CNRS - Toulouse INP - UT3 - UT Capitole - UT2

Institut de Recherche en Informatique de Toulouse

Prédiction de la performance et traitement sélectif des requêtes dans les moteurs de recherche d'information

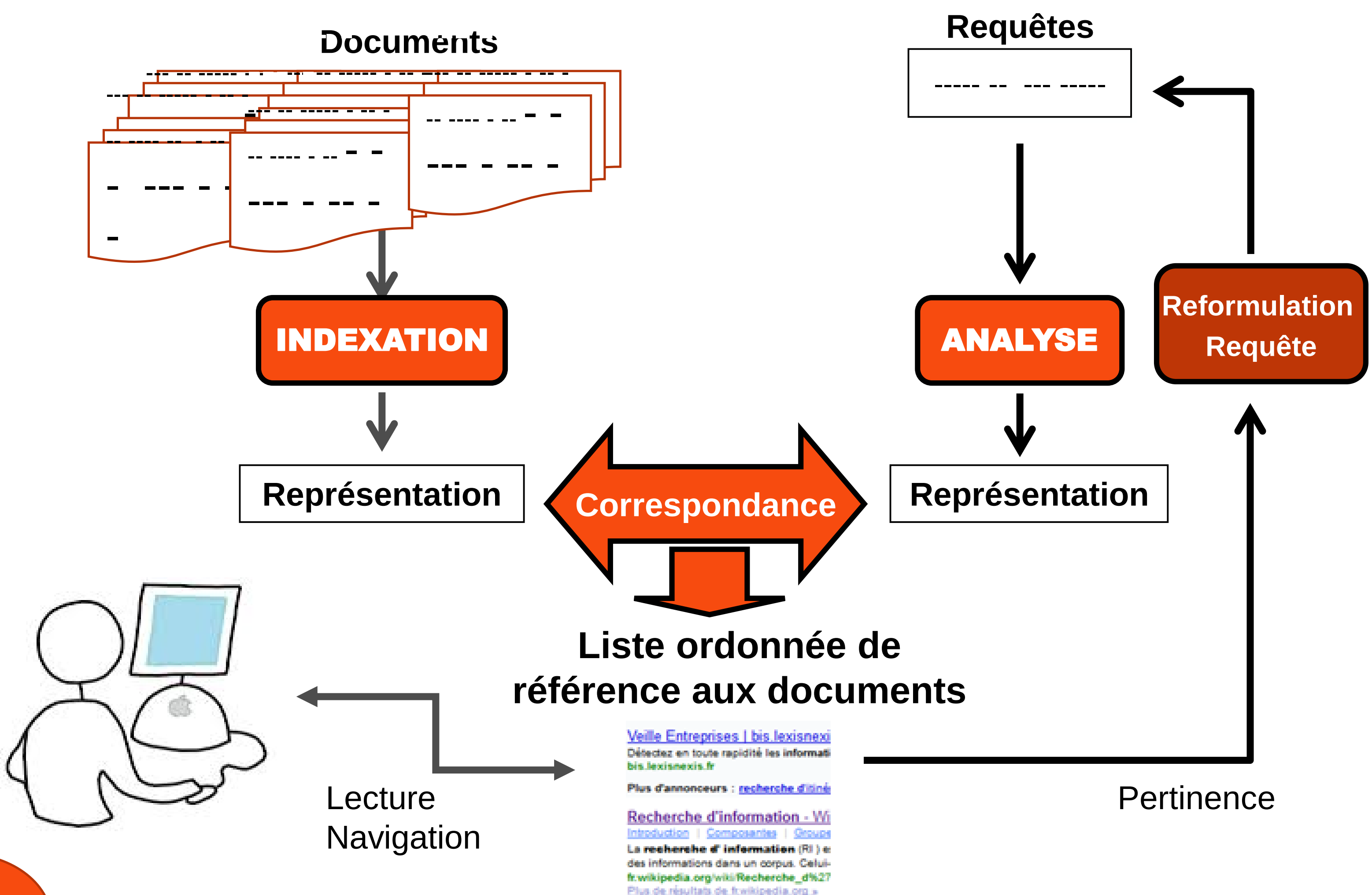
Josiane Mothe

Professeure

INSPE, UT2J

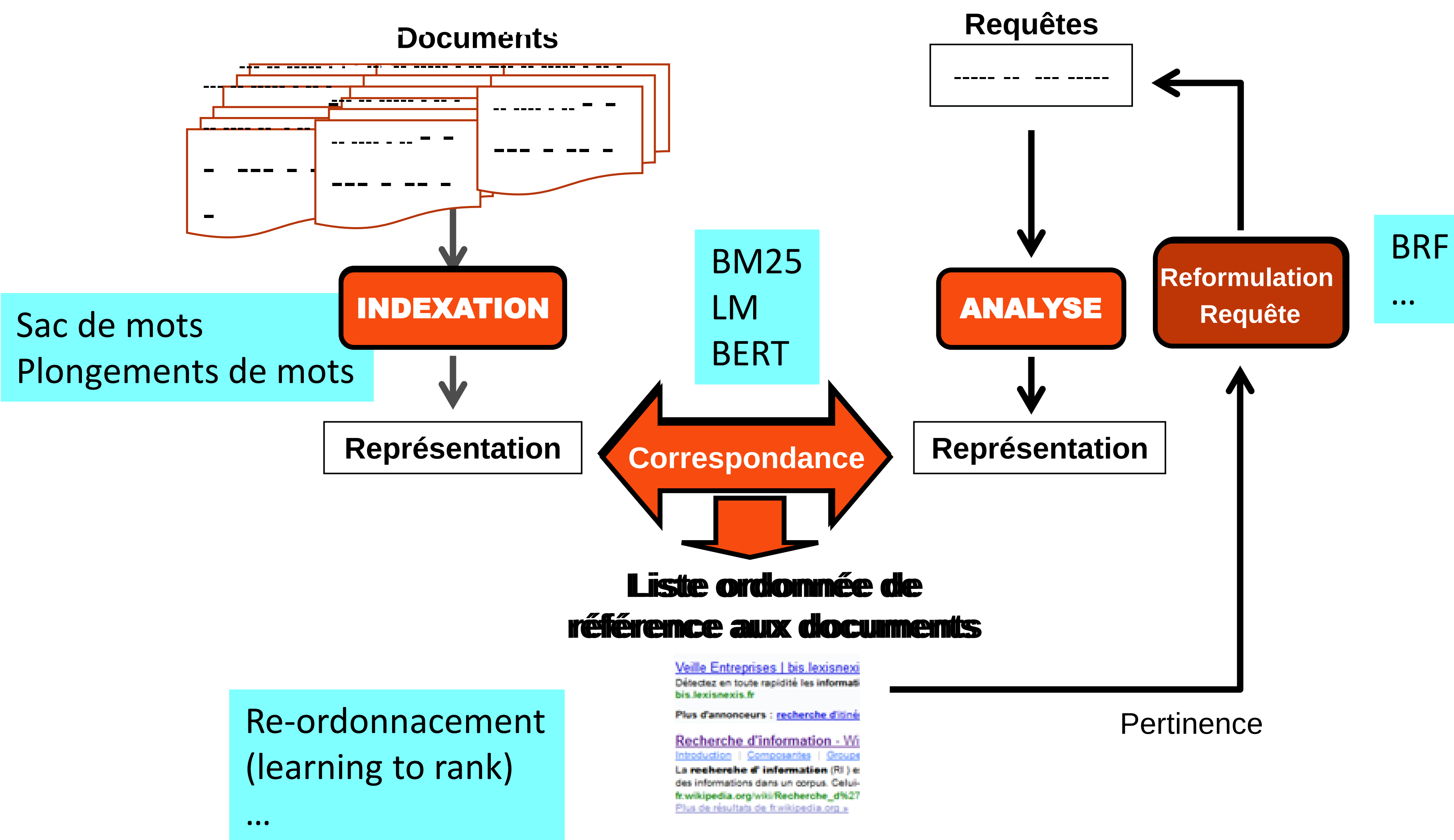
Novembre 2023

Recherche d'Information



Collecte de logs de connexion

Recherche d'Information - variantes



Exemples de paramètres des systèmes de RI dans Terrier

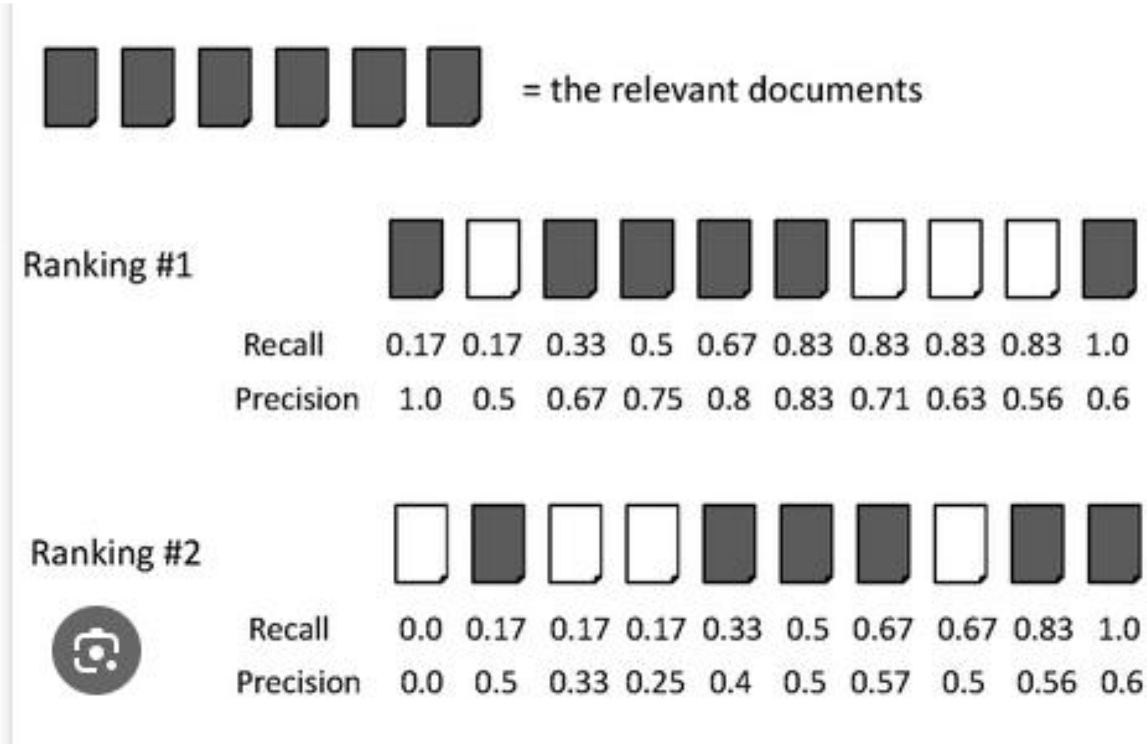
Parameters	Meaning	Values
Top	Topic number	351, …, 400
Field	Topic field	T, T+D, T+D+N
Bloc	Size of the indexing bloc	1, 5, 10
Idf	Inverse Document Frequency	FALSE, TRUE
Ref	Query reformulation	None, Bo1bfree, Bo2bfree, KLbfree
Model	Retrieval model	BB2c1, BM25b0.5, DFRBM25c1.0, IFB2c1.0, InexpB2c1.0, InexpC2c1.0, InL2c1.0, PL2c1.0, TFIDF
DocNb	Number of documents (reformulation)	0, 3, 5, 10, 50, 100, 200
qe_md	Minimum number of documents in which the term should appear to used in the query expension	0, 2
qe_t	Number of terms used in the query expension	0, 1

Evaluation de la Recherche d'Information

- Collections
  - Documents
  - Requêtes
  - Réponse attendue (QREL)
- Mesure d'évaluation
  - Rappel
  - Précision
  - P@10
  - AP
  - ....

- Test collections
  - TREC78 -- 100 Topics (351 – 450)
  - WT10G -- 100 Topics (451 – 550)
  - GOV2 -- 150 Topics (701 – 850)

Evaluation de la Recherche d’Information

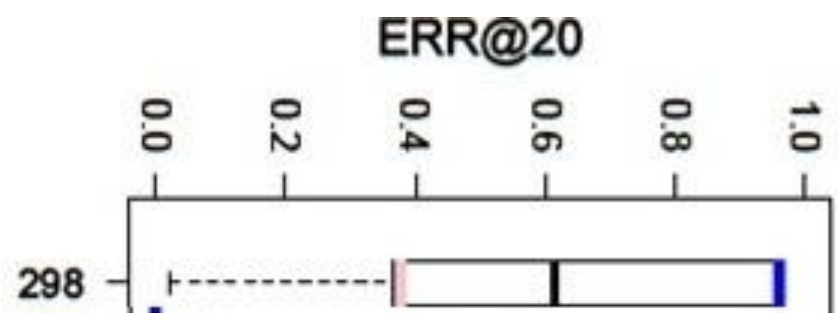


P@10
AP (Average precision)
nDCG (Normalized DCG)

https://stackoverflow.com/questions/40801196/some-ideas-and-direction-of-how-to-measure-ranking-ap-map-recall-for-ir-evalu

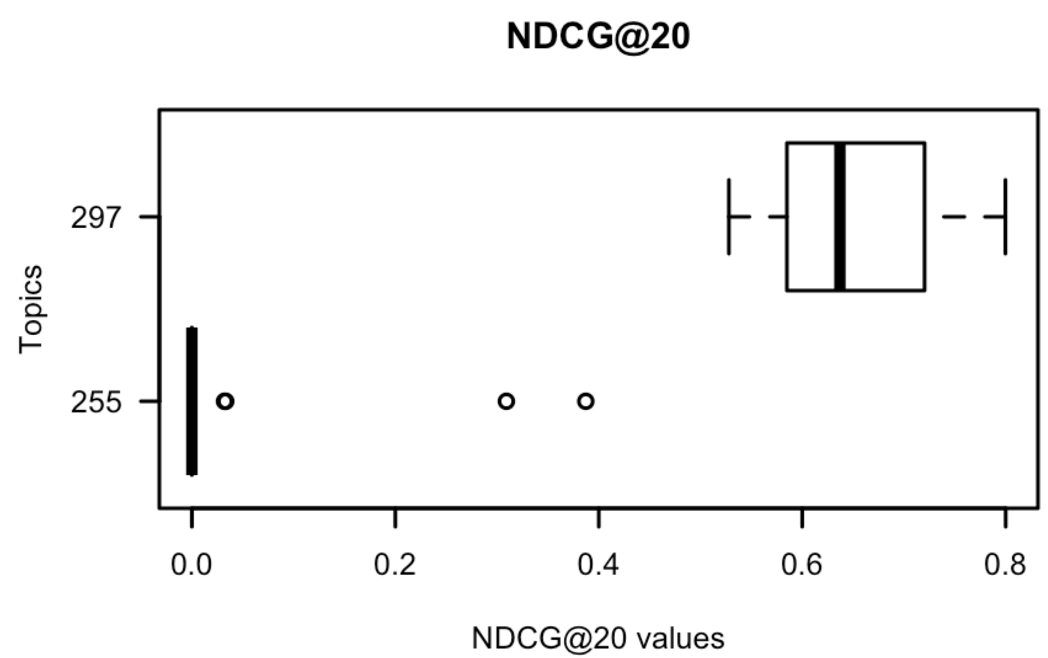
Query difficulty

- Search engines have an answer whatever the query is BUT
- Evaluation compaigns showed
  - System variety (the difficulty depends on the system



Query difficulty

- Evaluation compaigns showed
  - System variety
  - Some queries are easy, some are difficult



Query difficulty

- What is a difficult query ?

