On Correlation to Evaluate QPP

Josiane Mothe

On Correlation to Evaluate QPP

Josiane Mothe*

[1]Institut de Recherche en Informatique de Toulouse, IRIT, UMR5505, CNRS, Toulouse, France
[2]INSPE, UT2J, Université de Toulouse, Toulouse, France

Abstract

Correlation is widely used to test the hypothesis of the relationship between two variables. In this paper we chose to focus the discussion on query difficulty prediction for which correlation is often used to measure the accuracy of predictors. Here, the correlation is calculated between the actual system effectiveness and the predicted one. Although fairly simple to calculate, the Pearson correlation coefficient can be difficult to interpret and use correctly, especially because of its sensitivity to outliers. This paper illustrates the problem and opens discussion pathways.

Keywords

Information systems, Information retrieval, Query performance prediction, Evaluation, Correlation

1. Introduction

In many scientific domains where variables are used, it is often the case that we need to know whether two variables are independent or not. A correlation test is a hypothesis test for a relationship between two variables. In other words, it can be used to measure the dependence between two quantities.

In information retrieval (IR), correlation is used in various cases to measure the effectiveness of a proposed method such as in the following examples:

- When automatically generating a reliable set of relevance judgements (pseudo relevance judgements), Ravana *et al.* evaluate effectiveness by the correlation coefficient of two ranked system lists considering mean average precision scores between the original Text REtrieval Conference relevance judgements and pseudo relevance judgements [1];
- Correlation is used to evaluate the link between user satisfaction and system effectiveness [2] or, in the context of conversational search, to evaluate the link between the length of questions or answers and improvement in NDCG when terms are added to the query [3].
- In query difficulty prediction, accuracy is measured in terms of correlation between actual system effectiveness and the predicted effectiveness; Correlation is also often used to compare different variables (e.g. different query difficulty predictors) with regard to their link with a target variable (e.g. MAP or NDCG) [4, 5, 6, 7].

QPP++ 2023: Query Performance Prediction and Its Evaluation in New Tasks, co-located with The 45th European Conference on Information Retrieval (ECIR) April 2, 2023, Dublin, Ireland

*Corresponding author.
Josiane.Mothe@irit.fr (J. Mothe)
https://www.irit.fr/~Josiane.Mothe (J. Mothe)
0000-0001-9273-2193 (J. Mothe)

There are various methods to calculate the relationship between two variables; correlation coefficient is one of them. Among correlation coefficients, Pearson product-moment is the most used. Kendall and Spearman correlations are other measures used when two variables are to be analysed.

Correlation calculation results in a value that ranges between -1 (strong negative correlation) and 1 (strong positive correlation); 0 indicating that the two variables are not correlated. $p - value$ indicates the confidence or risk of error in rejecting the hypothesis that the two variables are independent.

This paper aims to discuss the possible misinterpretation of correlation through some examples. Here, we mainly focus on Pearson correlation which is the most used in QPP, although we also consider the other correlation coefficients.

In addition to some assumptions made on the variables, which we describe in section 2, one of the main problems in using Pearson correlation measure is its sensitivity to outliers [8, 9] that we illustrate in Section 3. The specific case of QPP is studied in Section 4. Section 5 concludes this paper.

2. Correlation measures

The most familiar measure of correlation is the Pearson product-moment correlation coefficient (also called the correlation coefficient and labelled ρ) which is a normalised form of the covariance. Covariance between two random variables measures their join distance to their expected values which can be the distance to the mean for numerical data. Pearson ρ assumes linear relationship between X and Y.

More formally, ρ is calculated by dividing the covariance of the two variables by the product of their standard deviations and 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

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

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

Thus, this correlation coefficient measures the link between the two variables by measuring the mean of the product of the distance of the two variables to their respective mean. When is it close to 1 or -1, the two variables are strongly correlated (positively or negatively); confirming the hypothesis that there is a linear relationship between the two variables.

Alternatively, Spearman's correlation (r_s) considers the ranks rather than the values and measures how far from each other variable ranks are. r_s is similar to Pearson on ranks ($\rho = r$ once column of X and Y are replaced by their ranks). Spearman's assumes monotonic relationship between X and Y.

