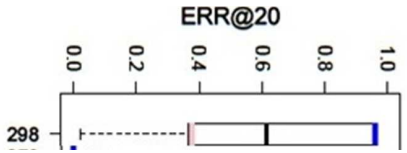
DIFFICULT QUERIES: DATA ANALYTICS MACHINE LEARNING AND USERS STUDIES

Josiane.Mothe@irit.fr

UNIVERSITÉ TOULOUSE Jean Jaurès
IRIT
CNRS - INPT - UPS - UT1 - UT2J

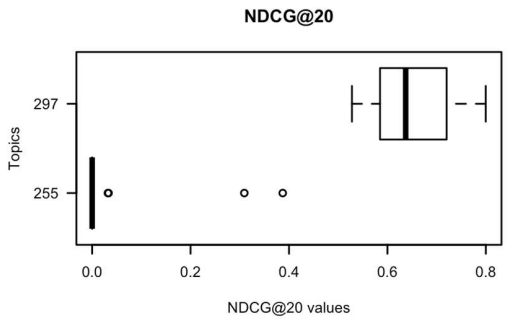
QUERY DIFFICULTY

- Geants have an answer whatever the query is BUT
- Evaluation compaigns showed
  - System variety



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

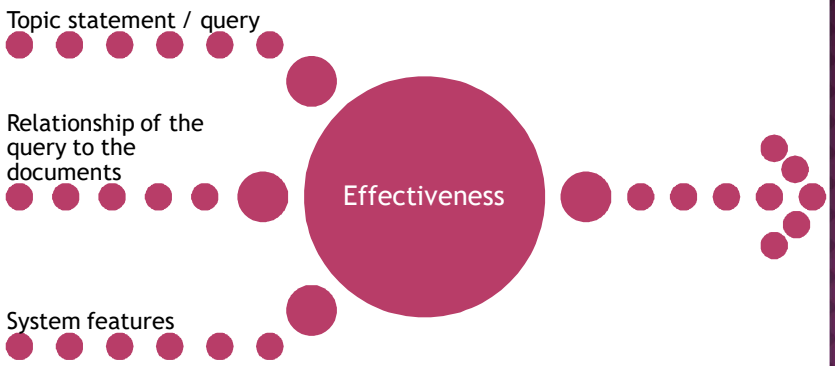
  Difficult topic = Poor effectiveness

- (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
  - Different systems (parameters) for different queries
- User studies
  - Measure users' abilities with regard to query difficulty

QUERY DIFFICULTY PREDICTORS

- 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 PREDICTORS

- 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 PREDICTORS

- 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]