- (IR) Defined regarding system effectiveness

Difficult topic = Poor effectiveness

- (Psy) Defined regarding human difficulty

Difficult task = hard for users (cognitive)

Query difficulty

- Back to the Reliable Information Access (RIA) Workshop (2004)

[Harman, 2009, IR journal]

Main research directions

- Query difficulty prediction
  - Predict whether a query is difficult or not
  - Performance prediction: Predict the value of the effectiveness measure
- Adaptive systems / selective query processing
  - Different systems (parameters) for different queries
- User studies
  - Measure users’ abilities with regard to query difficulty

Query difficulty prediction

- Why?

  - To handle differently queries

Examples?

***Selective query expansion**: the system decides whether the query should be expanded or not [Amati et al., 2004]*

***Adaptive system**: the system adjusts its parameters according to the query features [Deveaud et al., 2016]*

Query difficulty prediction

- Types

  - Pre-retrieval vs Post-retrieval

Definition
and examples?

Pre-retrieval:
does not need to process the query over the document collection

Post: does need

  - Based on Statistics vs Linguistics

Examples?

Query difficulty prediction

- Examples
  - IDF : min, max, mean, … of the IDF of the query terms
  - SynSet: … number of synonyms of the query terms [Mothe & Tanguy, 2005]
  - Query scope: ratio of the documents that contain at least one query term [Kanoulas et al., 2017]
  - Query Feedback (QF) : overlap between these two retrieved document lists [Zhou & Croft, 2007]
  - Weighted Information Gain (WIG) : divergence between the mean of the top-retrieved document scores and the mean of the entire set of document scores [Zhou & Croft, 2007]
  - Normalized Query Commitment (NQC) : standard deviation of the retrieved document scores [Shtok et al., 2009]
  - Clarity score: KL-divergence between the LM of the retrieved documents and the LM of the document collection [Cronen-Townsend & Croft, 2002]
  - Letor features: agregations of document scores [Chifu et al. 2018]

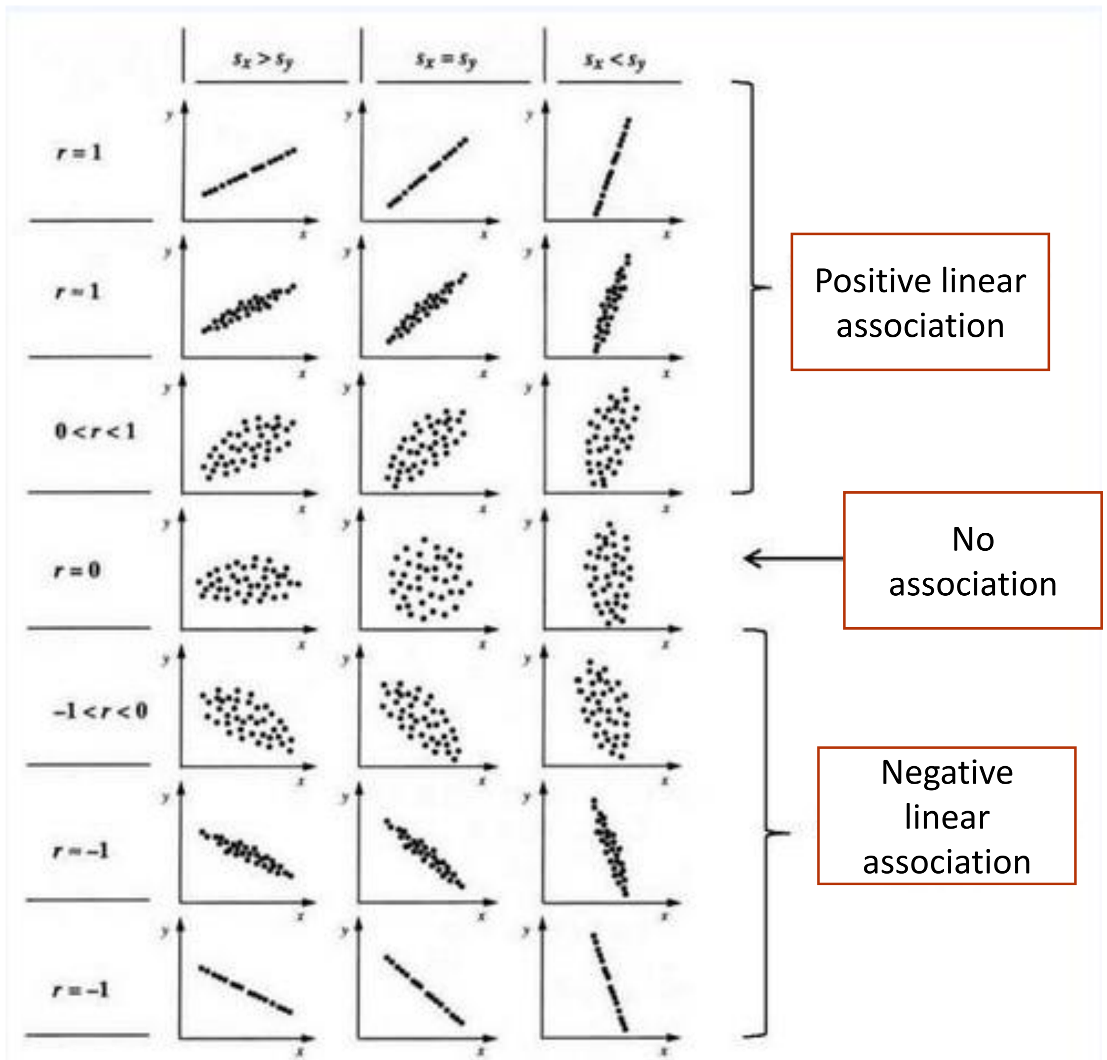
Evaluation of query difficulty predictors

- How to evaluate whether a feature is a good predictor?

  - Correlation on values (Bravais-Pearson) or on ranks (Kendall or Spearman)

Interpretation ?

Linear correlation Bravais-Pearson



From Wikipedia

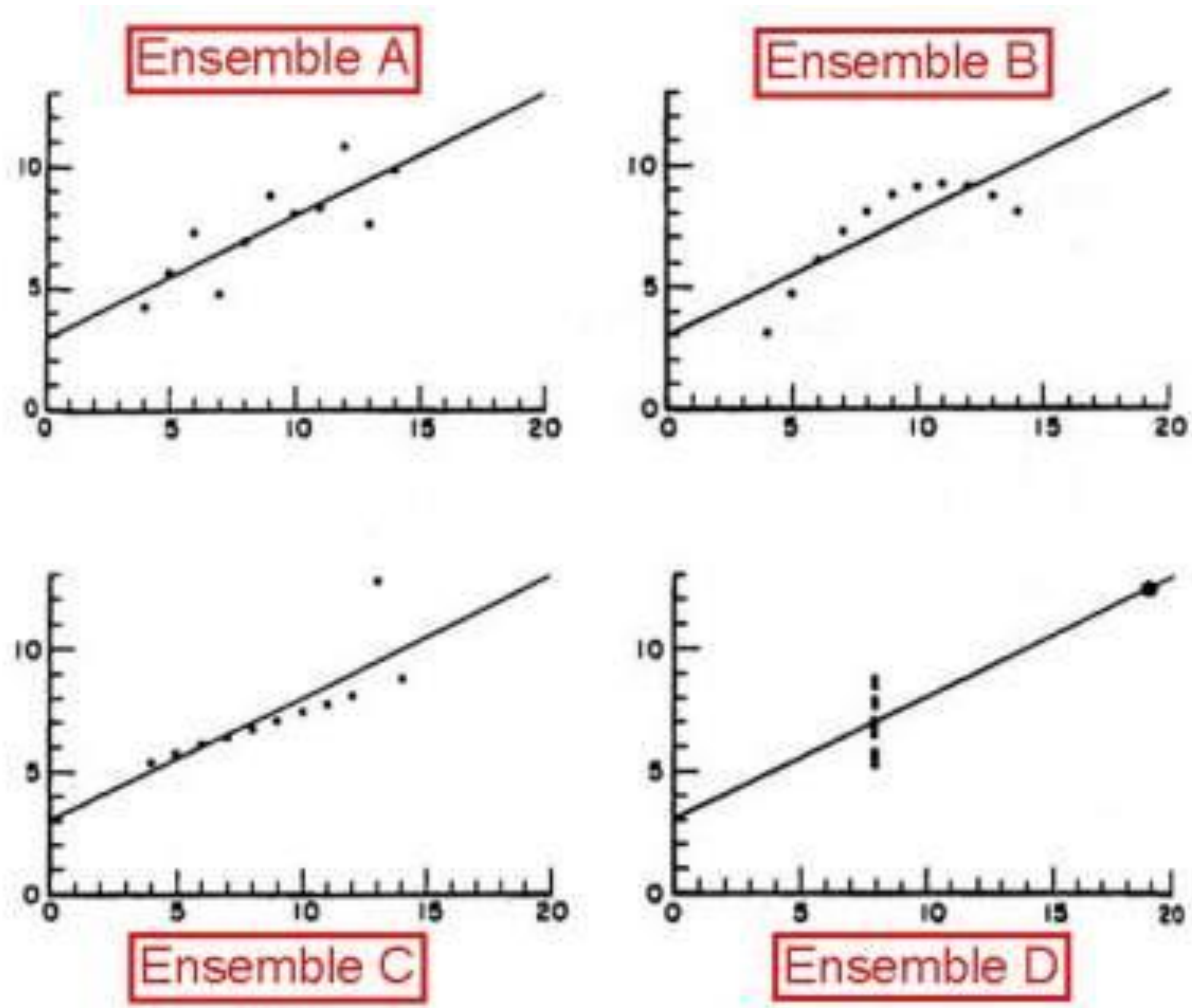
Linear correlation Bravais-Pearson

Anscombe data sets

Data set A		Data set B		Data set C		Data set D	
x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

$n = 11$, $\bar{x} = 9$, $\bar{y} = 7.5$, $s_x^2 = 10$, $s_y^2 = 3.75$, $s_{xy} = 5$. $r = 0.816$,

