand a value of $\theta_x < 0$ predicts a decrease of document scores with increasing values of the matched term frequencies, and is an example of an inconsistent observation. The same argument applies for the other two consistent and inconsistent cases, i.e. when $\Delta_{x,D,D_{top}} < 0$, and $\theta_x < 0$ and > 0 respectively. Similar arguments also apply for the fidelity measures corresponding to the other features, i.e., document length, term informativeness, etc.

To explicitly answer **Q1**, we compute the cases where the fidelity scores, $\xi_\gamma(D_2, D_1) > 0, \gamma = \{x, y, z\}$, and the matched values of the corresponding feature components (e.g. x representing term frequency) act as the plausible *explanation* for **Q1**.

Explanation across Ranking Models. In contrast to the earlier case, where we compared between document pairs retrieved at different positions by a single model, we now ask the question

Q2: Why does a model M_1 retrieve a document D at position r_1, whereas model M_2 retrieves D at r_2 for a query Q?

We require the notion of two reference documents, the top-documents retrieved by M_1 and M_2, to see if the functional representations are satisfactory. Specifically, we consider the case where the relative decrease in the score of D in M_1 (with respect to its top document) is higher than that of D in M_2, i.e.,

$$\begin{aligned} &\Delta_s(D, M_1, M_2) = \delta_s(Q, D, M_2) - \delta_s(Q, D, M_1), \text{ where} \\ &\delta_s(Q, D, M) = \frac{s(Q, D_{top}, M) - s(Q, D, M)}{s(Q, D_{top}, M)}. \end{aligned} \tag{12}$$

Considering M_1 as the reference (without loss of generality), we now define a fidelity score across different models M_1 and M_2 by

$$\begin{aligned} &\xi(M_1, M_2) = \Delta_s(D, M_1, M_2) \cdot \Delta(M_1, M_2), \text{ where} \\ &\Delta(M_1, M_2) = \theta(M_1, Q) - \theta(M_2, Q) \end{aligned} \tag{13}$$

Similar to the argument for document pairs within a model, it can be seen that the fidelity score corresponding to a component, e.g. $\xi_x > 0$ means an agreement between the relative decreases between the score values and the term-frequency coefficients between two models. If both of these are positive, it can be argued that a higher importance of term frequency in M_1 (since $\theta_x(M_1) > \theta_x(M_2)$) contributes to a higher decrease in the relative score of D in M_2. Similar arguments also apply for other feature components.

To explicitly answer **Q2**, we compute the cases where the fidelity scores, $\xi_\gamma(M_1, M_2) > 0, \gamma = \{x, y, z\}$, and the matched values of the corresponding feature components (e.g. x for term frequency) act as the plausible *explanation* for **Q2**.

5 EVALUATION

We conducted experiments on the MS-Marco dataset [1], a collection of over 8.8M passages (avg. length of 56.2 words). To quantitatively measure the fidelity of our explanation model, we sampled a

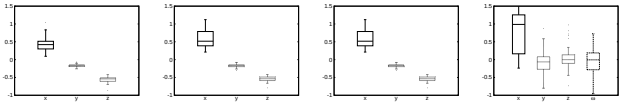

Figure 1: Box-plot of parameter vectors θ for BM25, LM-JM, LM-Dir and DRMM (in order from left-right).

Table 1: Explaining differences within a ranking model for a sample query in the MS-Marco dataset. Positive fidelity values ($\xi_\gamma > 0$) are bold-faced. ↑ (↓)'s indicate whether $\text{sgn}(\theta_\gamma) < 0$ (> 0), $\gamma \in \{x, \ldots, w\}$.

Model	Ranks	ξ_x	ξ_y	ξ_z	ξ_ω
BM25	10	-1.7000	**0.2148**↑	**0.1782**↑	
	20	0.8050	**0.3294**↑	**0.0913**↑	
LM-JM	10	-0.2016	-0.1489	**0.0324**↑	
	20	-0.1347	-0.2177	**0.0130**↑	
LM-Dir	10	-1.7479	**0.1126**↑	-0.0133	
	20	-0.9438	**0.0250**↑	-0.0133	
DRMM	10	**0.4490**↓	-0.0512	0.0000	**0.2235**↓
	20	**0.4490**↓	**0.0390**	0.0000	**0.0311**↓

Table 2: Explaining across different Ranking Models (fidelity scores computed by Equation 13 are bold-faced).