Similarly, Kendall correlation measures the correlation on ranks, that is the similarity of the orderings of the data when ranked by each of the variable values. It is affected by whether the ranks between observations are the same or not without considering how far they are as opposed to r. It is thus considered as more appropriate for discrete variables. Kendall measures the concordance of any pair of observations (x_i, y_i) and (x_j, y_j), where $i \neq j$. The pair is said to be concordant if the ranks for both elements agree ($x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$); discordant if the reverse occurs. The Kendall τ coefficient is defined as:

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

Whatever the correlation measure is, to be significant, the link between the two variables should not be due to the data sample only (i.e. random) but should reflect the link between the two variables on the entire population. Testing the null hypothesis aims at answering this issue.

Thus, when considering Pearson correlation (but the same holds for the other correlation measures) what is tested is $H_0 : \rho = 0$ (no statistical link between the two variables) vs. $H_1 : \rho \neq 0$ (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.

The null hypothesis $H_0 : \rho = 0$. (there is no relationship between the two variables X and Y) is usually rejected when $p - value < 0.05$ (and thus the variables are considered as related in that case). The p-value is a number between 0 and 1 representing the likelihood of the observation if the hypothesis is assumed to be correct. The statistically significance result is considered as highly improbable if the null hypothesis is assumed to be true.

Thus calculating $\rho(X, Y)$ and checking $p - value < 0.05$ is commonly used in order to conclude whether X and Y are related.

Correlation is easy to calculate although some misinterpretation or over-interpretation can occur as illustrated by the Anscombe's quartet and presented in the next section.

3. Anscombe's quartet

Anscombe illustrates the complementary aspect of correlation calculation with the graphical plotting of data [10].

Table 1 presents the 4 data sets Anscombe designed: each element is represented by two variables X and Y for which we want to know whether they correlate or not. Table 2 presents some statistics of the 4 data sets; it reports that various aggregation values are the same for the 4 data sets: the number of elements, the mean of the variable X and the one of Y, as well as the Pearson correlation and the associated P-values. In addition, from the same table, it can be observed that the ρ correlation value is higher than 0.816 (which is considered as a high value), p-value < 0.05 (which is considered as significant).

Because the real data may not respect the mathematical assumption (linear relationship between X and Y in the case of ρ) and because ρ is also sensitive to outliers, without having a look to the data and simply trusting the ρ value and the associate p-value, one could consider

Table 1
Anscombe's quartet - Data from [10]. Although different, the 4 data sets share various aggregation values as presented in Table 2.

Data set	#1	Data set	#2	Data set	#3	Data set	#4
X	Y	X	Y	X	Y	X	Y
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	7.04
14	9.96	14	8.1	14	8.84	8	8.47
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
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

Table 2
Statistical features of the Ascombe's data sets

Data set	#1	#2	#3	#4
#elements	11	11	11	11
Mean X	9	9	9	9
Mean Y	7.5009	7.5009	7.5009	7.5009
Pearson Correlation	0.8164	0.8162	0.8163	0.8165
P-value	0.0022	0.0022	0.0022	0.0022

the 4 cases are equivalent in terms of importance of the correlation. However, data plots tell a different story (see Figure 1).

When plotting the corresponding dots as in Figure 1, it is obvious that the 4 data sets are very different. For data set #1, 0.816 seems to reflect appropriately the linear correlation between X and Y. In data set #2, there is a clear correlation between X and Y but which is far for being linear. In this latter case, a different correlation measure may better reflect this perfect correlation. In data set #3, the correlation between X and Y would be 1 if the outlier was removed from the data set. This outlier abnormally lower the correlation value. Finally, in data set #4, there is no correlation at all but the high correlation value is due to an outlier. Removing this outlier would make the correlation 0.

Anscombe quartet illustrates that correlation value can not be considered without having a look at the plots. However, most of the time in IR studies (and in others areas as well), correlation is reported without considering plotting, but just "trusting" the associated p-value. There is thus a risk of misinterpretation. The more that many authors use the correlation coefficients without checking if the assumptions are met (e.g. linear correlation in case of Pearson correlation).

