ECIR, 2023

Josiane.Mothe
@irit.fr

On correlation to evaluate QPP

Objectives

- Open discussion on evaluation for QPP
- Question on the use of some measures that may not be appropriate to use

Evaluation of query difficulty predictors

- Evaluate if a feature / a model is a good predictor

Query/Topic	Predicted difficulty	Measured difficulty
Id1	0.60	0.50
Id2	0.45	0.45
Id3	0.70	0.80
Id4	0.20	0.10
Id5	0.10	0

- Two variables

→ are they independent ?

→ Correlation

Correlation

measures the strength and direction of association between two variables

- Used to evaluate different IR tasks
  - *Evaluate two ranked lists with automatic relevance judgments vs human ones*
  - *Evaluate users' satisfaction vs system effectiveness*
  - *QPP*

Correlation

- Pearson

correlation coefficient between two random variables $X(x_1, x_2, ...x_i, ..., x_N)$ and $Y(y_1, y_2, ...y_i, ..., y_N)$ is defined as:

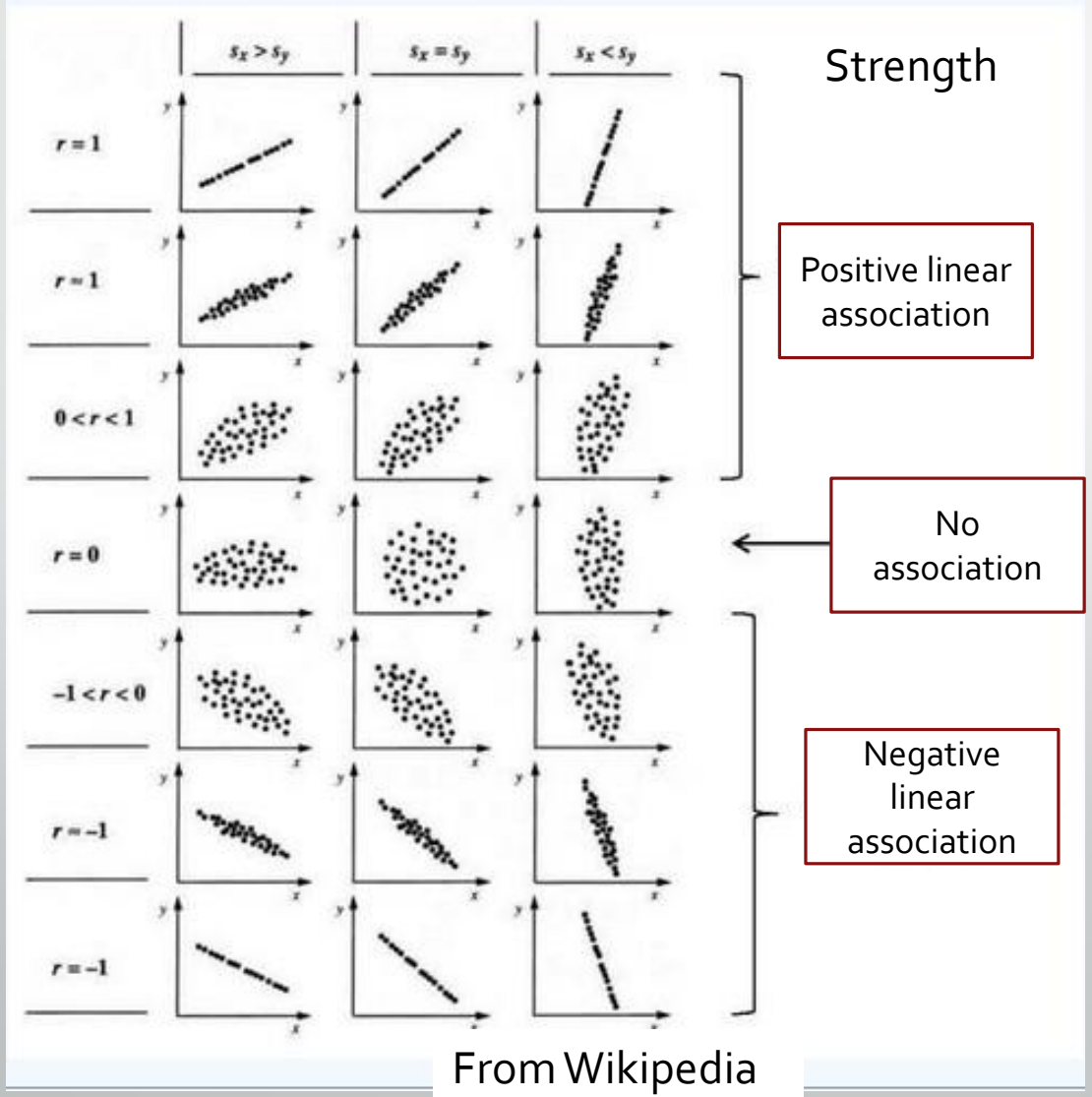
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}.$$