M_1	M_2	$\delta_s(D, M_1)$	$\delta_s(D, M_2)$	ξ_x	ξ_y	ξ_z
LM-JM	BM25	0.2353	0.4099	**1.7564**	-0.0840	**0.4215**
LM-Dir	LM-JM	0.1979	0.4095	-0.2656	**0.1024**	-0.1191
LM-Dir	BM25	0.1959	0.3993	**0.5364**	**0.0184**	**0.3711**

subset of 50 queries from the training set. We removed queries for each statistical model, we experimented with the optimal parameter settings (grid search) on the set of 50 queries. Figure 1 shows the distribution of the coefficient values for the individual components, namely x (term frequency) etc. for each query. To estimate the coefficient values, we set $k = 100$ in Equation 6 for each query. Figure 1 shows that $\bar{\theta}_x$, i.e., the average coefficient value corresponding to the *term frequency* (tf) feature, are positive indicating that each model satisfies constraints 'C1' of [3] (increase in tf leads to increasing document score). Moreover, a sub-linear slope (since these values are mostly less than 1) indicates that the models also satisfy 'C2' of [3] (rate of change of increase in document score decreases with increase in tf). Similarly, it can be seen from the negative values of $\bar{\theta}_z$ that an increase in document frequency (decrease in informativeness) leads to a decrease in the document score.

Comparisons within a model. Table 1 analyzes the relative differences in documents retrieved at positions 10 and 20 in comparison to the top-retrieved document. We seek to identify arguments tht could be provided to *explain* this observation. For the document ranked at position 10, we see that the values with up-arrows (i.e., $\xi_y > 0$ and $\xi_z > 0$) indicate that the increase of the y (document length) and the z (document frequency) values in D_{10} (with respect to D_{top}) contribute to decreasing its score. The bold-faced numbers associated with an arrow (up/down) constitute a set of valid explanations. Some more interesting observations are that in LM-Dir, increase in document frequency (decrease of informativeness) is the main contributing factor for retrieving D_{10} and D_{20} at lower ranks as compared to D_{top}, whereas the deciding factors in DRMM were tf, document length and semantic similarity.

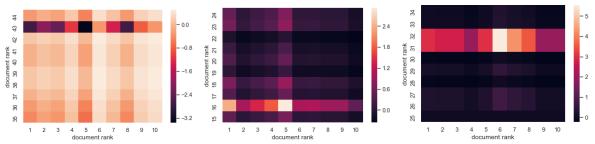

Figure 2: ξ_x, ξ_y and ξ_z distributions (left-right) for different rank pairs in BM25.

Comparisons across models. Table 2 shows that the decrease in relative score for a document using BM25 as compared to LM-JM can be attributed to the tf (x) and document frequency (z) factors.

In Figure 2, we show the fidelity score analysis for x, y and z values in BM25 through heat-map. Darker shades correspond to negative values for fidelity score. x and y axes correspond to two different range of ranks in BM25. For each feature value we show only those ranges of rank pairs where that particular feature was the reason for ranking difference. For example, in the leftmost figure, for $i = 0$ to 9 and for $j = 35$ to 45, term frequency (x) was the major reason for ranking difference. Most of the cells in the left most figure are of lighter shade because of positive values of the fidelity score for x.

ACKNOWLEDGEMENT

This work was supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

REFERENCES

[1] Payal et. al. Bajaj. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv:1611.09268* (2016).
[2] Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *Proc. of ICML'18*, Vol. 80. 882–891.
[3] Hui Fang and ChengXiang Zhai. 2005. An Exploration of Axiomatic Approaches to Information Retrieval. In *Proc. of SIGIR'05*. 480–487.
[4] Hui Fang and ChengXiang Zhai. 2006. Semantic term matching in axiomatic approaches to information retrieval. In *SIGIR'06*. ACM, 115–122.
[5] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *Proc. of CIKM '16*. 55–64.
[6] Donald Metzler and W. Bruce Croft. 2007. Linear Feature-based Models for Information Retrieval. *Inf. Retr.* 10, 3 (June 2007), 257–274.
[7] Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. 2019. Ablation Studies in Artificial Neural Networks. *CoRR arXiv* (2019).
[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proc. of KDD'16*. 1135–1144.
[9] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. (2009), 333–389.
[10] Jaspreet Singh and Avishek Anand. 2018. Interpreting search result rankings through intent modeling. *arXiv preprint arXiv:1809.05190* (2018).
[11] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally Interpretable Ranking Model Explanation. In *Proc. of SIGIR 2019*. 1281–1284.
[12] Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. *CoRR* abs/1311.2901 (2013). http://arxiv.org/abs/1311.2901