Linear correlation Bravais-Pearson

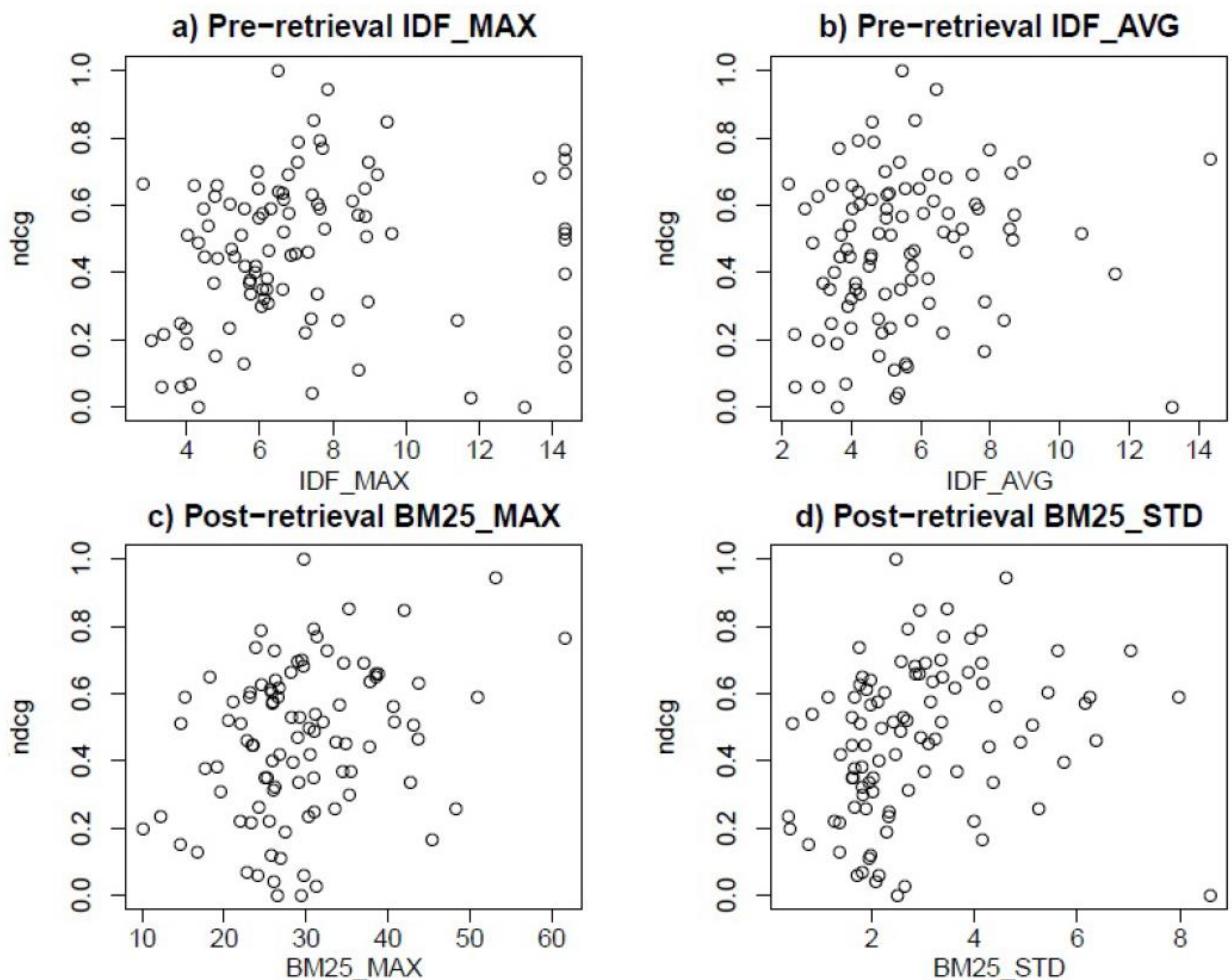


Correlation values should be consider with caution

Measure	Feature BM25_MAX	BM25_STD	IDF_MAX	IDF_AVG
Pearson ρ	$\boxed{0.294^*}$	0.232^*	0.095	0.127
Spearman r	0.260^*	0.348^*	0.236^*	0.196
Kendall τ	0.172^*	0.230^*	0.159^*	0.136^*

correlation	Feature BM25_MAX	BM25_STD	IDF_MAX	IDF_AVG
Removing topic 463 only				
ρ	0.294^*	$\boxed{0.339^*}$	0.142	0.225^*
r	0.268	0.342	0.234	0.183
τ	0.181^*	0.225	0.162^*	0.120

Correlation values should be consider with caution



Query difficulty prediction

		TREC Vol. 4+5			WT10g			GOV2		
		μ100	μ1500	μ5000	μ100	μ1500	μ5000	μ100	μ1500	μ5000
SPECIFICITY	AvQL[6]	0.13	0.14	0.16	-0.11	-0.14	-0.12	-0.05	0.02	0.03
	AvIDF[3]	*0.52**	0.53*	0.59*	0.21*	0.18	0.18	0.37*	0.32*	0.39*
	MaxIDF[9]	*0.52**	*0.54**	*0.60**	0.31*	0.30*	0.30*	0.35*	0.35*	0.43*
	DevIDF[4]	0.22*	0.24*	0.26*	0.21*	0.25*	0.27*	0.14	0.20*	0.27*
	AvICTF[4]	0.50*	0.50*	0.56*	0.20	0.16	0.16	0.34*	0.30*	0.37*
	SCS[4]	0.49*	0.49*	0.55*	0.15	0.13	0.13	0.31*	0.26*	0.34*
	QS[4]	0.42*	0.42*	0.47*	0.09	0.05	0.05	0.26*	0.18*	0.22*
	AvSCQ[11]	0.25*	0.27*	0.31*	0.32*	0.30*	0.30*	0.40*	0.36*	0.39*
	SumSCQ[11]	-0.01	0.00	0.00	0.20*	0.18	0.15	0.23*	0.23*	0.19*
	MaxSCQ[11]	0.32*	0.35*	0.38*	*0.36**	0.41*	0.45*	0.39*	0.42*	*0.46**
AMBI	AvQC[5]	0.45*	0.47*	0.51*	0.18	0.17	0.17	0.28*	0.31*	0.38*
	AvQCG[5]	0.33*	0.34*	0.37*	0.00	-0.03	-0.03	0.04	0.05	0.08
	AvNP[6]	-0.20*	-0.23*	-0.26*	-0.09	-0.10	-0.10	-0.06	-0.04	-0.05
	AvP	-0.11	-0.12	-0.14	-0.17	-0.18	-0.17	0.02	0.01	0.00
REL	AvPMI	0.37*	0.35*	0.39*	0.33*	0.28*	0.26*	0.26*	0.29*	0.33*
	MaxPMI	0.30*	0.30*	0.33*	0.31*	0.27*	0.24*	0.28*	0.31*	0.32*
	AvLesk[2]	0.24*	0.25*	0.27*	0.00	0.01	0.02	0.04	0.08	0.11
	AvPath[8]	0.12	0.14	0.16	0.01	0.04	0.05	-0.02	0.03	0.07
	AvVP[7]	0.25*	0.25*	0.27*	-0.06	-0.06	-0.05	-0.01	0.09	0.13
RNK	AvVAR[11]	0.50*	0.52*	0.56*	0.29*	0.29*	0.30*	*0.43**	0.40*	0.42*
	SumVAR[11]	0.28*	0.30*	0.31*	0.31*	0.29*	0.28*	0.33*	0.34*	0.30*
	MaxVAR[11]	0.48*	0.52*	0.54*	*0.36**	*0.42**	*0.47**	0.40*	*0.43**	*0.46**

Table 1: Results of the predictor evaluations given by the linear correlation coefficient.

Hauff et al., 2008, CIKM

Linguistic query difficulty predictors

- Pre-retrieval
- Linguistic-based

J. Mothe and L. Tanguy. Linguistic features to predict query difficulty. In *Predicting query difficulty - methods and applications Workshop, Int. Conf. on Research and Development in Information Retrieval, SIGIR*, pages 7–10, 2005.

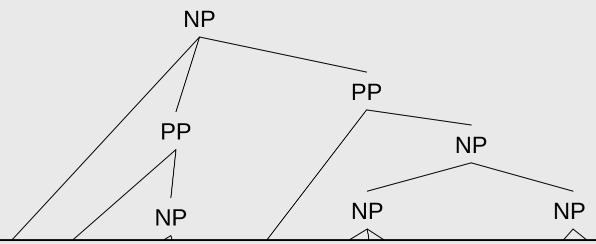
Linguistic query difficulty predictors

Method and data

- Queries
  - 200 TREC queries (TREC 3, 5, 6 and 7)
  - Title query (closest to real users'queries)
  - Feature extraction
- Participants' runs – adhoc task

	TREC 3	TREC 5	TREC 6	TREC 7
# runs	40	61	80	103
# queries	50	50	50	50

Syntactic depth vs span (2)



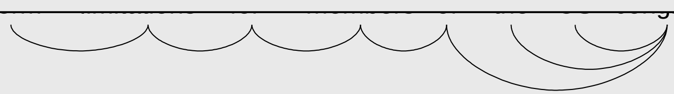
Polysemy value:

WordNet

# of synset s a term belongs to
Default value : 1

Syntactic depth

syntactic complexity
in terms of hierarchy



TREE TAGGER (Schmidt)
part-of-speech tagger

For example, topic 158
Term limitations for members of the U.S. Congress

Term/NN
limitations/NNS
for/IN
members/NNS
of/IN
the/DT
U.S./NP

TREE TAGGER (Schmidt)

terms that are not in
its reference wordlist

Example:
“postmenopausal”, “multilingualism”

Linguistic query difficulty predictors

Analysis

- Correlations
  - Correlation between recall and features
  - Correlation between precision and features
  - Pearson coefficient [-1,1]
    - The higher => the stronger correlation
    - Positive or negative correlation
  - Significance p-value
    - Estimate prob. of correlation being due to random
    - The smaller => the higher confidence

Linguistic query difficulty predictors

Analysis

- Results

TREC Campaign	*Significant variables for Recall*	*Significant variables for Precision*
TREC 3	- PREP - SYNTDEPTH **- SYNSETS**	- SUFFIX - NBWORDS - CC
TREC 5		**- SYNTDIST** - SYNTDEPTH
TREC 6	**- SYNSETS** + PN	
TREC 7	**- SYNSETS**	+ PN - LENGTH **- SYNTDIST**

**Significant correlations (p-value <= 0.05) between
linguistic features and
recall / precision**

Letor features as predictors

- Letor features:
  - query-document scores, aggregated over the documents for a query

Table Pearson correlation of the WMODEL:DFIZ_std [13] SLF predictor according to n, the number of top-ranked retrieved documents considered when calculating the feature.

	n	5	10	50	100	500	1000
AP	Robust	.056	.065	.089	.081	.146*	.191*
	WT10G	.027	-.084	.163	.142	.175	.217*
	GOV2	.261*	.385+	.409+	.407+	.453+	.453+
	ClueWeb12B	.320*	.255*	.266*	.243*	.269*	.298*
NDCG	Robust	.081	.119	.183*	.191*	.252+	.306+
	WT10G	.069	-.032	.224*	.217*	.294*	.321*
	GOV2	.285+	.397+	.426+	.418+	.453+	.447+
	ClueWeb12B	.246*	.234*	.253*	.224*	.265*	.301*

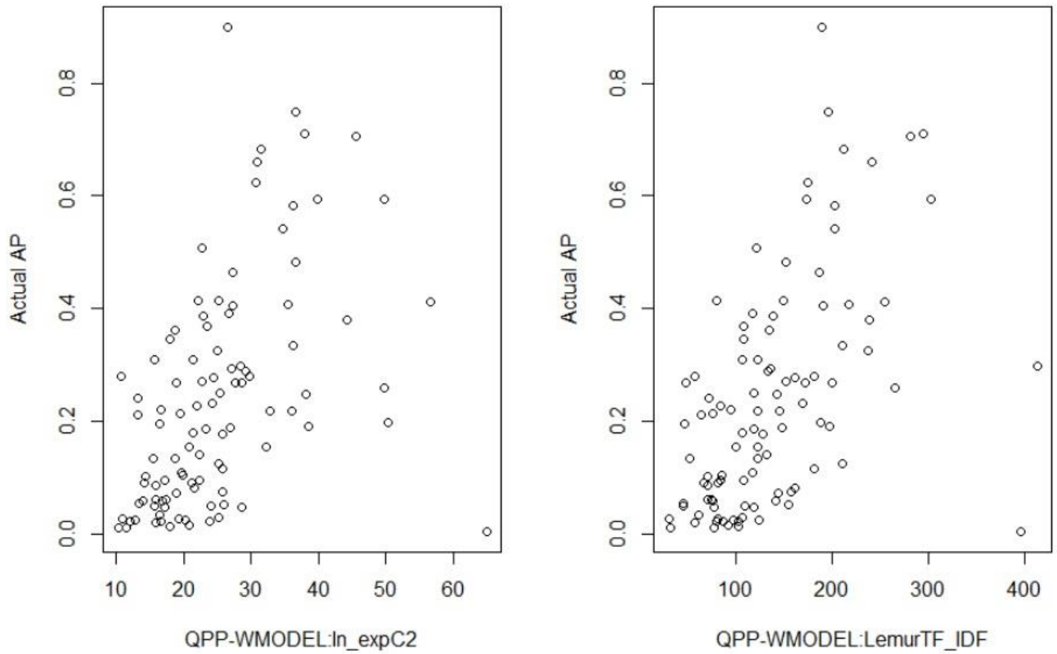
Combination of Letor features

Table Performance of linear regression of combined post-retrieval predictors according to Pearson correlation.

