Desirable Properties for Diversity and Truncated Effectiveness Metrics

Ameer Albahem[1] Damiano Spina[1] Falk Scholer[1] Alistair Moffat[2] Lawrence Cavedon[1]

[1] RMIT UNIVERSITY

[2] THE UNIVERSITY OF MELBOURNE

The 23rd Australasian Document Computing Symposium
11-12 December 2018, Dunedin, New Zealand

Problem

- Complex search scenarios such as dynamic search are evaluated based on several dimensions: relevance, novelty and effort
- As metrics model multiple dimensions, understanding their behavior and suitability to search tasks is critical
- Existing meta-analysis approaches either do not model the dimensions or do not quantify the behavior

Contribution

A meta-analysis framework that can:

- Model desirable evaluation dimensions separately or simultaneously
- Quantify metrics behavior
- Provide cases for counter-intuitive behavior.

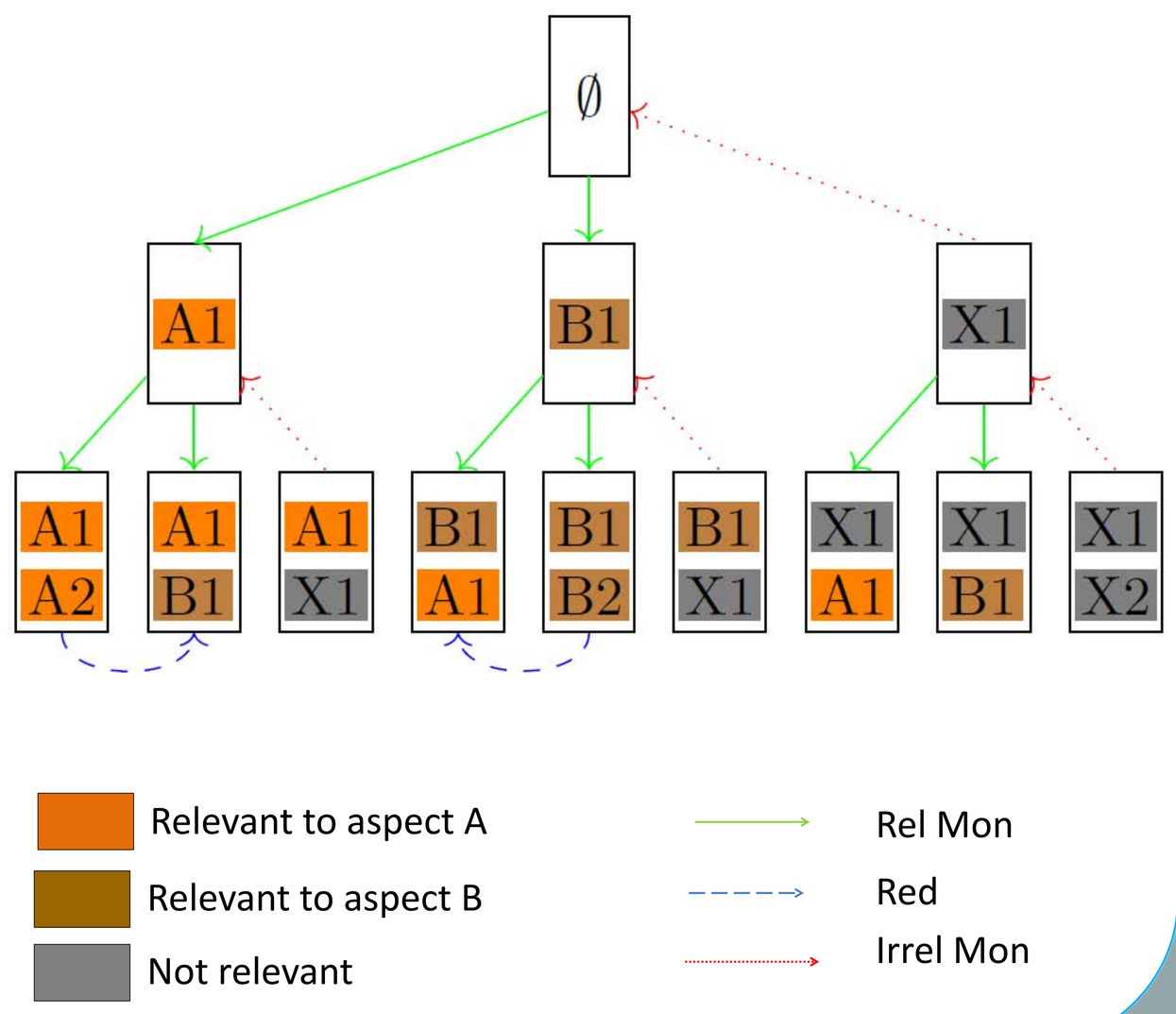
Framework

Properties:

- **Relevance Monotonicity (Rel Mon):** If a ranking is extended by a single relevant document, then the computed effectiveness score should not decrease, but might remain the same
- **Irrelevance Monotonicity (Irrel Mon)**: if a ranking is extended by a single non-relevant document, then the computed effectiveness score should not increase
- **Redundancy (Red)**: A diversification metric should not favor a ranking that adds a document relevant to an already-seen aspect over a ranking that adds a relevant document to an unseen aspect
- **Induction**: If there is a directed path from the ranking *Y* to the ranking *Z*, then *Y* is inferior or equal to ranking *Z*

Case-analysis steps:

- Generate all cases of a truncated and diversified ranking of a maximum length of ***m***
- Build tree-based relationships of rankings using properties.
- Evaluate all generated rankings using the metrics
- For every metric, check and count cases where a metric violates any of the properties proposed



Analysis

Setup:

- Truncated and Diversified rankings of up to *m*=10 documents
- Two aspects (A and B)

Metrics:

- **Ad hoc**: RR, P@5, P@10, NDCG@5, NDCG@10 and AP
- **Diversity**: Subtopic Recall; Intent-Aware: Precision (P-IA), Average Precision (AP-IA), and Diversity-aware Expected Reciprocal Rank, α-NDCG, and Novelty and Rank Biased Precision (NRBP)
- **Diversity and Truncation**: Cube Test (CT), Normalized Cube Test (nCT), and the Average Cube Test (ACT)

Number of cases (number of edges)

Table. Violation summary of properties and metrics

Metric	Properties				
	Irrel Mon (29,523)	Rel Mon (59,046)	Red (2,026)	Induction (90,595)	Total (1,033,032)
ACT	29,496	0	0	81,340	110,836
AP-IA	0	0	2,026	6,892	8,918

Number of violations