Where

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

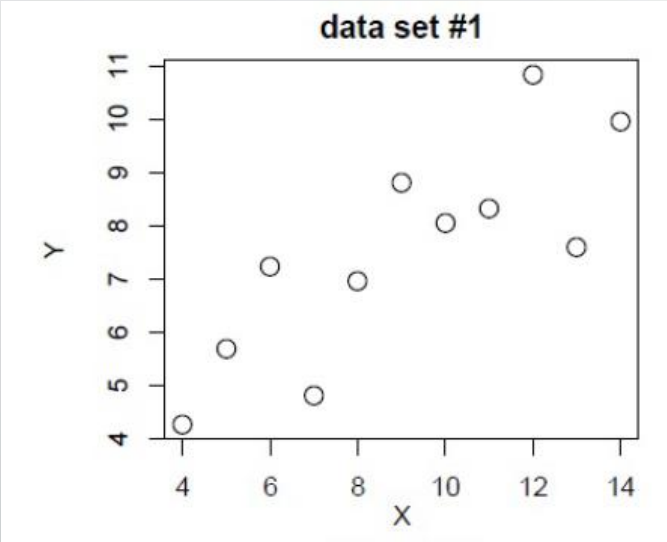
$$\sigma(X)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

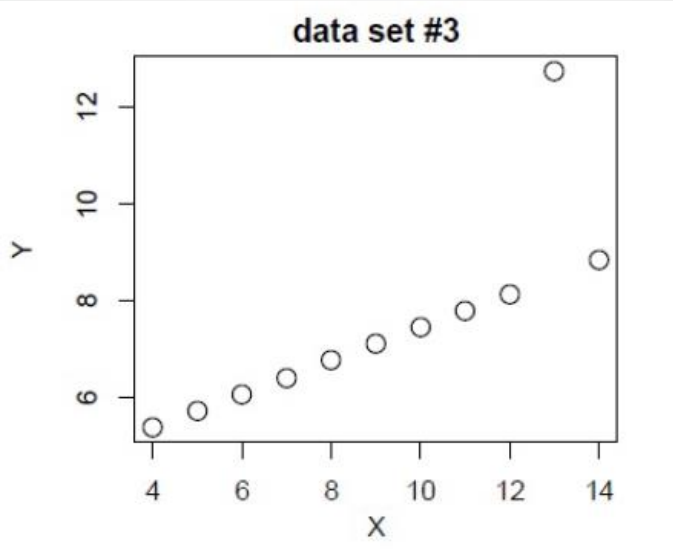
Linear correlation Bravais-Pearson



From Wikipedia

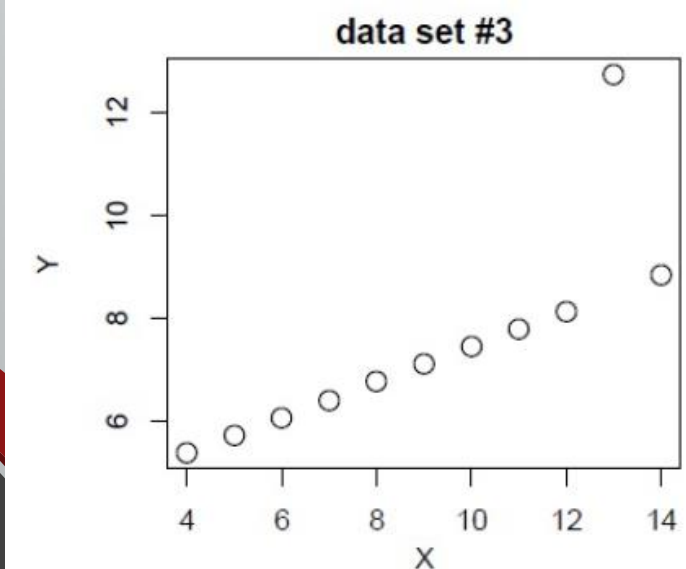
Guess the value of ρ

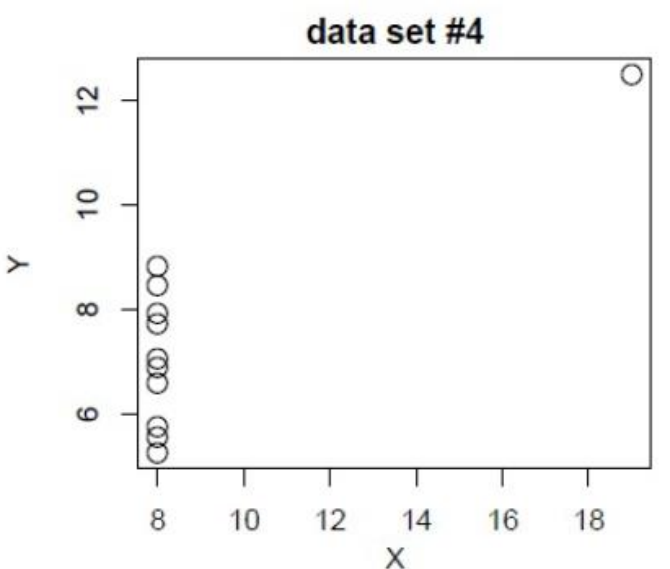




0.8164 (P-value 0.0022)

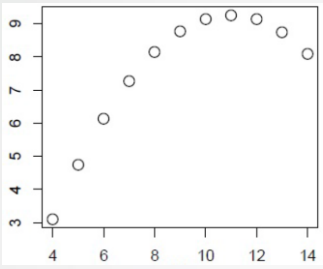
Anscombe's quarte





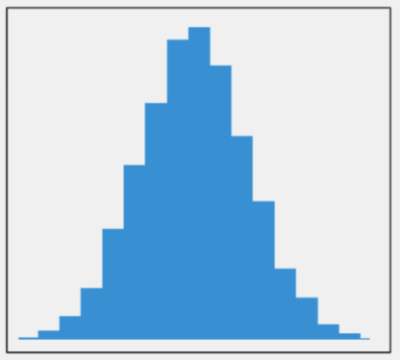
Correlation

- Assumptions
  - Linear link
  - Outlier free
  - Continuous
  - Normally distributed
  - Similar spread across range



Non linear
(& non monotonic)

