Estimating Topic Difficulty Using Normalized Discounted Cumulated Gain

Lukas Gienapp   Benno Stein   Matthias Hagen   Martin Potthast

Leipzig University

Bauhaus-Universität Weimar

Martin-Luther-Universität Halle-Wittenberg

webis.de

29th ACM International Conference on Information and Knowledge Management (CIKM 2020)

How can we identify topics in offline IR evaluation for which systems (systematically) face retrieval problems?

Experimental Setting

Offline IR Evaluation

$$\begin{pmatrix} p_{1,1} & \cdots & \cdots & p_{t,1} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{1,s} & \cdots & \cdots & p_{t,s} \end{pmatrix}$$

Topic-System-Matrix:

- $p_{t,s}$ denotes effectiveness score of system s on topic t w.r.t. a measure on the relevance judgements

Experimental Setting

Offline IR Evaluation

$$\begin{pmatrix} p_{1,1} & \cdots & \cdots & p_{t,1} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{1,s} & \cdots & \cdots & p_{t,s} \end{pmatrix}$$

Topic-System-Matrix:

- $p_{t,s}$ denotes effectiveness score of system s on topic t w.r.t. a measure on the relevance judgements
- System performance: row-based *aggregation* (mean) of system s over all topics T

Experimental Setting

Offline IR Evaluation

$$\begin{pmatrix} p_{1,1} & \cdots & \cdots & p_{t,1} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{1,s} & \cdots & \cdots & p_{t,s} \end{pmatrix}$$

Topic-System-Matrix:

- $p_{t,s}$ denotes effectiveness score of system s on topic t w.r.t. a measure on the relevance judgements
- System performance: row-based *aggregation* (mean) of system s over all topics T
- Topic difficulty: column-based *aggregation* of topic t over all systems S

Experimental Setting

Offline IR Evaluation

$$\begin{pmatrix} p_{1,1} & \cdots & \cdots & p_{t,1} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ p_{1,s} & \cdots & \cdots & p_{t,s} \end{pmatrix}$$

Topic-System-Matrix:

- $p_{t,s}$ denotes effectiveness score of system s on topic t w.r.t. a measure on the relevance judgements
- System performance: row-based *aggregation* (mean) of system s over all topics T
- Topic difficulty: column-based *aggregation* of topic t over all systems S

Research Questions:

→ What is a suitable *aggregation* method for topic difficulty estimation?

→ How can it be applied in practice with minimal overhead?

Limitations of Existing Approaches

(1) Local inconsistency:

- Problem: results are incomparable between experiments
- Solution: standardized aggregation techniques

Limitations of Existing Approaches

(1) Local inconsistency:

- Problem: results are incomparable between experiments
- Solution: standardized aggregation techniques

(2) Topic set instability:

- Problem: topic ratings depend on each other
- Solution: aggregation method using only information within topic

Limitations of Existing Approaches

(1) Local inconsistency:

- Problem: results are incomparable between experiments
- Solution: standardized aggregation techniques

(2) Topic set instability:

- Problem: topic ratings depend on each other
- Solution: aggregation method using only information within topic

(3) Experimental inconsistency

- Problem: different measures used for topic difficulty and system performance
- Solution: use nDCG for both system performance & topic difficulty

Limitations of Existing Approaches

(1) Local inconsistency:

- Problem: results are incomparable between experiments
- Solution: standardized aggregation techniques

(2) Topic set instability:

- Problem: topic ratings depend on each other
- Solution: aggregation method using only information within topic

(3) Experimental inconsistency

- Problem: different measures used for topic difficulty and system performance
- Solution: use nDCG for both system performance & topic difficulty

(4) Discrete class labeling

- Problem: difficulty expressed as classes (“easy”, “hard”, ...)
- Solution: aggregation resulting in numerical scale

Ratio-based Topic Difficulty

Requirements:

- ❑ Aggregation method should not be subject to mentioned limitations
- ➜ Any kind of aggregation derived from a distribution over all topics is unsuitable

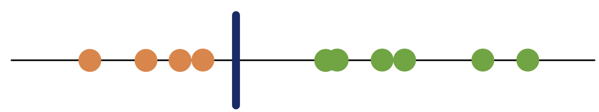
Ratio-based Topic Difficulty

Requirements:

- ❑ Aggregation method should not be subject to mentioned limitations
- ➜ Any kind of aggregation derived from a distribution over all topics is unsuitable

Solution:

- ❑ Difficulty is expressed as ratio
- ➜ Systems scoring higher than a baseline to overall number of systems



= 0.6

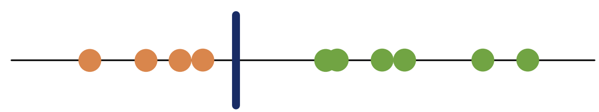
Ratio-based Topic Difficulty

Requirements:

- ❑ Aggregation method should not be subject to mentioned limitations
- ➜ Any kind of aggregation derived from a distribution over all topics is unsuitable

Solution:

- ❑ Difficulty is expressed as ratio
- ➜ Systems scoring higher than a baseline to overall number of systems



= 0.6

Issues solved:

- ❑ *Topic set instability* – topics are now scored independently
- ❑ *Discrete class labeling* – ratio is numerical value between 0 and 1

Problem: What is a sensible baseline?

Ratio-based Topic Difficulty

Hypothetical Random Baseline Ranking

Requirements for a baseline:

- Domain-agnostic and comparable
- Should be applicable to every experiment
- Does not create experimental overhead

Ratio-based Topic Difficulty

Hypothetical Random Baseline Ranking

Requirements for a baseline:

- Domain-agnostic and comparable
- Should be applicable to every experiment
- Does not create experimental overhead

Proposed: hypothetical random ranking as as baseline

= A system drawing documents at random

- Restricted to random permutations of the pooling for practicability
- Its nDCG performance approaches the mean of the relevance label distribution

→ **Baseline**: mean relevance of judged documents to compare systems to

Ratio-based Topic Difficulty

Baseline Standardization

Procedure:

- Standardize the relevance label distribution (z-transformation)
- Baseline nDCG is 0 across all experiments
- Standardization affects nDCG scores linearly (proof in the paper)

Ratio-based Topic Difficulty

Baseline Standardization

Procedure:

- Standardize the relevance label distribution (z-transformation)
- Baseline nDCG is 0 across all experiments
- Standardization affects nDCG scores linearly (proof in the paper)

Benefits:

- Improves on the *local inconsistency* issue
- Intra-experiment results are unaffected
- Inter-experiment comparability is improved

→ transforms baseline into well-defined reference point

Ratio-based Topic Difficulty

Summary

Our novel measure can be simplified to the following three steps:

(1) Standardize the relevance label distribution of the topics' pooling

➜ improves *local inconsistency* issue

(2) Calculate nDCG scores

➜ solves *experimental inconsistency* issue

(3) Ratio of positive-scoring systems to total number of systems denotes difficulty

➜ solves *topic set instability* issue

➜ solves *discrete class labeling* issue

Conclusion

Our contribution:

- novel method of scoring difficulty of topics
- overcomes several existing limitations
- does not add any experimental requirements

Also included in the paper:

- reevaluation of TREC data to illustrate the practical advantages
- formal proof of the linear shift property of nDCG
- concept of random baseline ranking with potential applications beyond topic difficulty estimation

Conclusion

Our contribution:

- novel method of scoring difficulty of topics
- overcomes several existing limitations
- does not add any experimental requirements

Also included in the paper:

- reevaluation of TREC data to illustrate the practical advantages
- formal proof of the linear shift property of nDCG
- concept of random baseline ranking with potential applications beyond topic difficulty estimation

Thank you!