	Combination	Robust	WT10G	GOV2	ClueWeb
AP	S_1: WIG + QF	.459+	.274*	.399+	.287*
	S_2: 2 Best SLF	.382+	.404+	.438+	.237*
	S_3: Best SLF	.402+	.339+	.420+	.302*
	S_4: $S_1 \cup S_2$	.478+	.420+	.465+	.260*
	$S_1 \cup S_3$: All	.454+	.427+	.509+	.208*
NDCG	S_1: WIG + QF	.537+	.303*	.405+	.286*
	S_2: 2 Best SLF	.430+	.449+	.469+	.211*
	S_3: Best SLF	.458+	.457+	.464+	.312*
	S_4: $S_1 \cup S_2$	.556+	.468+	.514+	.293*
	S_5: $S_1 \cup S_3$	.526+	.446+	.487+	.188

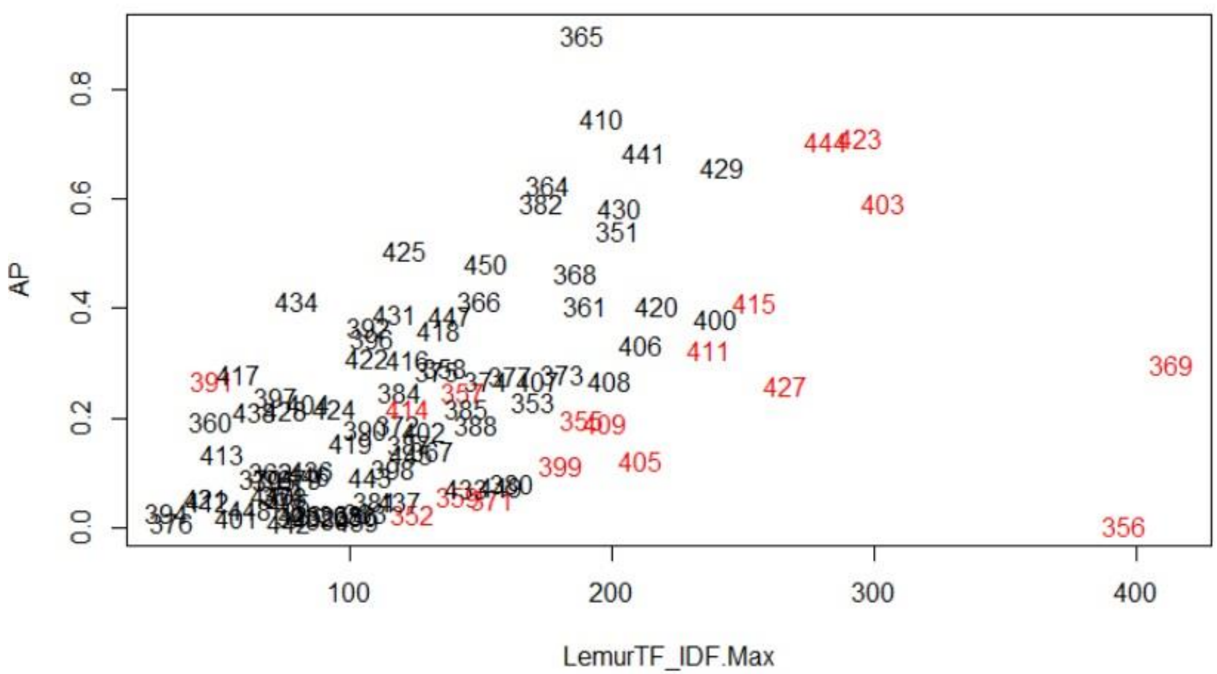
Some queries are difficult to predict

- Outliers (effectiveness prediction)



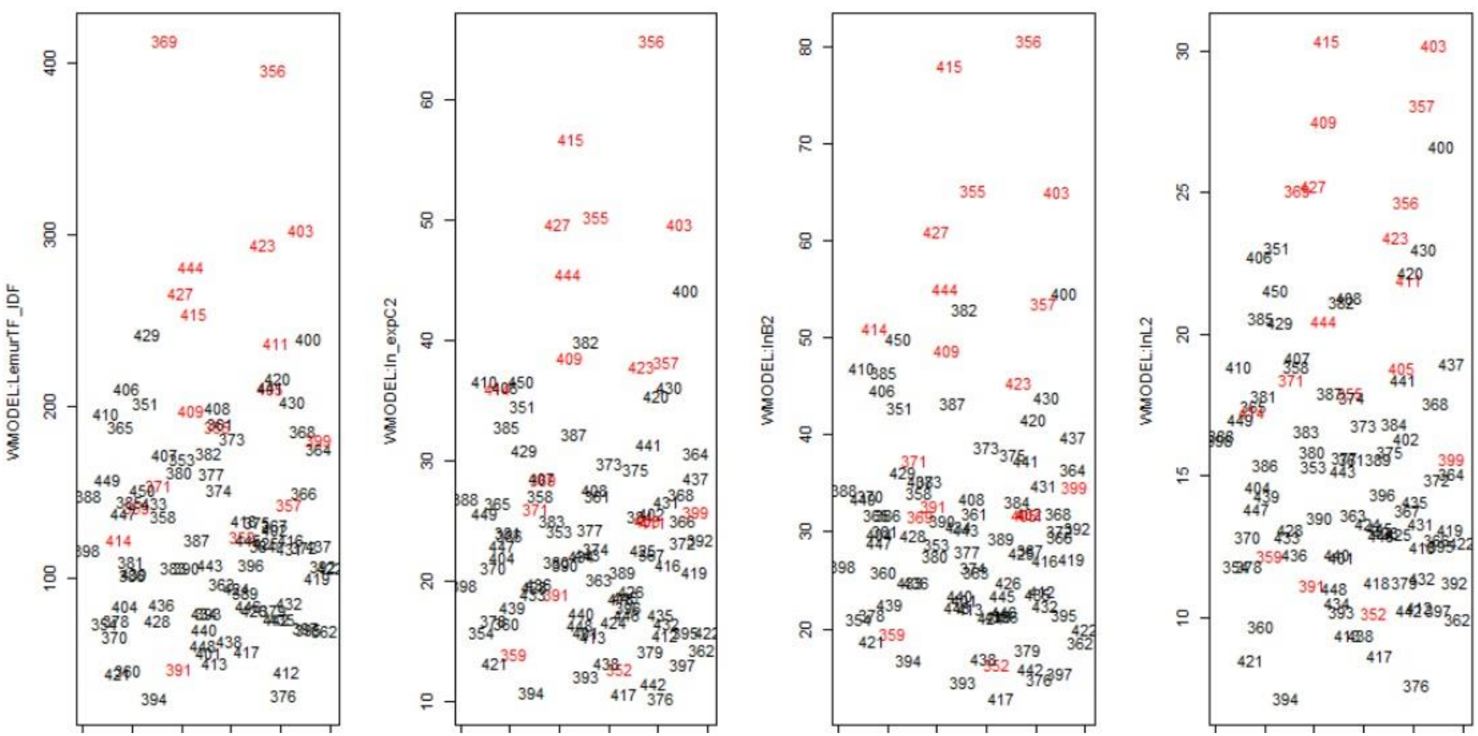
Some queries are difficult to predict

- Multi-variate outliers detection



Some queries are difficult to predict

- Multi-variate outliers detection



Some queries are difficult to predict

- Multi-variate outliers detection

Collection	WT10G	WT10G	TREC78	TREC78
Measure	NDCG	AP	NDCG	AP
Best system	0.444	0.236	0.524	0.238
Ouliers	16	16	19	18
LemurTF_IDF - Univariate	0.337	0.342	0.544	0.658
LemurTF_IDF - Outliers only	0.206	0.292	0.095	0.350
LemurTF_IDF - No Ouliers	**0.438**	**0.468**	**0.601**	**0.700**
LemurTF_IDF - All	0.365	0.393	0.381	0.522
In_expC2 - - Univariate	0.423	0.368	0.607	0.631
In_expC2 - No Outliers	0.391	0.350	**0.607**	**0.635**
In_expC2 - All	0.425	0.371	0.418	0.484
InB2 - Univariate	0.329	0.286	0.542	0.536
InB2 - No Outliers	0.286	0.214	0.530	**0.543**
InB2 - All	0.336	0.274	0.372	0.416
InL2 - Univariate	0.264	0.341	0.380	0.426
InL2 - No Outliers	0.258	0.347	**0.458**	**0.491**
InL2 - All	0.340	**0.353**	0.398	0.446

Collection	WT10G
Measure	NDCG
Ref system	0.4528
Outliers	24
NQC - No Outliers	**0.330**
NQC - All	0.097
UQC - No Ouliers	**0.350**
UQC - All	0.206
WIG - No Outliers	-0.015
WIG - All	0.077
QF - No Ouliers	**0.357**
QF - All	0.283

Main research directions

- Query difficulty prediction
- **Adaptive system / selective query processing**
- User studies

What are the most influential system parameters

- Descriptive analysis of results

Mining Information Retrieval Results: Significant IR parameters

J. Compaoré, S. Déjean, A.-M. Gueye, J. Mothe, J. Randriamparany

The First International Conference on Advances in Information Mining and Management - IMMM 2011

Studying the variability of system setting effectiveness by data analytics and visualization.
Déjean, S., Mothe, J., & Ullah, M. Z. (2019).
In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10* (pp. 62-74). Springer International Publishing.

What are the most influential system parameters

Parameters	Meaning	Values
Top	Topic number	351, …, 400
Field	Topic field	T, T+D, T+D+N
Bloc	Size of the indexing bloc	1, 5, 10
Idf	Inverse Document Frequency	FALSE, TRUE
Ref	Query reformulation	None, Bo1bfree, Bo2bfree, KLbfree
Model	Retrieval model	BB2c1, BM25b0.5, DFRBM25c1.0, IFB2c1.0, InexpB2c1.0, InexpC2c1.0, InL2c1.0, PL2c1.0, TFIDF
DocNb	Number of documents (reformulation)	0, 3, 5, 10, 50, 100, 200
qe_md	Minimum number of documents in which the term should appear to used in the query expension	0, 2
qe_t	Number of terms used in the query expension	0, 1

What are the most influential system parameters

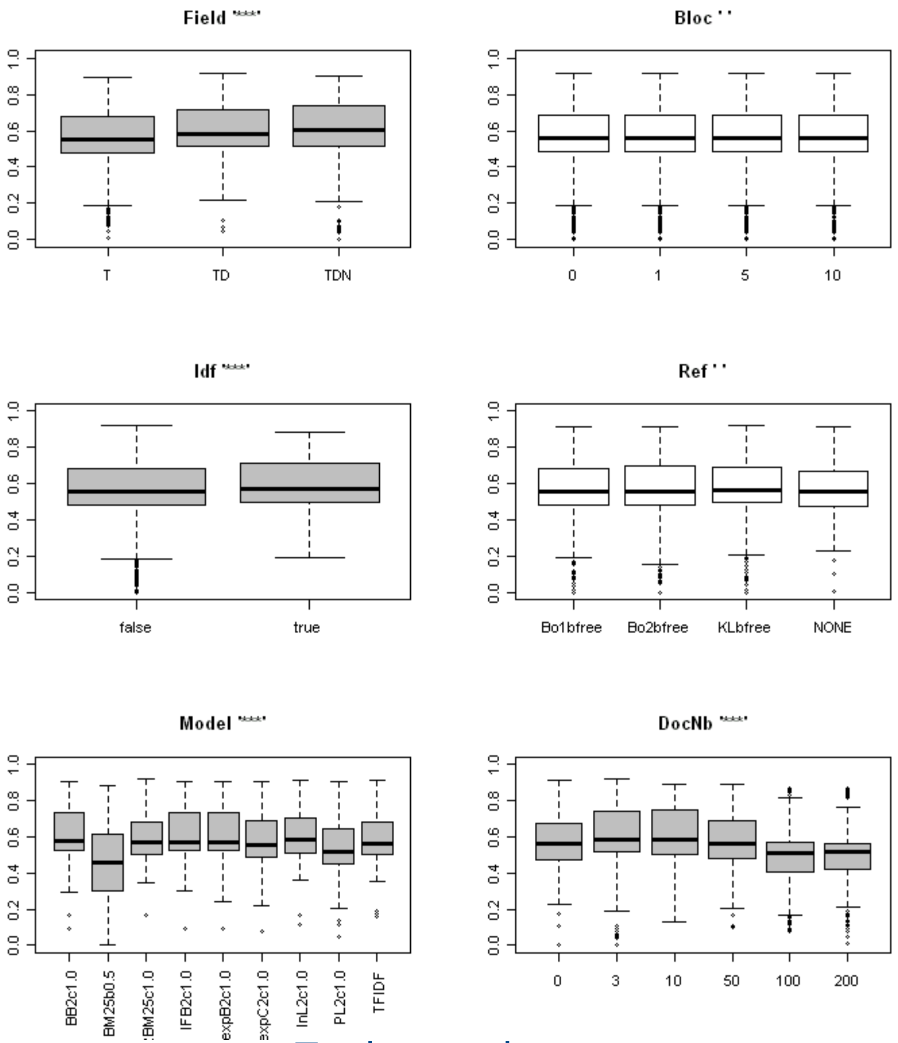
- Data

98650 rows: 1 row = one topic processed by a chain of modules
8 columns: 7 parameters + 1 performance measure (map)

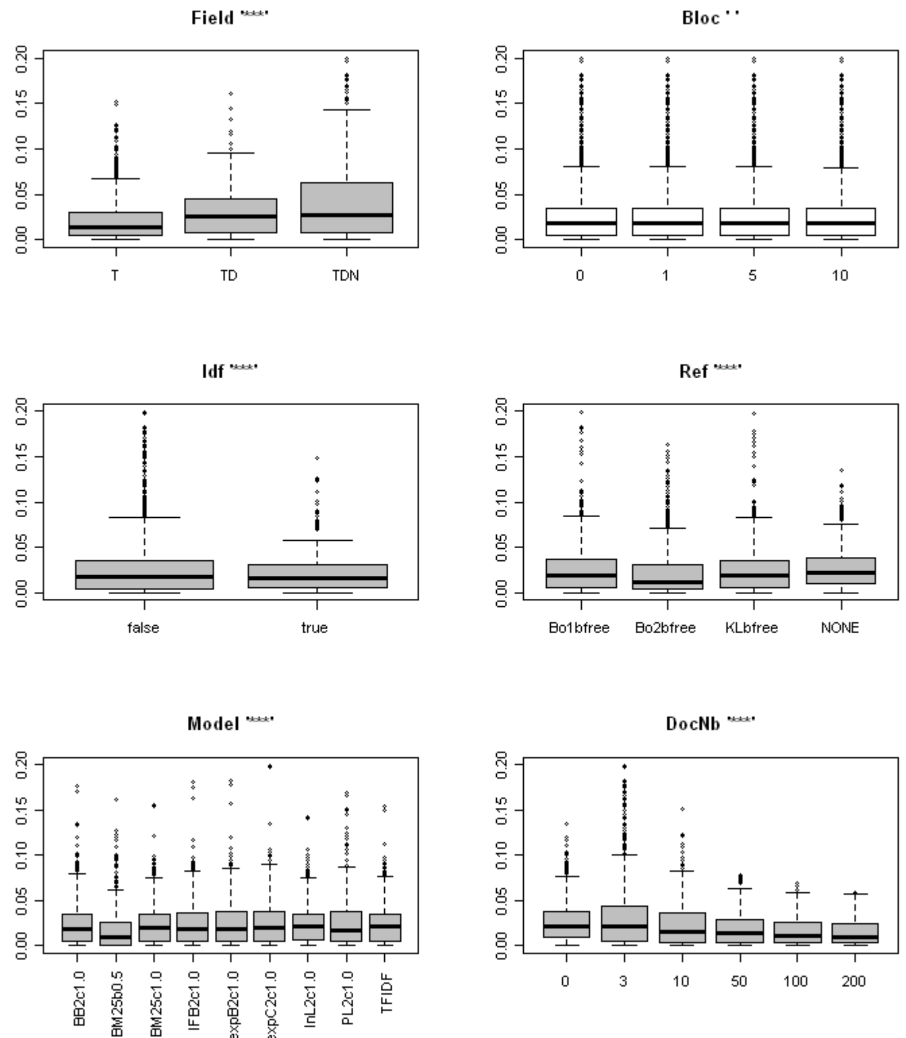
#	Top	Field	Bloc	Idf	Ref	Weight	DocNb	map
1	351	T	1	false	Bo1bfree	BB2c1.0	3	0.6134
2	352	T	1	false	Bo1bfree	BB2c1.0	3	0.3412
3	353	T	1	false	Bo1bfree	BB2c1.0	3	0.3479
4	354	T	1	false	Bo1bfree	BB2c1.0	3	0.0662
5	355	T	1	false	Bo1bfree	BB2c1.0	3	0.2794
6	356	T	1	false	Bo1bfree	BB2c1.0	3	0.0460
...								
98645	445	T	0	true	NONE	TFIDF	1	0.1514
98646	446	T	0	true	NONE	TFIDF	1	0.2234
98647	447	T	0	true	NONE	TFIDF	1	0.1121
98648	448	T	0	true	NONE	TFIDF	1	0.0114
98649	449	T	0	true	NONE	TFIDF	1	0.0714
98650	450	T	0	true	NONE	TFIDF	1	0.3226