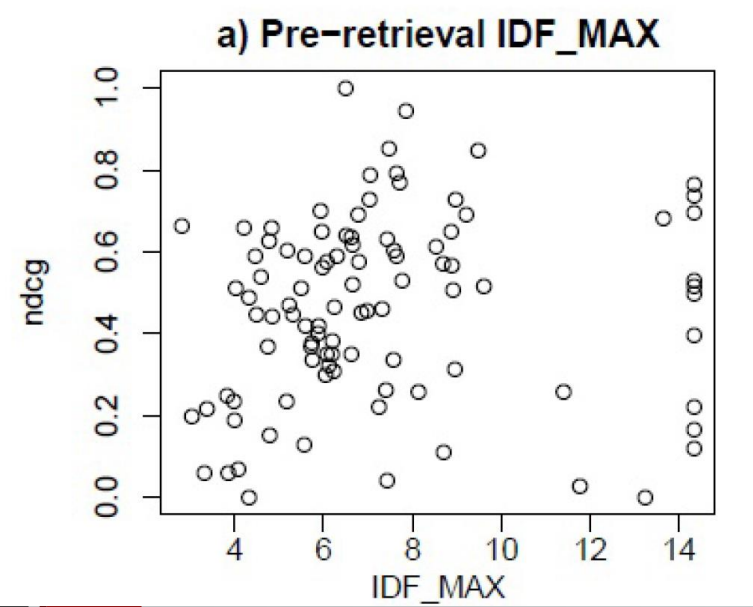
Normal distribution

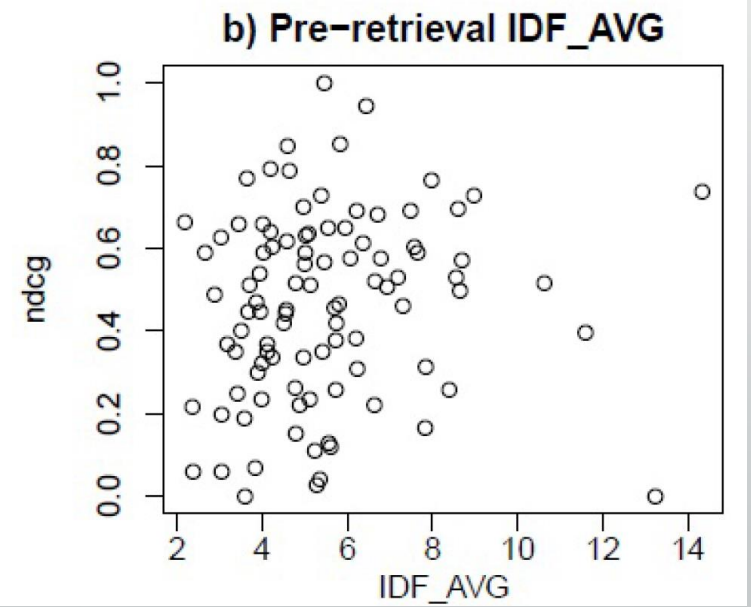


Correlation

- Assumptions
  - Linear link
  - Outlier free
  - Continuous
  - Normally distributed
  - Similar spread across range

measure of how close the observations are to a line of best fit.





Significance

- Null hypothesis:
  $H0 : \rho = 0$
  (no statistical link between the two variables) vs.

  $H1 : \rho <> 0$
  There is a statistical link between the two variables

  In bivariate normal data, $\rho = 0$ if and only if X and Y are independent. So testing for independence is equivalent to testing $\rho = 0$ in this situation.
- P-value: the null hypothesis is rejected if the p-value is less than or equal to a predefined threshold value (0.05) - is due to chance 5%

Correlation other than Pearson

- **Spearman** considers ranks rather than values
  thus measures how far from each other variable ranks are

Query/Topic	Predicted difficulty	Measured difficulty	Predicted Rank	Measured Rank
Id1	0.60	0.50	2	2
Id2	0.45	0.45	3	3
Id3	0.70	0.80	1	1
Id4	0.20	0.10	4	4
Id5	0.10	0	5	5

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)} \sigma_{\mathrm{R}(Y)}}$$

H_0: There is no [monotonic] association between the two variables.

- Data are ordinal (numerical or categorial)
- No assumption on linearity of the link (but monotonic, can have outliers

