

# Component-level IR Evaluation

## A large-scale study across text collections

### Approach

- 1) from system to **component-based evaluation** of IR systems and modern search engine APIs
- 2) aggregating cumulated observations from different evaluations using a **large-scale experiment** testing about **12,000 different IR systems** against numerous collections
- 3) investigating three main parameters of IR systems: **Stemming, Ranking, and Relevance Feedback**

### Findings

- 1) **empirical verification** of dependency relationship between test collections and IR system configurations
- 2) some IR system configurations achieve high effectiveness consistently across test collections and could serve as **standard baselines in evaluation**
- 3) hard to account for **variance in test collection** setup

### Experimental Results

Stemmer	Ranking	PRF(d,t)*	Avg. MAP**
-	-	-	0.3765
Porter	DLH13	KL(6,30)	0.3578
Porter	TF_IDF	KL(9,80)	0.3518
Krovetz	DLH13	KL(6,20)	0.3503
Porter	In_expB2	KL(6,40)	0.3501
Porter	DFR_BM25	KL(6,100)	0.3497
Porter	BM25	KL(6,100)	0.3495
Porter	DFR	KL(3,20)	0.3480
Porter	IFB2	KL(6,40)	0.3464
UeaLite	DLH13	KL(6,100)	0.3450
Porter	BB2	KL(6,10)	0.3448

\* pseudo relevance-feedback using d documents and t terms and Kullback-Leiber (KL) model

\*\* averaged mean average precision across test collections used in a grid-like experiment setup

### Future Work

- 1) deep analysis: **statistical significance tests** using randomization, comparing system parameters independently
- 2) applications: **learning to select a system configuration** using test collection classification or query performance prediction



CHEMNITZ UNIVERSITY  
OF TECHNOLOGY

Empirical analysis of differences in key components of modern IR toolkits and APIs: **Why do we NOT have commonly accepted baselines for standard test collections in IR?**

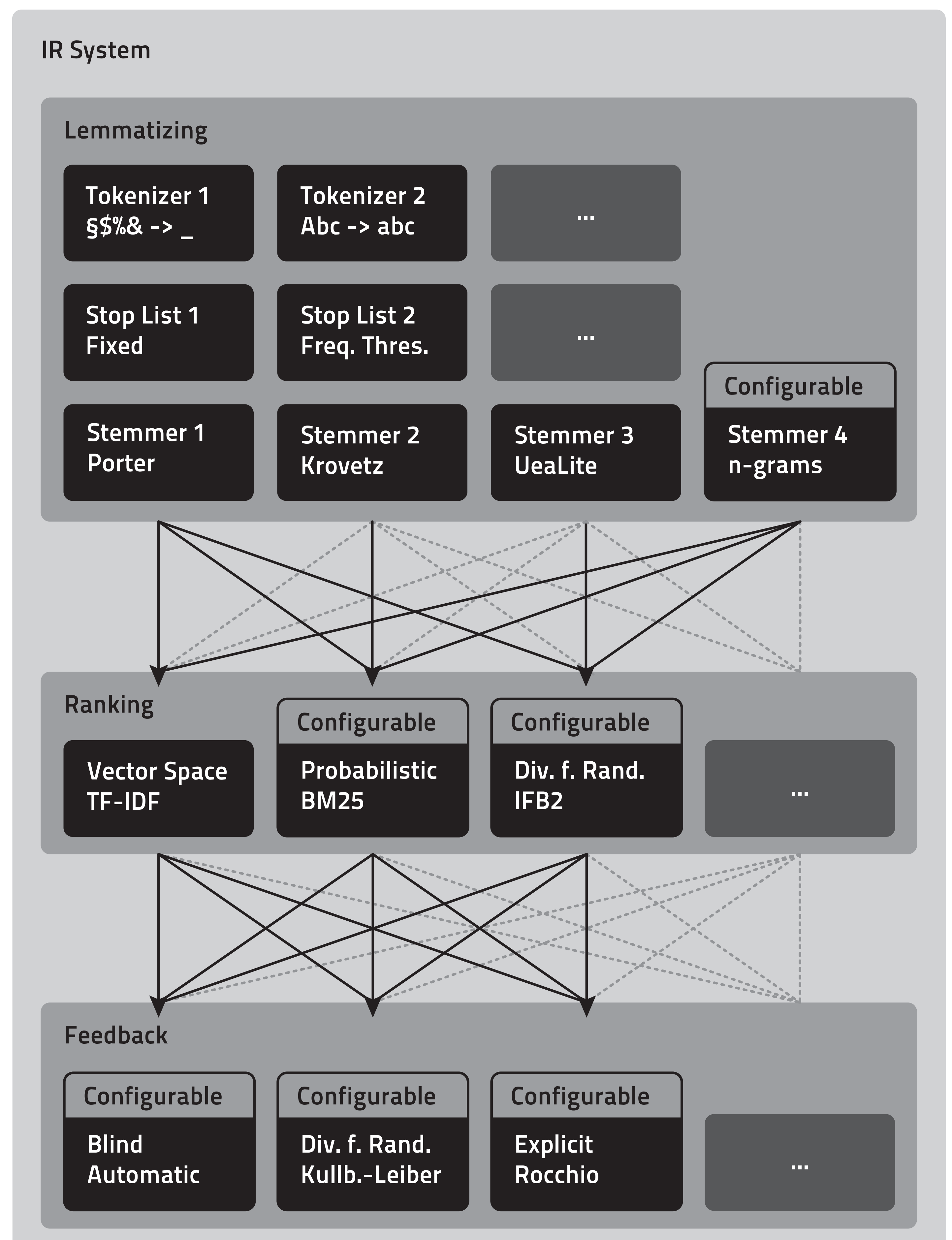
### IR Evaluation Methodology: Researchers' Perspective

Due to the complexity and number of components in current IR systems improving retrieval performance is usually limited to optimize a single component (by developing, implementing and evaluating it) and leaving out the remaining components or at least fix them to a certain extent.

**Therefore it is almost impossible to draw generally valid conclusions.**

### Possible Directions

The research community should develop a standard format to **capture component-level configuration information at every IR evaluation task**. It should be easy to generate and process, but also flexible. All results of the contributions to a specific task could **serve as ground truth to derive a common baseline for reference**. That baseline might be the average/median/best over all contributions.



INNOPROFILE  
UNTERNEHMEN  
REGION  
Die BMBF-Innovationsinitiative  
Neue Länder

SPONSORED BY THE



Federal Ministry  
of Education  
and Research

Jens Kürsten · Maximilian Eibl  
firstname.lastname@  
informatik.tu-chemnitz.de

Professur Medieninformatik  
Technische Universität Chemnitz  
Fakultät für Informatik  
Straße der Nationen 62  
09107 Chemnitz · Germany

This publication was prepared as a part of the research initiative **sachsMedia**, which is funded by the German Federal Ministry of Education and Research under the grant reference number 03IP608. The authors take sole responsibility for the contents of this publication.