What are the most influential system parameters

Significant effect (1-factor ANOVA)



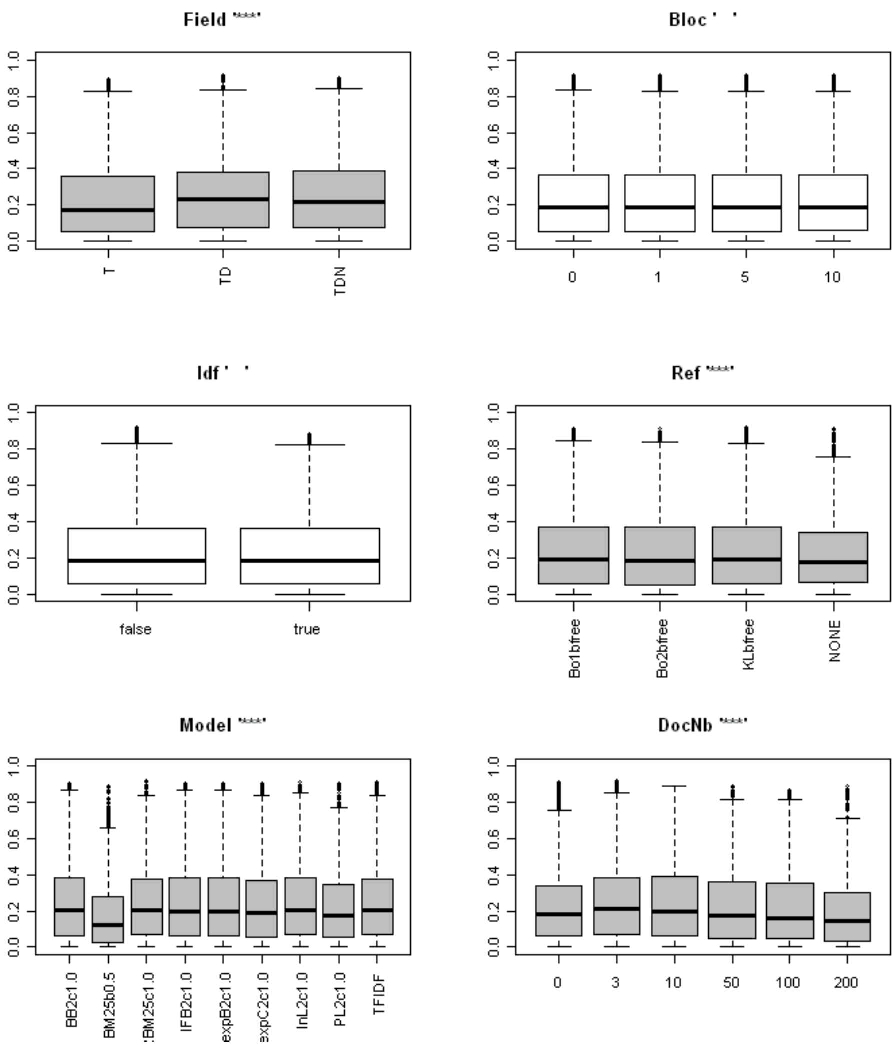
Easiest topics

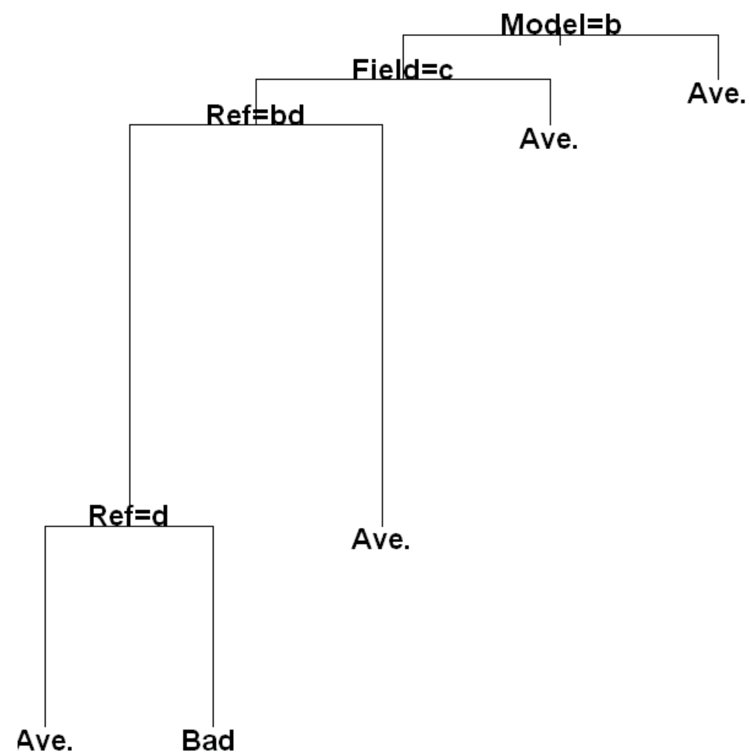


Hardest topics

CLEF, Déjean et al., 2019

What are the most influential system parameters

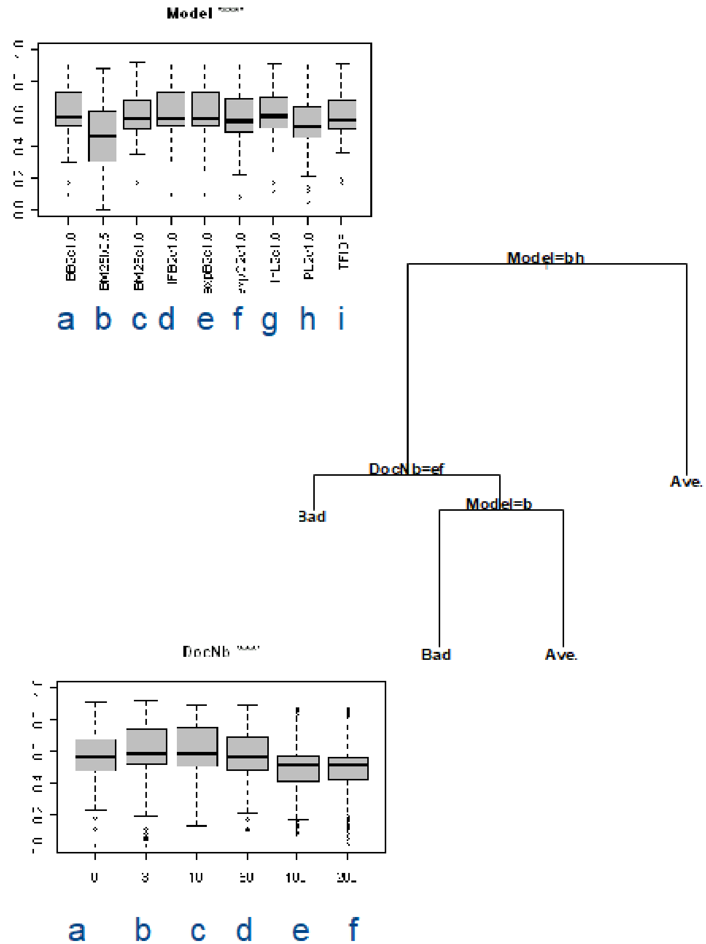




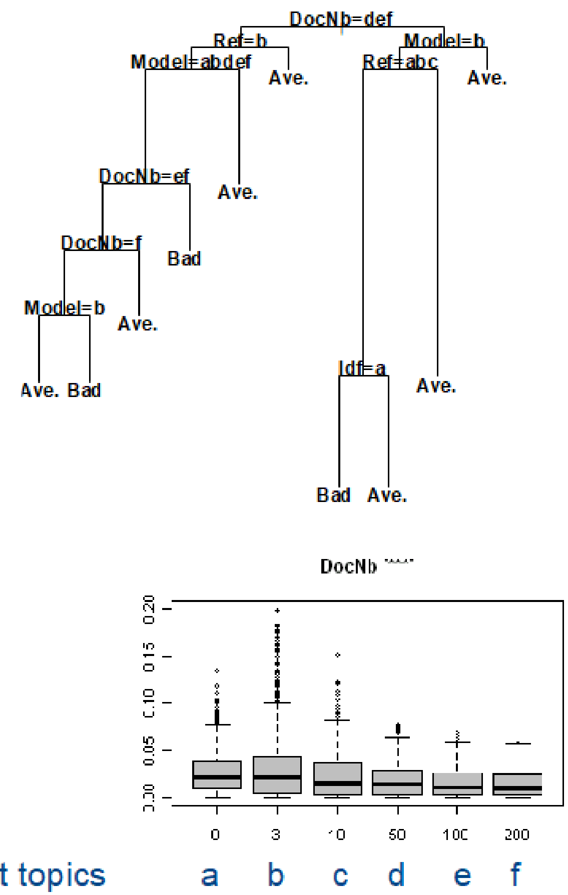
What are the most influential system parameters