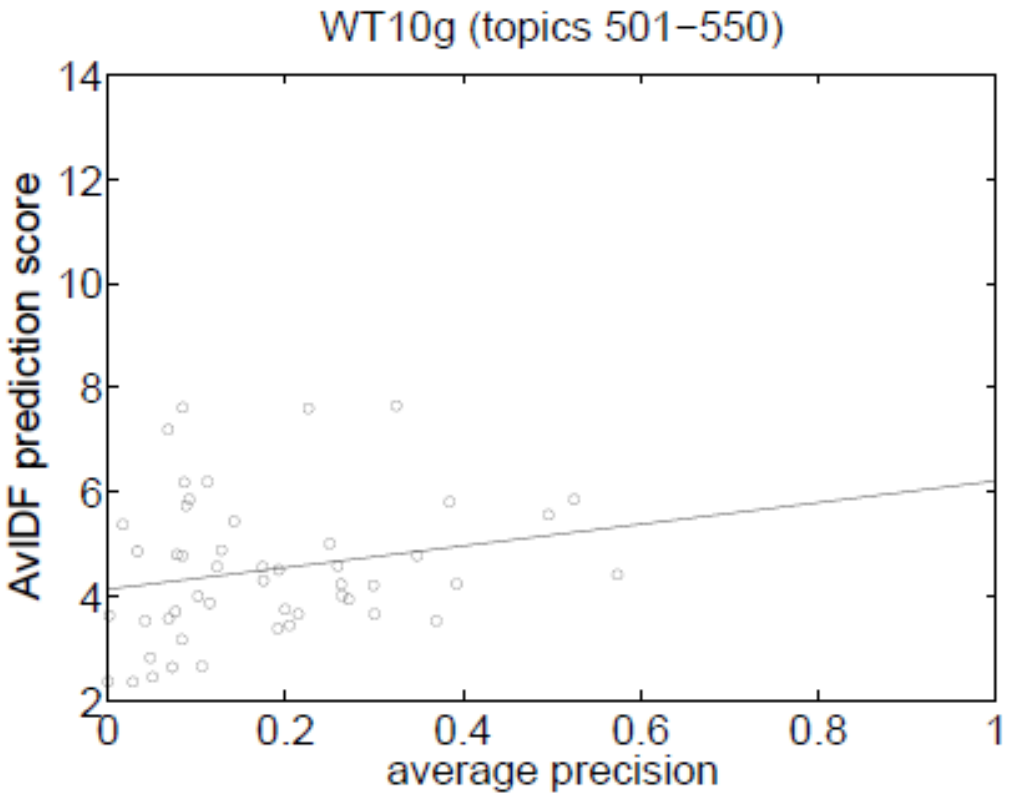
Correlation other than Pearson

- Kendall measures the correlation on ranks

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

query difficulty predictors

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



