RANKSHAP

SHAPLEY VALUE BASED FEATURE ATTRIBUTIONS FOR LEARNING TO RANK

Tanya Chowdhury, Yair Zick, James Allan
University of Massachusetts Amherst

Feature Attribution Methods for Ranking

1. Popular methods designed for **classification/regression** ; limited focus on **ranking**.
2. Understanding ordering in search engines/rec systems critical for :
   - Building user trust and ensuring fairness.
   - Verifying model functionality.

Ranking Feature Attribution Problem

- Let f_R be a black-box ranking model. For query $\vec{q}$ and a set of documents $D = \{\vec{d}_1, \ldots, \vec{d}_k\}$, the model produces an ordered list $f_R(\vec{q}, D)$.
- Given instance $(\vec{q}, D)$, objective is to compute post-hoc feature attributions $\phi_R(f_R, \vec{x})$, where $\vec{x} = (\vec{q}, D)$.

Challenges with Existing Methods

1. **Inconsistencies in Empirical Systems :**
   - Extensions of classification/regression methods, like EXS/Rank-LIME, produce contradictory attributions.
   - DeepSHAP shows low correlation between attributions for the same query, using different reference values.
   - Methods fail to be consistent with basic sanity checks.

Proposed Solution

Introduce RankSHAP, an axiomatic Shapley value-based framework for generalizable, consistent, and human-aligned feature attributions.

RankSHAP Framework : Axioms

Desirable Axioms for Ranking Effectiveness Metrics :

- Relevance Sensitivity, Position Sensitivity

Generalized Ranking Effectiveness Metric (*GREM*) (Theorem 1)

An effectiveness metric satisfies *Relevance* and *Position Sensitivity* if and only if it can be represented as

$$GREM_n = \sum_{j=1}^{n} g(\text{rel}_j) \cdot h(j)$$

where $g(\text{rel}_j)$ is a non-decreasing function and $h(j)$ is a non-increasing function.

Desirable Shapley Properties for Ranking Attributions :

- Rank-Efficiency, Rank-Missingness, Rank-Symmetry, Rank-Monotonicity

RankSHAP Framework : Shapley Value

The Ranking Shapley Value (Theorem 2)

Let V_R be a ranking effectiveness metric that belongs to $GREM_n$.
The Shapley value ϕ_R, computed with respect to V_R, is the **unique** feature attribution that satisfies the Shapley ranking axioms.

$$\phi_R(f_R, \vec{x}, i) = \sum_{S \subseteq M \setminus \{i\}} \frac{|S|!\,(m - |S| - 1)!}{m!} \left[V_R(f_R(S \cup \{i\}, \vec{x})) - V_R(f_R(S, \vec{x}))\right],$$

KernelSHAP approximation

$$L(f_R, g, \pi_{\vec{x}}) = \sum_{\vec{z} \in Z} [NDCG(f_R(\vec{z})) - NDCG(g(\vec{z}))]^2 \pi_{\vec{x}}(\vec{z})$$

RankSHAP Framework : Axiomatic Analysis

$$L_{\text{RANKLIME}}(f_R, g, \pi_{\vec{x}}) = \sum_{\vec{z} \in Z} ApproxNDCG(f_R(\vec{z}), g(\vec{z}))\ \pi_{\vec{x}}(\vec{z})$$

$$L_{\text{RANKINGSHAP}}(f_R, g, \pi_{\vec{x}}) = \sum_{\vec{z} \in Z} [\tau(f_R(\vec{z}), g(\vec{z}))]^2\ \pi_{\vec{x}}(\vec{z})$$

$$L_{\text{RANKSHAP}}(f_R, g, \pi_{\vec{x}}) = \sum_{\vec{z} \in Z} [NDCG(f_R(\vec{z})) - NDCG(g(\vec{z}))]^2\ \pi_{\vec{x}}(\vec{z})$$

Table 1 – Analyzing competing attribution methods for axiomatic compliance.

Algorithm	**R-Efficiency**	**R-Missingness**	**R-Symmetry**	**R-Monotonicity**
EXS	✗	✗	✗	✗
RankLIME	✗	✗	✗	✗
RankingSHAP	✗	✗	✓	✗
RankSHAP	✓	✓	✓	✓

RankSHAP Experiments : Key Insights

- **Across Systems :** RankSHAP outperforms competitors by :
  - **25.78%** on Fidelity and **19.68%** on weighted Fidelity.
  - Consistently demonstrates positive correlation with original rankings (except Random).
- **Impact of Document Set Size :**
  - **20% drop** from 10 to 20 documents and **14.6% drop** from 20 to 100 documents.
  - Performance declines logarithmically with document set size.
- **Across Datasets :**
  - Performance **5.7% lower** on Robust04 compared to MS MARCO.
  - Lower performance attributed to dataset size and cross-dataset fine-tuning.
- **Across Ranking Models :**
  - Best performance on BM25. Fidelity on BERT/T5 is **13%-15% lower** than BM25.
  - Slightly better performance on LLAMA2 compared to BERT/T5.

RankSHAP : User Study

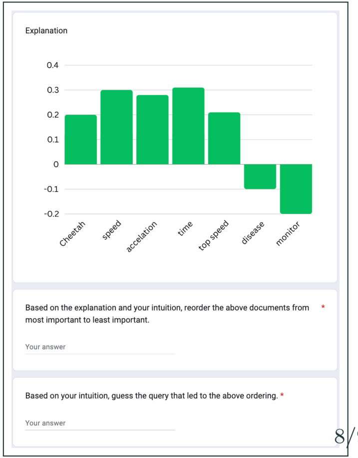
Goal : Evaluate if attributions help understand *why* items are ranked a particular way.

Tasks :

1. Reorder passages by **perceived importance** (using attributions).
2. Estimate the **query** from displayed feature attributions.

Table 2 – Higher is better (Fidelity & Semantic Sim.)

	Random	EXS	LIME	RiSHAP	**Ours**
Q1 (τ)	0.23	0.43	0.47	0.52	**0.56**
Q2 (Sim)	0.30	0.48	0.52	0.58	**0.69**



Thank you !

Questions ? Feel free to reach out at

tchowdhury@cs.umass.edu