➲ Multivariate analysis : CART

Classification **A**nd **R**egression **T**ree



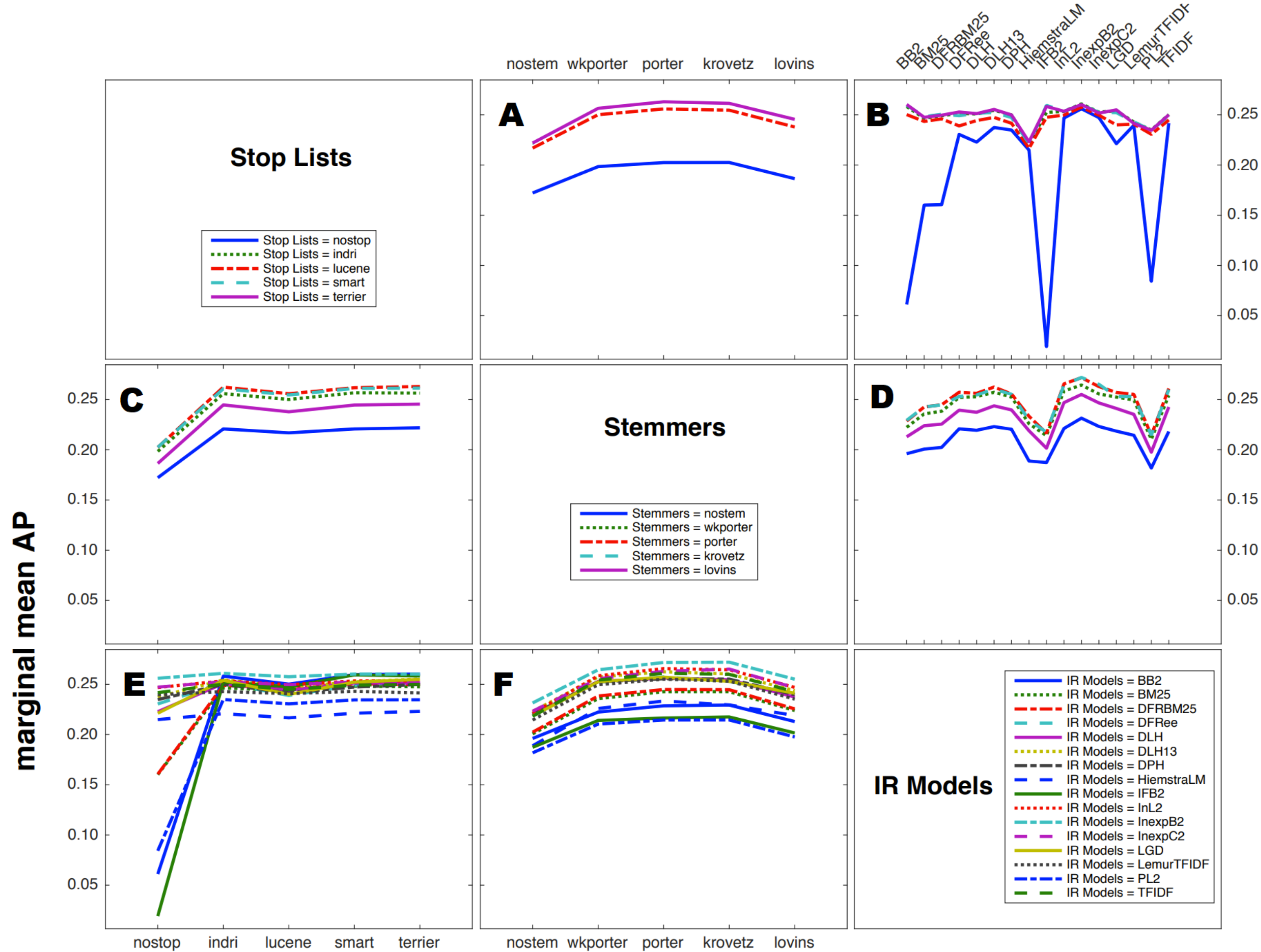
Easiest topics



Hardest topics

Ferro, SIGIR 2018

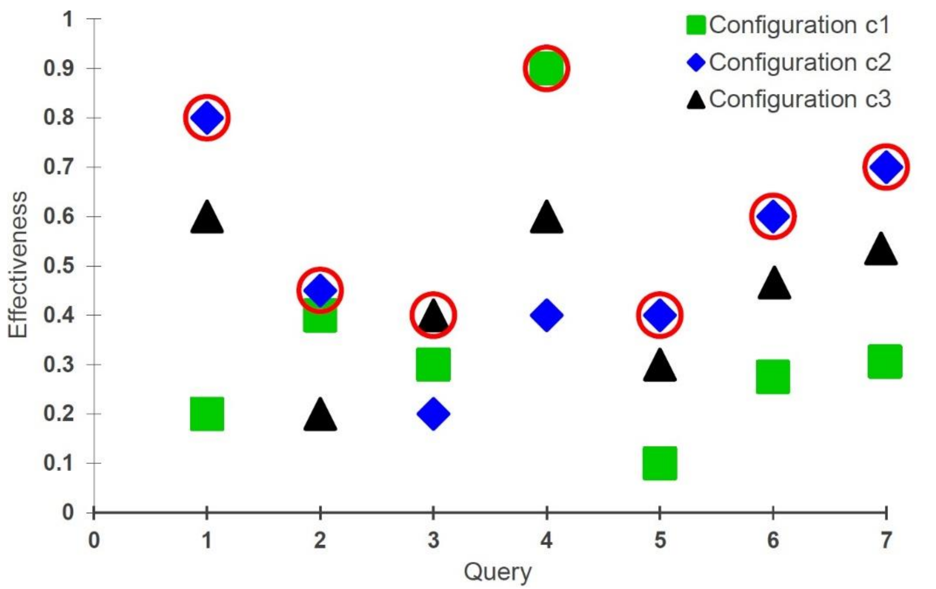
What are the most influential system parameters



Selective query processing strategies

- Observations on IR system
  - Systems perform differently on queries
  - One size does not fit all

	S_1	S_2	S_3	S_4
Q_1	$\boxed{0.8}$	$\boxed{0.9}$	0.2	0.1
Q_2	0.2	0.1	$\boxed{0.9}$	$\boxed{0.8}$



- Solution
  - Selective search strategy
    - **Different systems or system configurations are used for different queries**

Selective query processing strategies

- Early methods
  - Selective query expansion
    - [Cronen-Townsend et al., 2004] [Amati, 2003] [Yom-Tov et al., 2005]
      Decide whether a query should be expanded

    Effectiveness is limited to two configurations
    - [Xu et al., 2009]

    Different types of expansion according to queries

    The performance is still bounded by the three expansion strategies used
  - Model selection
    - [He and Ounis, 2004]

    Best matched query-cluster to select the search model

    The performance is also limited to those search models (8)

Selective search strategies

- More recently
  - Selective search model approach
    - [Arslan and Dincer, 2019]

    Used the frequency distribution of query terms to select the best search models

    Performance improved than SQE but limited to the search models

  - Selective search based on various configurations
    - [Mothe and Washha, 2017]

    Predicts the best value for a set of system parameters for a query – classifier-based approach

    Does not consider the dependency of the parameters

Which System to use to process a query?

- Parameter values make different system configurations
- Effectiveness differs according to configurations
- Can we learn the configuration to use?
- Learning to rank query-documents -> L2R query-configurations
- E-risk based function

Learning to Rank System Configurations
Romain Deveaud, Josiane Mothe, Jian-Yun Nie.
Conference on Information and Knowledge Management (CIKM), 2016.

Predicting the Best System Parameter Configuration: the (Per Parameter Learning) PPL method
Josiane Mothe, Mahdi Washha
International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES), Elsevier, 2017.

Defining an Optimal Configuration Set for Selective Search Strategy-A Risk-Sensitive Approach
Mothe, J., & Ullah, M. Z.
In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 1335-1345), 2021

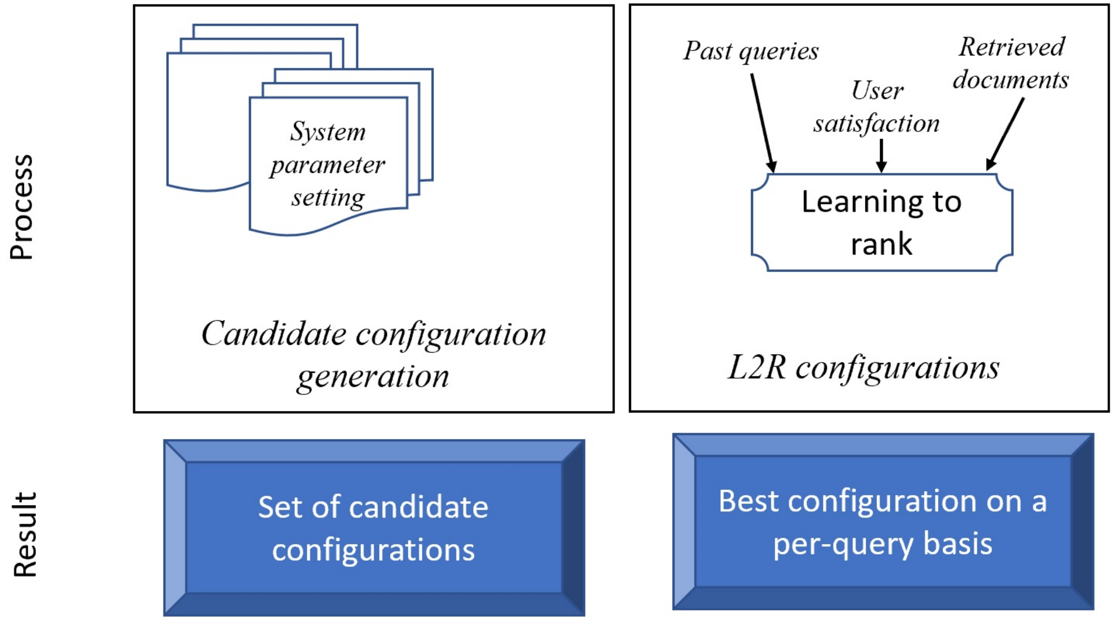
Selective search strategies

- [Deveaud et al, 2018]

Learning to rank system configurations

20 000 configurations : A specific setting of an ensemble of components and their hyper-parameters

e.g. BM25 with Bo2 query expansion using 5 documents and 10 query terms



Which System to use?

- System parameters

Table 1: Description of the system parameters that we use to build our dataset

Parameter	Description & values[2]
Retrieval model	21 different retrieval models: DirichletLM, JsKLs, BB2, PL2, DFRee, DFI0, XSqrAM, DLH13, HiemstraLM, InL2, DLH, DPH, IFB2, TFIDF, InB2, InexpB2 , DFRBM25 , BM25, LGD, LemurTFIDF, InexpC2.
Expansion model	7 query expansion models: nil, Rocchio, KL, Bo1, Bo2, KLCorrect, Information, KLComplete.
Expansion documents	Number of documents used for query expansion: 2, 5, 10, 20, 50, 100.
Expansion terms	Number of expansion terms: 2, 5, 10, 15, 20.
Expansion min-docs	Minimal number of documents an expansion term should appear in: 2, 5, 10, 20, 50.

Which System to use?

- Training examples
  - Query-configurations with effectiveness as label
  - Query: set of features (query difficulty predictors)
    - Linguistics based
    - Statistics based
- Machine learning methods
  - Train to know what is the best system configuration according to query features

Which System to use?

Table 2: Results with different L2R models and feature ablations. △ indicates statistically significant improvements over the Grid Search baseline, according to a paired t-test ($p < 0.05$). ▼ indicates statistically significant decreases induced by a feature ablation with respect to the corresponding (All) models.

	MAP		P@100		RPrec	
BM25	0.1942		0.1719		0.2330	
Grid Search	0.2480		0.2213		0.2835	
Random Forests (All)	0.3319 △		0.2785 △		0.3439 △	
- QueryStats	0.3180 △	(-4.17%)	0.2947 △	(+5.80%)	0.3658 △	(+6.35%)
- QueryLing	0.3367 △	(+1.43%)	0.2835 △	(+1.80%)	0.3507 △	(+1.96%)
- RetModel	0.3210 △	(-3.28%)	0.2746	(-1.44%)	0.3462 △	(+0.65%)
- Expansion	0.2201 ▼	(-33.68%)	0.1843 ▼	(-33.84%)	0.2384 ▼	(-30.69%)
SVM^{rank} (All)	0.3073 △		0.2529		0.3204	
- QueryStats	0.2820 △	(-8.23%)	0.2667 △	(+5.48%)	0.3304 △	(+3.12%)
- QueryLing	0.2918 △	(-5.03%)	0.2501	(-1.11%)	0.3498 △	(+9.19%)
- RetModel	0.3118 △	(+1.48%)	0.2628 △	(+3.91%)	0.3400 △	(+6.10%)
- Expansion	0.1723 ▼	(-43.92%)	0.1203 ▼	(-52.43%)	0.1914 ▼	(-40.28%)
GBRT (All)	0.3338 △		0.2803 △		0.3400 △	
- QueryStats	0.3375 △	(+1.11%)	0.2699	(-3.71%)	0.3275 △	(-3.71%)
- QueryLing	0.2982 △	(-10.68%)	0.2908	(+3.75%)	0.3288 △	(-3.31%)
- RetModel	0.3299 △	(-1.17%)	0.2702	(-3.62%)	0.3581 △	(+5.32%)
- Expansion	0.2345 ▼	(-29.75%)	0.1775 ▼	(-36.66%)	0.2505 ▼	(-26.32%)
LambdaMART (All)	0.3271 △		0.2772 △		0.2873	
- QueryStats	0.3272 △	(+0.03%)	0.2705 △	(-2.42%)	0.2692	(-6.28%)
- QueryLing	0.3324 △	(+1.62%)	0.2695 △	(-2.78%)	0.3486 △	(+21.34%)
- RetModel	0.3144 △	(-3.87%)	0.2713 △	(-2.13%)	0.3528 △	(+22.78%)
- Expansion	0.2188 ▼	(-33.11%)	0.1456 ▼	(-47.49%)	0.2078 ▼	(-27.67%)
Upper bound (oracle performance)	0.4136		0.3434		0.4490	

Which System to use?

- Learning to rank system configurations

Table 2: Results with different L2R models and feature ablations. △ indicates statistically significant improvements over the Grid Search baseline, according to a paired t-test ($p < 0.05$). ▼ indicates statistically significant decreases induced by a feature ablation with respect to the corresponding (All) models.

	MAP				RPrec	
BM25	0.1942				0.2330	
Grid Search	0.2480				0.2835	
					0.3439 △	
					0.3658 △	(+6.35%)
					0.3507 △	(+1.96%)
					0.3462 △	(+0.65%)
					0.2384▼	(-30.69%)
					0.3204	
					0.3304 △	(+3.12%)
					0.3498 △	(+9.19%)
					0.3400 △	(+6.10%)
					0.1914▼	(-40.28%)
GBRT (All)	0.3338 △		0.2803 △		0.3400 △	
- QueryStats	0.3375 △	(+1.11%)	0.2699	(-3.71%)	0.3275 △	(-3.71%)
- QueryLing	0.2982 △	(-10.68%)	0.2908	(+3.75%)	0.3288 △	(-3.31%)
- RetModel	0.3299 △	(-1.17%)	0.2702	(-3.62%)	0.3581 △	(+5.32%)
- Expansion	0.2345▼	(-29.75%)	0.1775▼	(-36.66%)	0.2505▼	(-26.32%)
LambdaMART (All)	0.3271 △		0.2772 △		0.2873	
- QueryStats	0.3272 △	(+0.03%)	0.2705 △	(-2.42%)	0.2692	(-6.28%)
- QueryLing	0.3324 △	(+1.62%)	0.2695 △	(-2.78%)	0.3486 △	(+21.34%)
- RetModel	0.3144 △	(-3.87%)	0.2713 △	(-2.13%)	0.3528 △	(+22.78%)
- Expansion	0.2188▼	(-33.11%)	0.1456▼	(-47.49%)	0.2078▼	(-27.67%)
Upper bound (oracle performance)	0.4136		0.3434		0.4490	

	MAP	
GBRT (All)	0.3338 △	
- QueryStats	0.3375 △	(+1.11%)
- QueryLing	0.2982 △	(-10.68%)
- RetModel	0.3299 △	(-1.17%)
- Expansion	0.2345▼	(-29.75%)

Which System to use?