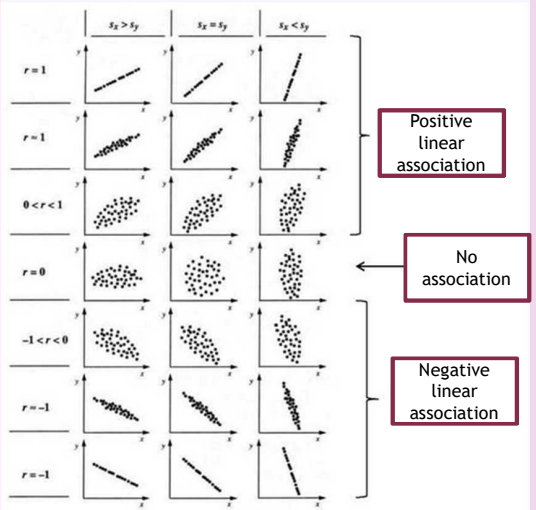
EVALUATION OF QUERY DIFFICULTY PREDICTORS

- 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

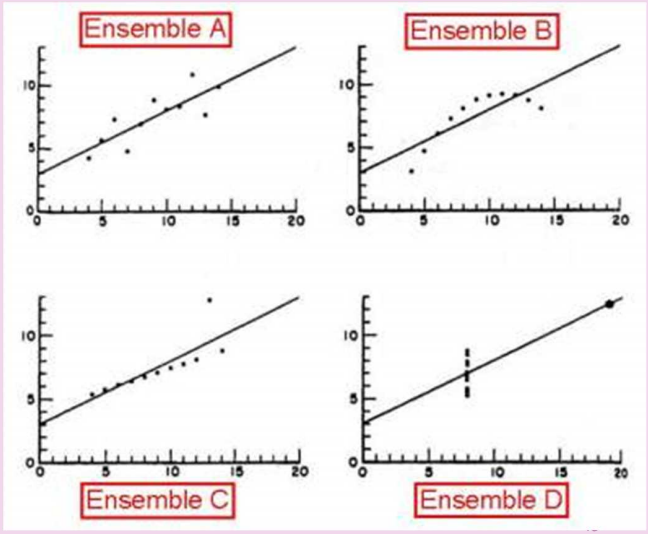
LINEAR CORRELATION BRAVAIS-PEARSON

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

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

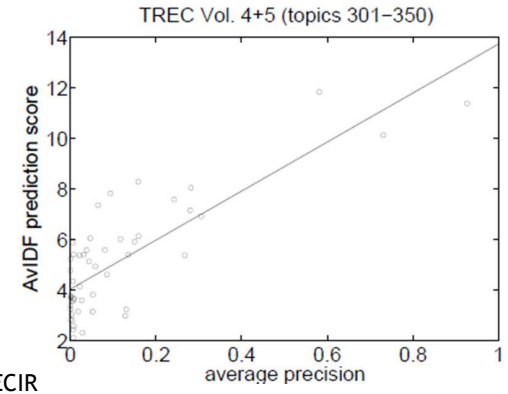
LINEAR CORRELATION BRAVAIS-PEARSON



QUERY DIFFICULTY PREDICTORS

- IDF

TF.IDF based retrieval system, $MAP = 0.11$, $r = 0.81$

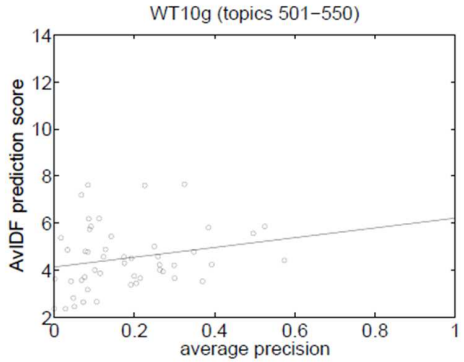


Hauff et al., 2009, ECIR

QUERY DIFFICULTY PREDICTORS

- IDF

Language Modeling based retrieval system, $MAP = 0.18$, $r = 0.22$



Hauff et al., 2009, ECIR

QUERY DIFFICULTY PREDICTORS

- IDF

		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**
AMB	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.

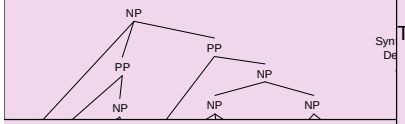
LINGUISTIC QUERY DIFFICULTY PREDICTORS

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

MAIN RESEARCH DIRECTIONS

- Query difficulty prediction
- **Adaptive system**
- 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

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

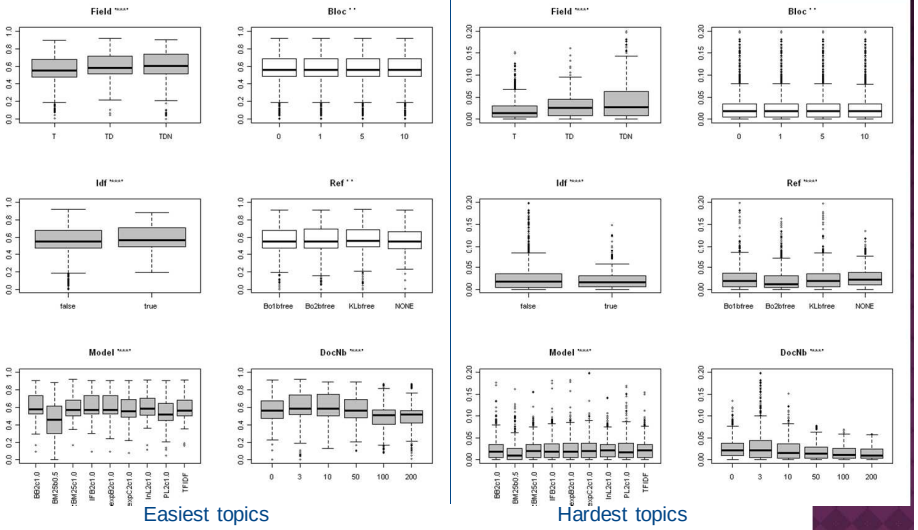
WHAT ARE THE MOST INFLUENTIAL SYSTEM PARAMETERS

- **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	Bolbfree	BB2c1.0	3	0.6134
2	352	T	1	false	Bolbfree	BB2c1.0	3	0.3412
3	353	T	1	false	Bolbfree	BB2c1.0	3	0.3479
4	354	T	1	false	Bolbfree	BB2c1.0	3	0.0662
5	355	T	1	false	Bolbfree	BB2c1.0	3	0.2794
6	356	T	1	false	Bolbfree	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