Hauff et al., 2009, ECIR

query difficulty predictors

		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**

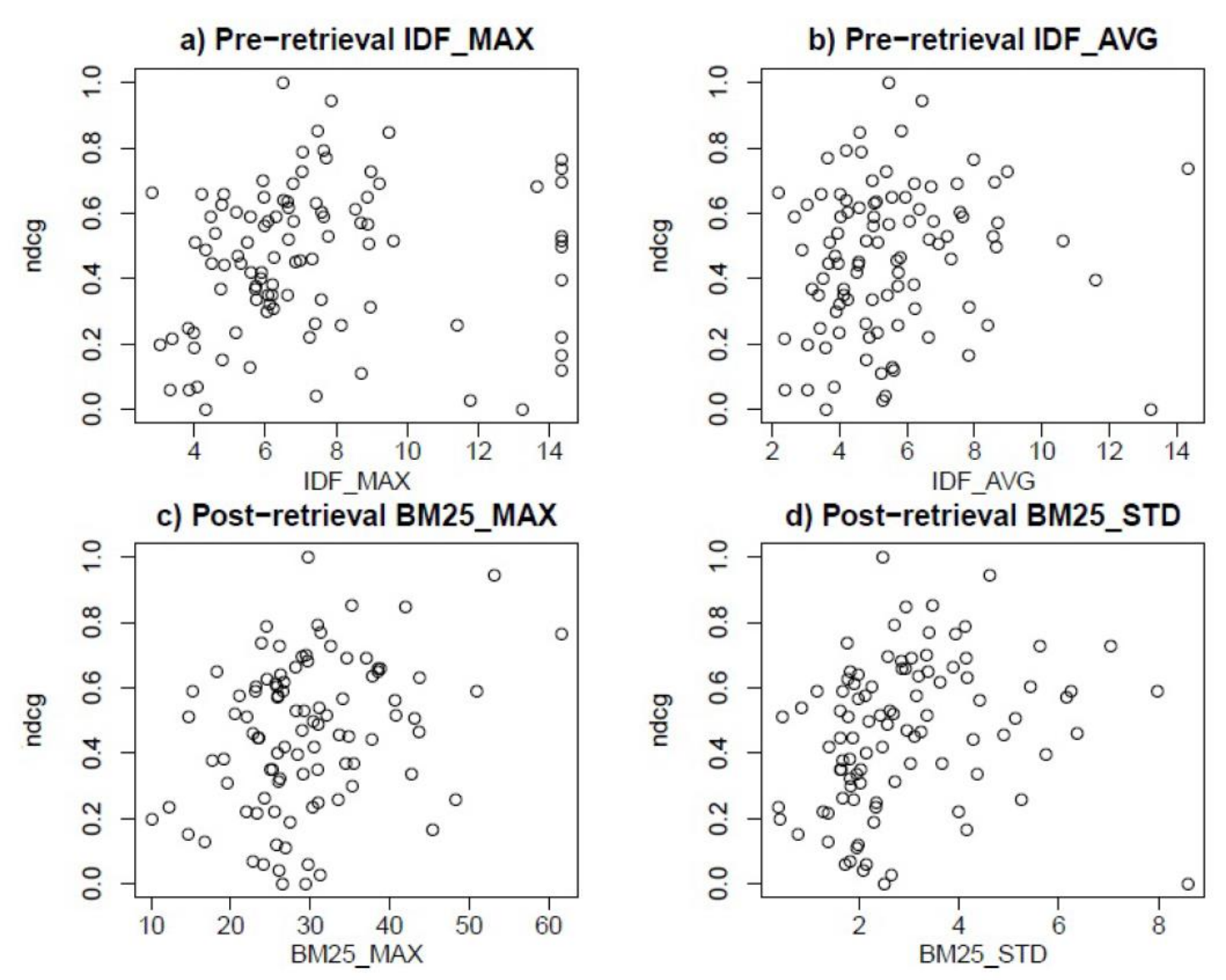
Table 1: Results of the predictor evaluations given by the linear correlation coefficient.