BUT

- Selective search strategy (SSS)
  - Number of possible configurations is very large
  - Too many configurations are difficult to maintain
  - Some configurations are good for a few queries
  - Some configurations could be risky for important queries

Objective

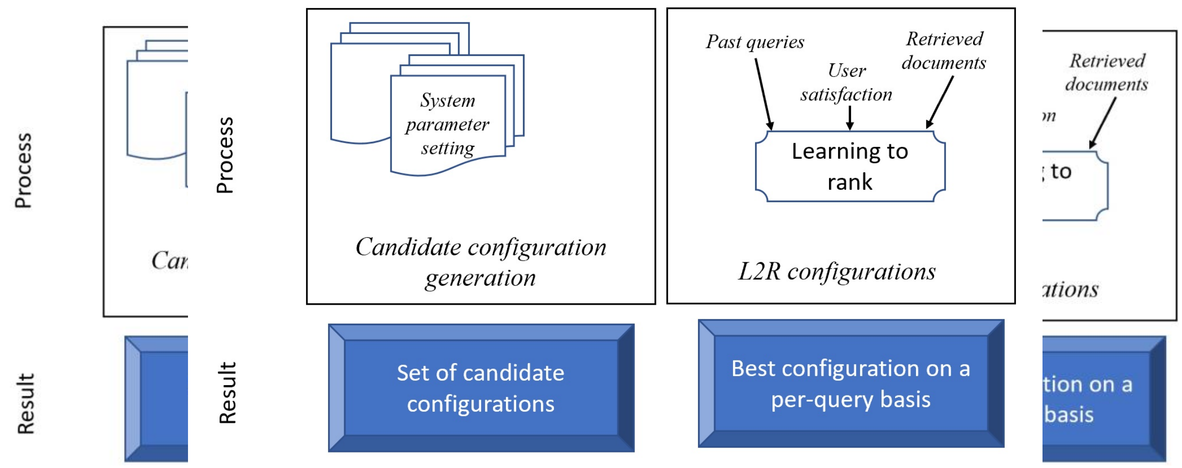
- Select a representative set of system configurations

Solution

- Random selection – poor [Deveaud et al, 2018]
- Advanced selection

Which System to use?

Selection of a limited number of configurations



Selection of a limited number of configurations
from the initial pool
for the selective search strategy

Which System to use?

Selection of a limited number of configurations

- Greedy approach
- Iteratively selects one representative configuration at a time

- We used risk and reward functions to select the reduced set of candidate configurations

Direct

Which System to use?

Risk-sensitive criteria

- **Risk-averse ranking** algorithm considering mean-variance analysis of a ranked list [Wang and Zhu, 09]
- **Risk-reward trade-off function Urisk** based on Frisk to optimize learning to rank model [Wang et al, 2012]
- **Student's T-distribution-wise** risk-reward trade-off function **Trisk** [Dincer et al., 2014]
- **Zrisk** and **Grisk** to compare the risk-reward trade-off of a system against multiple baselines [Dincer et al., 2016]
- Risk-reward trade-off in rank fusion [Benham et al, 2017]
- **Risk functions for feature selection in learning to rank documents [De Sousa et al., 2016]**

Risk sensitive criteria to select candidate configurations

- Definition of System Risk
  - The risk of performing a given particular query less effectively than a given baseline system

- F_{RISK} is defined as follows:

$$F_{Risk}(Q_T, M) = \frac{1}{|T|} \sum_{q_i \in Q_T} \max(0, B(q_i) - M(q_i))$$

- Q_T is the training query set
- $B(q_i)$ is the baseline effectiveness for query q_i, and
- $M(q_i)$ is the effectiveness of the model for which the risk is estimated

Model design

- Best Query-Configuration Fit
  - For each query, select the most appropriate configuration
  - Cast as a problem of ranking the candidate configurations

Set of configurations

(a, b, c)
Q1 $(1_a, 1_b, 1_c)$
Q2 $(2_a, 2_b, 2_c)$
Q3 $(3_a, 3_b, 3_c)$

(α, δ, β)
S1 $(\alpha_1, \delta_1, \beta_1)$
S2 $(\alpha_2, \delta_2, \beta_2)$
S3 $(\alpha_3, \delta_3, \beta_3)$
S4 $(\alpha_4, \delta_4, \beta_4)$

Learning to rank
Q1: **S2**,S1,S3,S4
Q2: **S4**,S2,S1,S3
Q3: **S4**,S1,S2,S3

Query features

Configuration features

Model design

- Query Features (the (a,b,c))
  - Summarized LETOR features based on BM25
    - 38 LETOR features [Adrian et al., 2018]
    - Aggregated functions
      - Mean, Standard deviation, and Maximum
- Configuration features (the (α,δ, β))
  - 21 retrieval models
  - 7 expansion models
  - 6 variants of number of expansion documents
  - 5 variants of number of expansion terms
  - 5 variants of minimum number of expansion documents

Model design

- Training based on
  - Query-System configuration pairs + label (effectiveness)
  - Learning-to-rank algorithms for point-wise, pairwise, and listwise approaches

- RankLib library
  - Random Forest, GBRT, and LambdaMART
- SVM-rank library
  - SVMrank
- Scikit-learn
  - Linear regression

Experiments and evaluation

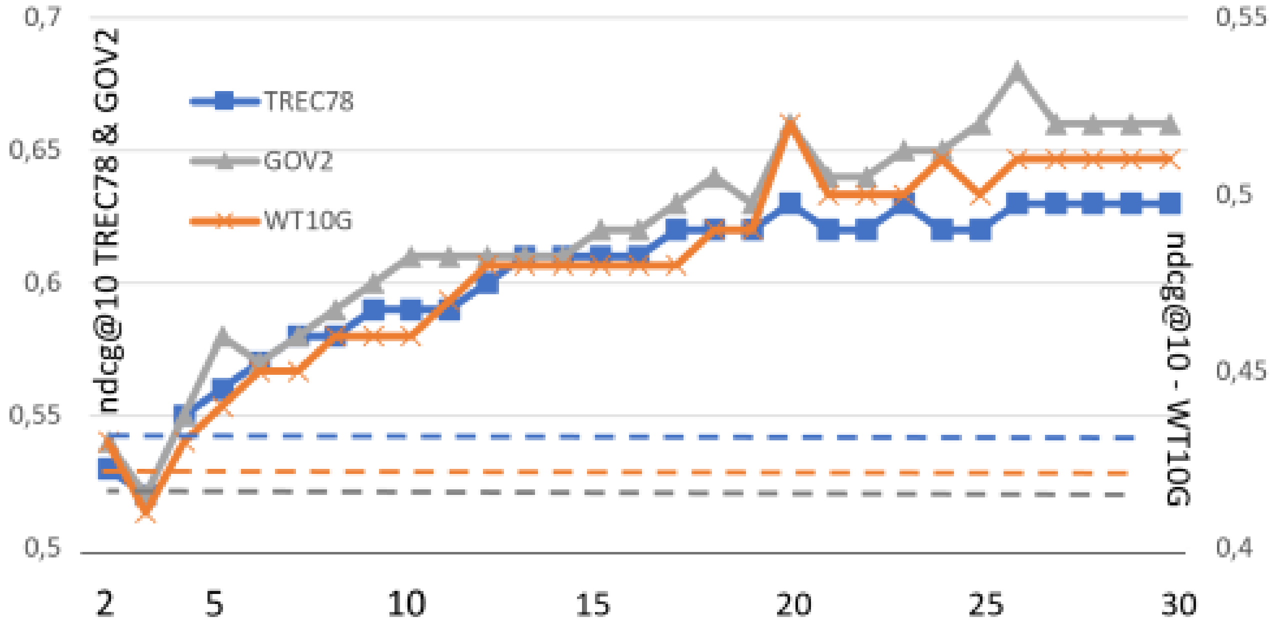
- Test collections
  - TREC78 -- 100 Topics (351 – 450)
  - WT10G -- 100 Topics (451 – 550)
  - GOV2 -- 150 Topics (701 – 850)
- Metrics
  - AP, nDCG@10, and P@10
- Evaluation
  - Two-fold cross-validation for three trials (Q_A and Q_{A^-})
- Significance testing
  - Two-tailed paired t-test with Bonferroni correction
    - P-value < 0.05

Experiments and evaluation

- Baselines
  - Single Configuration
    - BM25
    - L2R-D SVM-rank
    - Grid Search
    - Best trained
  - Selective Search Strategy
    - Trained SQE
    - Deveaud et al. [2018]
- Oracles
  - Best Conf.
  - Oracle20SS
  - Oracle

Results

- Impact of k on effectiveness and cost:



Performance for E_{RISK} function on the three collections while varying the number of candidate configurations. The dotted dash horizontal lines are the single best configuration

Results

- Effectiveness on TREC78 with 20 candidate configurations by E_{RISK} function

	Methods	TREC78 MAP	TREC78 nDCG@10	TREC78 P@10
Baselines	BM25	.21	.47	.43
	L2R-D SVM^r	.22 [.000]	.48 [.001]	.46 [.004]
	GS	.24 [.003]	.51 [.019]	.47 [.003]
	Best trained	.25 [.010]	.52 [.008]	.47 [.009]
SelSS	Trained SQE	.24 [.002]	.53 [.007]	.49 [.006]
	Deveaud *et al.* [16]	.24 [.002]	.56 [.003]	.52 [.004]
	ERisk-RF	**.28**$^{\Delta\uparrow}$ [.007]	**.63**$^{\Delta\uparrow}$ [.005]	**.60**$^{\Delta\uparrow}$ [.012]
	Best conf.	.26	.54	.51
	Oracle	.39	.83	.80
	Oracle20SS	.29	.63	.61

Results

- Effectiveness on GOV2 with 20 candidate configurations by E_{RISK} function

	Methods	MAP	nDCG@10	P@10
		GOV2		
Baselines	BM25	.27	.46	.54
	L2R-D SVM^r	.28 [.001]	.49 [.002]	.57 [.003]
	GS	.35 [.005]	.52 [.003]	.62 [.008]
	Best trained	.35 [.005]	.49 [.012]	.59 [.010]
SelSS	Trained SQE	.35 [.009]	.52 [.002]	.63 [.005]
	Deveaud *et al.* [16]	.40 [.003]	.66 [.001]	.77 [.005]
	ERisk-RF	**.41**$^{\triangle}$ [.002]	**.67**$^{\triangle}$ [.002]	**.79**$^{\triangle}$ [.010]
	Best conf.	.36	.52	.63
	Oracle	.50	.85	.94
	Oracle20SS	.42	.68	.80

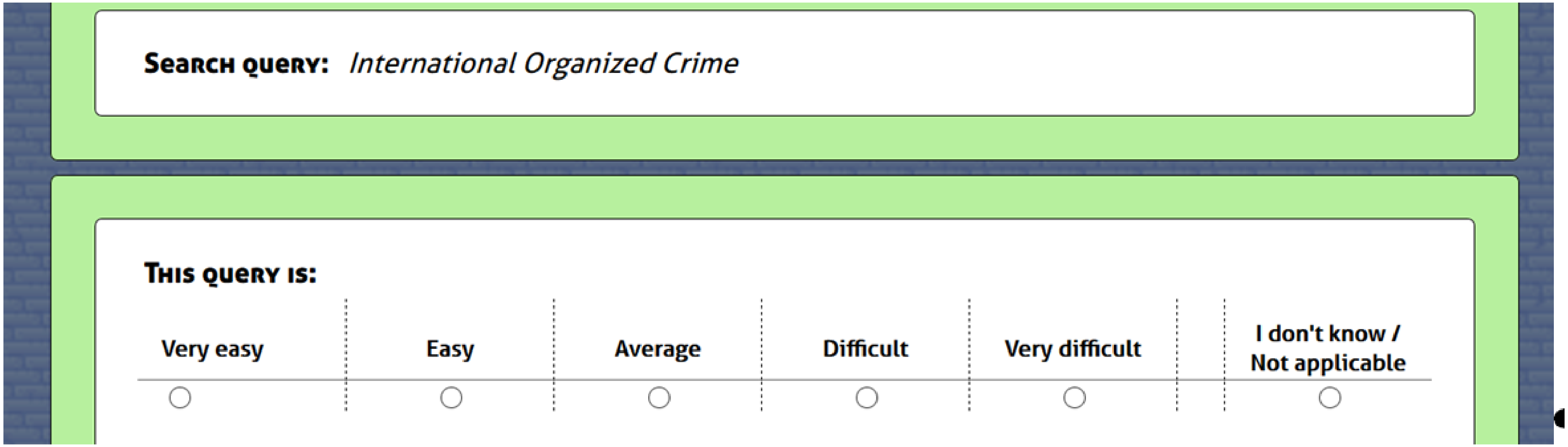
Main research directions

- Query difficulty prediction
- Adaptive systems
- **User studies**

Human-Based Query Difficulty Prediction: Is There Any Hope?

- Can we learn something from human?
- From the crowd ? From labs?

mbq.irit.fr



Human studies

- TREC 7 & 8 (old data)
  - Crowd: No correlation
  - Lab (students in libraries): No correlation
  - While little correlation with IDF (0.5) and STD (0.6)

#	Participants	Scale	Collection	# of topics	Metrics	Amount of info	Explanations	Topics
E1	Crowd (IN + US) 120 (60 + 60)	3	TREC 6-8	30	AP	Q, Q+D	Free text	310 311 312 313 314 315 316 351 352 353 354 355 356 357 358 360 403 404 406 414 420 421 422 424 426 427 428 430 433 434
E2	Lab 38 (29 + 9)	3	TREC 6-8	91 (*)	AP	Q, Q+D	Free text (**)	321-350 in TREC 6, 351-381 in TREC 7, 421-450 in TREC 8 (*)
E3	Crowd (IN, US) 100 (50 + 50)	5	TREC 2014	25	ERR@20 NDCG@20	Q, Q+D	Free text	251 255 259 261 267 269 270 273 274 276 277 278 282 284 285 286 287 289 291 292 293 296 297 298 300
E4	Lab 22	5	TREC 2014	25	ERR@20 NDCG@20	Q, Q+D	Categories (**) + Free text	Same as E3

Human studies

- TREC 2012 (web data)
  - Crowd: Little correlation (0.4)
  - Lab (IRIT + others): no correlation
  - While no correlation with IDF and little with STD (0.4)

#	Participants	Scale	Collection	# of topics	Metrics	Amount of info	Explanations	Topics
E1	Crowd (IN + US) 120 (60 + 60)	3	TREC 6-8	30	AP	Q, Q+D	Free text	310 311 312 313 314 315 316 351 352 353 354 355 356 357 358 360 403 404 406 414 420 421 422 424 426 427 428 430 433 434
E2	Lab 38 (29 + 9)	3	TREC 6-8	91 (*)	AP	Q, Q+D	Free text (**)	321-350 in TREC 6, 351-381 in TREC 7, 421-450 in TREC 8 (*)
E3	Crowd (IN, US) 100 (50 + 50)	5	TREC 2014	25	ERR@20 NDCG@20	Q, Q+D	Free text	251 255 259 261 267 269 270 273 274 276 277 278 282 284 285 286 287 289 291 292 293 296 297 298 300
E4	Lab 22	5	TREC 2014	25	ERR@20 NDCG@20	Q, Q+D	Categories (**) + Free text	Same as E3

Why do you think a query is easy/difficult?