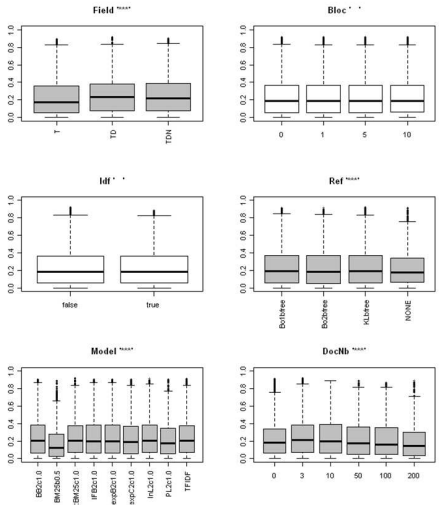
Significant effect (1-factor ANOVA)

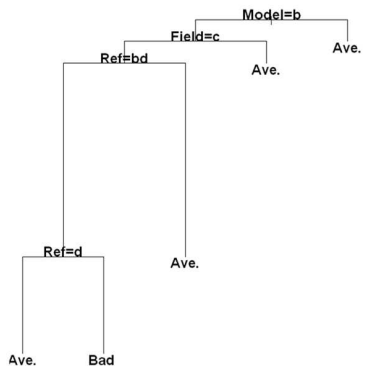


Easiest topics

Hardest topics

WHAT ARE THE MOST INFLUENTIAL SYSTEM PARAMETERS

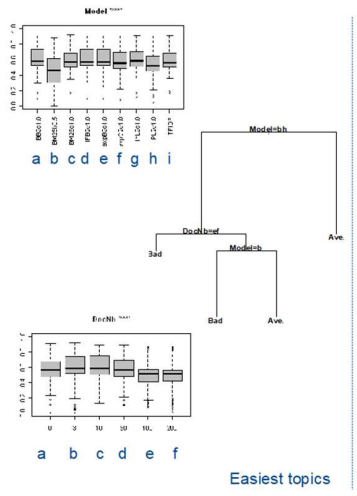




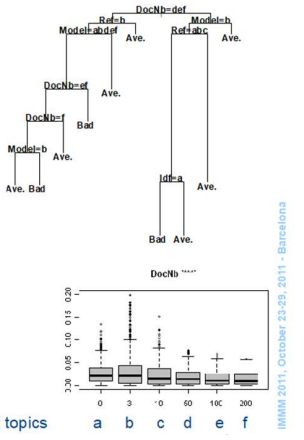
WHAT ARE THE MOST INFLUENTIAL SYSTEM PARAMETERS

➲ Multivariate analysis : CART

Classification And Regression Tree



Easiest topics



Hardest topics

WHICH SYSTEM TO USE?

- 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

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.

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?

- 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?

- 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	
GBRT (All)	0.3338 △		0.3439 △	
- QueryStats	0.3375 △	(+1.11%)	0.3658 △	(+6.35%)
- QueryLing	0.2982 △	(-10.68%)	0.3507 △	(+1.96%)
- RetModel	0.3299 △	(-1.17%)	0.3462 △	(+0.65%)
- Expansion	0.2345▼	(-29.75%)	0.2384▼	(-30.69%)
GBRT (All)	0.3338 △		0.2803 △		0.3204	
- QueryStats	0.3375 △	(+1.11%)	0.2699	(-3.71%)	0.3304 △	(+3.12%)
- QueryLing	0.2982 △	(-10.68%)	0.2908	(+3.75%)	0.3498 △	(+9.19%)
- RetModel	0.3299 △	(-1.17%)	0.2702	(-3.62%)	0.3400 △	(+6.10%)
- Expansion	0.2345▼	(-29.75%)	0.1775▼	(-36.66%)	0.1914▼	(-40.28%)
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	