Hauff et al., 2008, CIKM

Table 3

Correlation between query features and ndcg. * marks the usual <0.05 P-Value significance

	Feature			
Measure	BM25_MAX	BM25_STD	IDF_MAX	IDF_AVG
Pearson ρ	0.294*	0.232*	0.095	0.127
P-Value	0.0034	0.0224	0.3531	0.2125
Spearman r	0.260*	0.348*	0.236*	0.196
P-Value	0.0100	<0.001	0.0202	0.0544
Kendall τ	0.172*	0.230*	0.159*	0.136*
P-Value	0.0128	<0.001	0.0215	0.0485


a) Pre−retrieval IDF_MAX
ndcg
0.0
0.2
0.4
0.6
0.8
1.0
4
6
8
10
12
14
IDF_MAX
b) Pre−retrieval IDF_AVG
ndcg
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
12
14
IDF_AVG
c) Post−retrieval BM25_MAX
ndcg
0.0
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
60
BM25_MAX
d) Post−retrieval BM25_STD
ndcg
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
BM25_STD

Measure	Feature BM25_MAX	BM25_STD	IDF_MAX	IDF_AVG
Pearson ρ	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*	0.339*	0.142	0.225*
r	0.268	0.342	0.234	0.183
τ	0.181*	0.225	0.162*	0.120

Conclusion

- **Disagreement among methods be seen as a warning**
- **Plot the data to make sure that the calculated coefficients are meaningfull and comparable**
- **Outliers**
- **Coefficients should be used with caution when comparing different predictors**

References

- F. J. Anscombe, Graphs in statistical analysis, The american statistician 27 (1973)
- D. Carmel, E. Yom-Tov, Estimating the query difficulty for information retrieval, (2010)
- C. Hauff, D. Hiemstra, F. de Jong, A survey of pre-retrieval query performance predictors (2008)
- J. Mothe, Analytics methods to understand information retrieval effectiveness—a survey (2022)
- J. Mothe, On correlation to evaluate QPP (ECIR,2023)
- E Poesina, R.T. Ionescu, J. Mothe, iQPP: A Benchmark for Image Query Performance Prediction (SIGIR, 2023)