- Can human predict difficulty?
  - No [Hauff et al., 2010] [Mizzaro & Mothe, 2016]
- Difficulty Reasons:
  - Why is a query difficult?
  - Can human identify the reasons?
  - Do reasons correlate to automatic predictors?
- Amount of information:
  - Do description change the difficulty prediction? (compared to the query only)
- Links with actual system difficulty

Why do you think a query is easy/difficult?

Why do you Think this Query is Difficult? A User Study on Human Query Prediction
Stefano Mizzaro, Josiane Mothe.
ACM SIGIR, 2016.

Human-Based Query Difficulty Prediction
Adrian-Gabriel Chifu, Sébastien Déjean, Stefano Mizzaro, Josiane Mothe
European Colloquium on Information Retrieval (ECIR), 2017.

Why do you think a query is easy/difficult?

- Aim: ***what are the reasons?***
- Participants: 39 MS (library and teaching studies)
- Choose among 150 topics (TREC adhoc)
- Evaluate difficulty (3 levels scale)
  + free text explanation

```
easy because:
difficult because:
```

- First using T, then using T+D

Annotation analysis

- Recoding free text

Comment	Recoding
A single word in the query	One-Word
The term exploration is polysemous	Polysemous-Word
Far too vague topic	Too-Vague-Topic
Is it in US? Elsewhere?	Missing-Where
Few searches on this topic	Unusual-Topic
Risk of getting too many results	Too-Many-Documents
There are many documents on this	Many-Documents

Table 2. Most frequent: (a) words in free text comments; (b) comments after recoding.

(a)

Easy because		Difficult because	
Precise	113	Missing	64
Clear	48	Broad	62
Many	45	Risk	56
Polysemous	36	Context	34
Usual	16	Polysemous	33
Specialist	15	Vague	26
Simple	11	Many	21

(b)

Easy because		Difficult because	
Precise-Topic	66	Risk-Of-Noise	50
Many-Documents	45	Broad-Topic	43
No-Polysemous-Word	31	Missing-Context	34
Precise-Words	25	Polysemous-Words	22
Clear-Query	19	Several-Aspects	20
Usual-Topic	16	Missing-Where	16

Why do you think a query is easy/difficult?

- Master students in libriary studies

Is this query easy?
Why ?

easy: clear query without ambiguity since there is no alternative synonyms

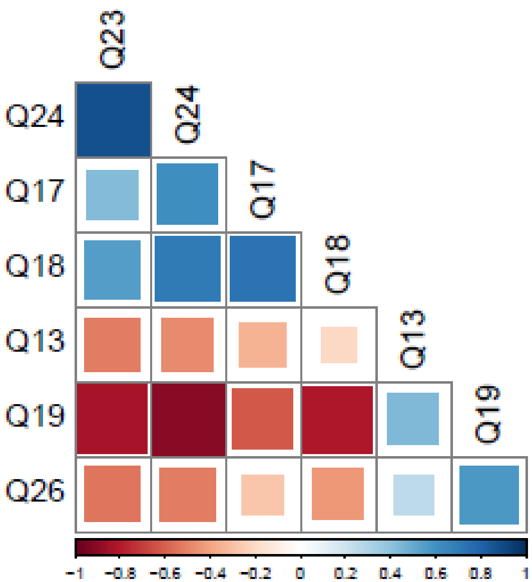
R4: The query contains generic word(s)
R10: The topic is Unusual/uncommon/unknown
R11: The topic is too broad/general/large/vague
R12: The topic is specialized
R26: The number of query words is too high
R16 The topic is too precise/specific/focused/delimited/clear
R23: Many of the relevant documents will be retrieved
R27 The query is concrete/explicit

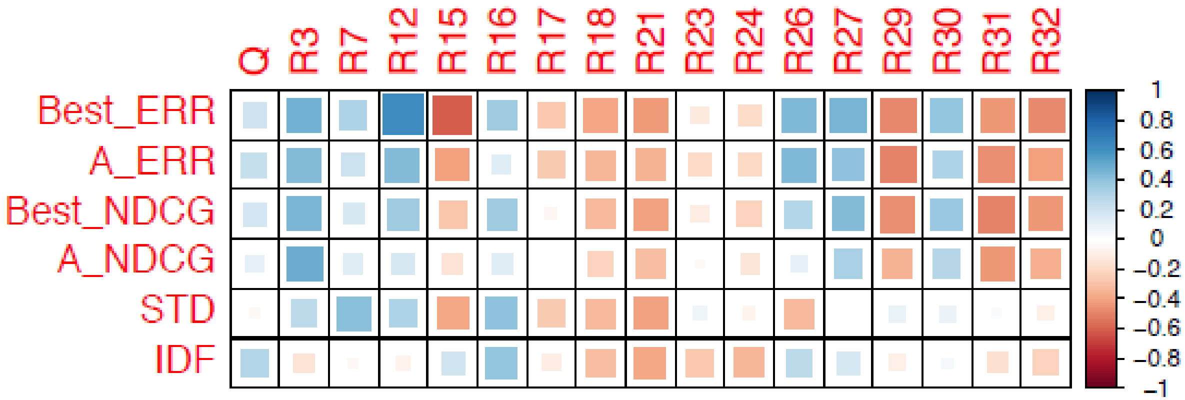
Figure 3: Examples of reasons resulting from the recoding of free text annotation on query difficulty comments.

CloseD-questions as reasons

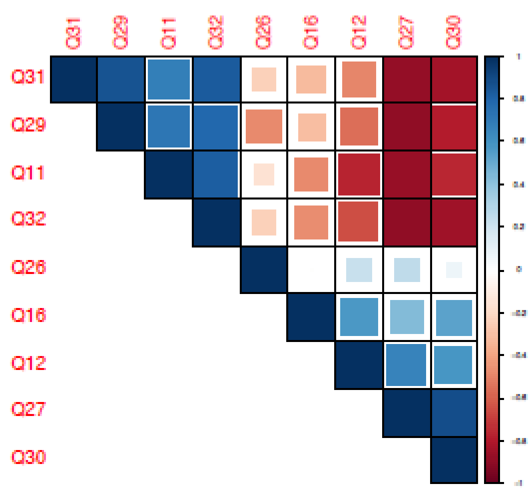
- Reasons as 32 closed-questions (ClueWeb12)
- 25 topics (10 hard, 10 easy, 5 avg), 22 part.
- 8 annotations per topics (5-levels scale for difficulty + Questions)

Question
Q1: The query contains vague word(s)
Q3: The query contains word(s) relevant to the topic/query
Q10: The topic is unusual/uncommon/unknown
Q13: The topic has several/many aspects
Q17: The topic is usual/common/known
Q18: The number of documents on the topic in the web is high
Q19: None or very few relevant documents will be retrieved
Q20: Only relevant documents will be retrieved
Q23: Many of the relevant documents will be retrieved
Q24: Many relevant documents will be retrieved
Q26: The number of query words is too high
Q28: The query contains various aspects
Q30: The query is clear




Q
R3
R7
R12
R15
R16
R17
R18
R21
R23
R24
R26
R27
R29
R30
R31
R32
Best_ERR
A_ERR
Best_NDCG
A_NDCG
STD
IDF
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1


Q
R3
R7
R12
R15
R16
R17
R18
R21
R23
R24
R26
R27
R29
R30
R31
Best_ERR


Q31
Q29
Q11
Q32
Q26
Q16
Q12
Q27
Q30

R12: The topic is specialized
R26: The number of query words is too high
R16 The topic is too precise/specific/focused/delimited/clear
R27 The query is concrete/explicit

Table 4: Pearson's correlations between actual system effectiveness, automatic predictors and reasons. Bold indicates a p-value < 0.05, * <0.005.

	Best ERR	TREC AERR	Best NDCG	TREC ANDCG	STD	IDF
STD	0.335	0.171	**0.438**	**0.450**	**1***	0.087
IDF	0.209	0.133	0.296	0.178	0.087	**1***
R12	**0.622***	**0.436**	0.359	0.180	0.302	-0.066
R16	0.349	0.140	0.345	0.137	**0.393**	**0.390**
R26	**0.445**	**0.447**	0.295	0.101	-0.321	0.261
R27	**0.460**	**0.409**	**0.434**	0.323	-0.005	0.171

Close questions analysis

- Correlation with human « prediction »

	Reason	Correlation Q	Correlation Q+D
None	R2: The query contains polysemous/ambiguous word(s)	0.342	0.145
	R8: The words in the query are inter-related or complementary	-0.028	0.187
	R12: The topic is specialized	-0.103	-0.136
	R10: The topic is Unusual/uncommon/unknown	0.526	0.496
Some	R13: The topic has several/many aspects	0.614	0.708
	R19: None or very few relevant document will be retrieved	0.880	0.800
	R30: The query is clear	-0.532	-0.631

Some reasons clearly correlate with the perception of difficulty.
S/he predicts the query difficult when:

- **The topic has several aspects**
- **S/he has a idea on the number of retrieved documents**
- **The query is not clear**

Close questions analysis

- Link *system* query features and *human* reasons

R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25	R26	R27	R28	R29	R30	R31	R32	
18	3	11	23	-6	-2	-3	7	5	10	-14	-7	6	-23	-15	32	1	7	11	7	-39	11	3	1	-8	26	17	11	-9	4	-16	2	avg_idf
20	14	-13	-1	17	12	-15	-31	-2	9	29	0	-19	3	-56	-38	15	-11	18	11	-13	21	16	16	14	20	-21	5	9	-14	23	-22	hypernyms
20	14	-13	-1	17	12	-15	-31	-2	9	29	0	-19	3	-56	-38	15	-11	18	11	-13	21	16	16	14	20	-21	5	9	-14	23	-22	meronyms
-23	-22	27	-45	8	11	42	12	-39	-16	-9	30	6	-33	-15	28	-40	-10	-6	-13	-40	-27	-6	-8	-35	-32	-1	2	9	9	2	0	STD
32	33	5	6	-21	-21	-31	29	20	14	11	-18	20	-28	13	23	15	21	0	7	36	8	-5	-6	55	15	-20	-4	9	-24	17	25	hyponyms
45	43	-7	8	-54	-54	-15	18	13	-9	20	-29	-12	-22	8	10	-5	18	-24	6	8	-13	-31	-32	28	9	-22	-42	22	-34	18	12	sister.terms
39	41	-14	-3	-62	-70	-2	18	-3	-29	24	-21	-27	-15	12	6	-22	7	-34	17	14	-30	-36	-36	16	-5	-32	-62	34	-37	35	17	synonyms
19	36	9	-27	-42	-47	-1	-8	-27	-6	41	-25	-41	-17	8	-10	-16	-4	-76	-52	1	-23	-76	-75	-5	-40	-53	-51	62	-41	44	-9	holonyms

Some reasons clearly correlate with query features

- **The number of holonyms seems related to the predicted number of retrieved documents [*many document when many parts*]**
- **The variety of aspects (R28) and synonyms [*topic ambiguity*]**
- **Specialization (R6) and synonyms [*few senses when specialized*]**

Close questions analysis

- Links between reasons and percieved difficulty/actual difficulty

Question	Correl.	
Q1: The query contains vague word(s)	.52	-.30
Q3: The query contains word(s) relevant to the topic/query	-.41	.43
Q10: The topic is unusual/uncommon/unknown	.52	.26
Q13: The topic has several/many aspects	.61*	-.07
Q17: The topic is usual/common/known	.62*	-.25
Q18: The number of documents on the topic in the web is high	-.69*	-.34
Q19: None or very few relevant documents will be retrieved	.88*	.32
Q20: Only relevant documents will be retrieved	-.47	.09
Q23: Many of the relevant documents will be retrieved	-.86*	-.20
Q24: Many relevant documents will be retrieved	-.87*	-.21
Q26: The number of query words is too high	.62*	.45
Q28: The query contains various aspects	.46	-.12
Q30: The query is clear	-.53	.30

While some reasons clearly correlate with human perception of difficulty, they are poor indicator of actual difficulty.

Conclusion

- Human can not predict query difficulty
- Reasons of difficulty make sense to them

No need to ask them

Use this when :

Designing system

Training users

- Enlarge the panel
- Various level of system/domain knowledge
- Compute features on *human* reasons

Future work

General conclusion

- Query difficulty prediction
  - Still not solved
  - Too many factors, including users
  - Evaluation is better with performance prediction than correlation with effectiveness
- Adaptive systems
  - Face real application constraints
- User studies
  - Many hope to find cross effects

General conclusion

- Descriptive analysis
  - Help understanding
  - Help discovering unknown trends
  - Calculations and visualisations are complementary
  - Methods should be used when appropriate
- Machine Learning
  - Extract models to predict
  - Evaluation is crutial

More at

www.irit.fr/~Josiane.Mothe

Josiane.mothe@irit.fr