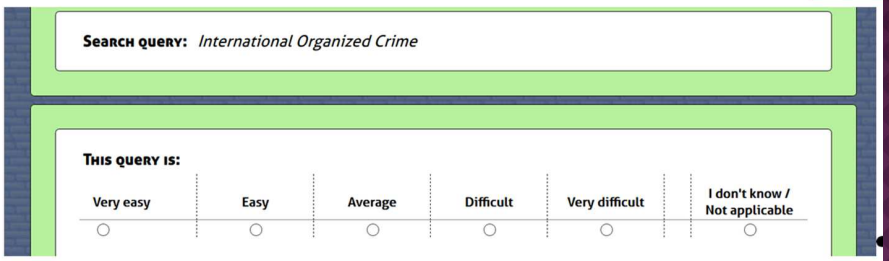
MAIN RESEARCH DIRECTIONS

- 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

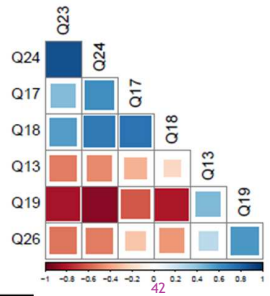
R4: The query contains generic word(s)
R10: The topic is Unusual/uncommon/unknown
R11: The topic is too broad/general/large/vague
R12: The topic is specialized
R26: The number of query words is too high
R16 The topic is too precise/specific/focused/delimited/clear
R23: Many of the relevant documents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

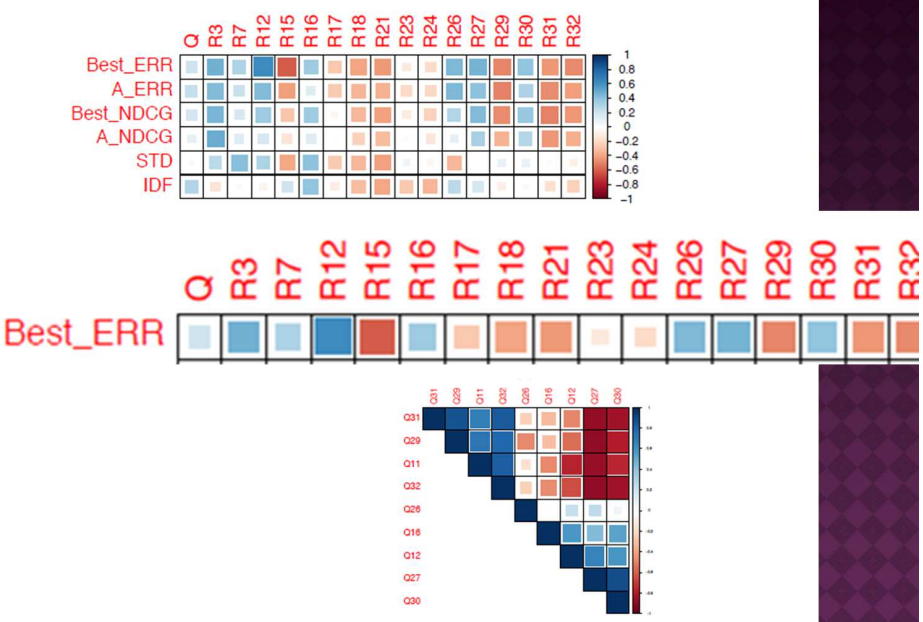
Figure 3: Examples of reasons resulting from the recoding of free text annotation on query difficulty comments.

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





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

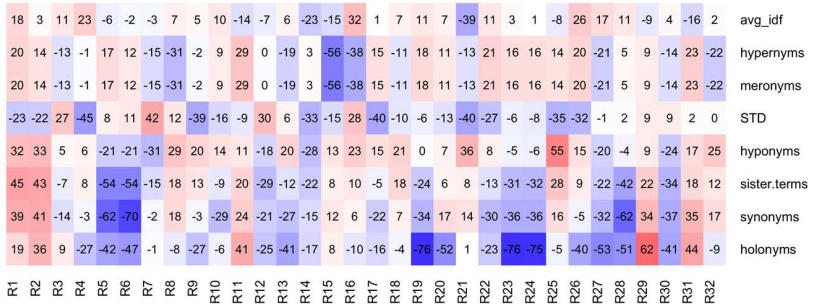
- Correlation with human « prediction »

	Reason	Correlation Q	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
Some	R10: The topic is Unusual/uncommon/unknown	0.526	0.496
	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



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

No need to ask them

- Reasons of difficulty make sense to 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

@JosianeMotheFr