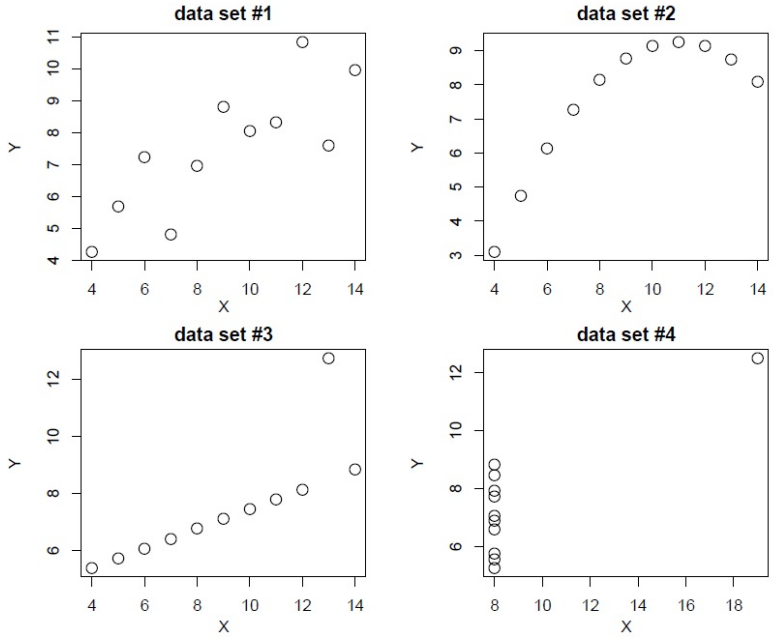

Figure 1: Anscombe's quartet and Pearson correlation. Although different, the 4 data sets share various aggregation values (such as the correlation between X and Y (cor=0.816).

4. Query difficulty predictors and correlation

In query difficulty prediction, the accuracy of a predictor is often measured in terms of how much the values of the predictor correlates with the actual system effectiveness.

In this section, we consider NDCG as the system effectiveness measure and thus as the value to be predicted by the query difficulty predictor. The system we used here is a simple BM25 weighting schema. We also consider as illustrative examples two well known query difficulty predictors BM25 and IDF. BM25 is based on the scores retrieved documents obtained; it is thus a post retrieval feature. IDF on the other hand is a pre-retrieval feature based on query word IDF. We consider two variants that have been used in the literature for these two features: maximum and standard deviation for BM25 later referenced as BM25_MAX and BM25_STD (the maximum and standard deviation of BM25 weights for document - query pairs for that query); and maximum and average for IDF later referenced as IDF_MAX and IDF_AVG (the maximum and average inverse document frequency of the query terms).

4.1. Measuring correlation

A typical problem is to compare the accuracy of different variables (here it would be these four features) to predict query difficulty. One common solution is to consider correlation between each variable that corresponds to a predictor and the target variable that represents the system effectiveness (e.g. NDCG).

Table 3 reports the Pearson correlation as well as Kendall τ and Spearman correlation of the 4 query features with NDCG on WT10G TREC collection which consists of topics 451-550 and

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

Table 4
Correlation between query features and NDCG. P-Value is indicated by * mark using the usual < 0.05 threshold

	Feature			
correlation	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

about 1.7 millions of web pages. The three calculations agree on the fact that the correlation values are weak; which is often the case in this task [11]. They also agree on that BM25 post-retrieval features are better predictors than IDF pre-retrieval features and that IDF_AVG is weakly correlated with NDCG and generally not significantly; IDF_MAX's correlation is also weak. However, the three correlation measures disagree on the best predictor: while Pearson suggests BM25_MAX is the best, Kendall and Spearman prefer BM25_STD.

Should the disagreement among methods be seen as a warning when discussing the results and making conclusions? We believe so.

4.2. Plotting the data

Figure 2 displays the WT10G topics (NDCG as Y-axis and the displayed predictor as X-axis). Visually, it becomes difficult to see which is the best predictor for NDCG.

We can see that IDF_Max has many outliers (right side of Figure 2a). If we removed these outliers with very high IDF_Max, then, the rest of the measures are much more correlated than the measures of BM25_Max where it is difficult to identify any correlation.

Should we plot the data to make sure that the calculated coefficients are meaningful and comparable? We think so.

4.3. Impact of outliers

While observing the (Pearson) correlation value only (first line of Table 3, BM25_MAX is more correlated to NDCG than BM25_STD; both being statistically significantly correlated. When

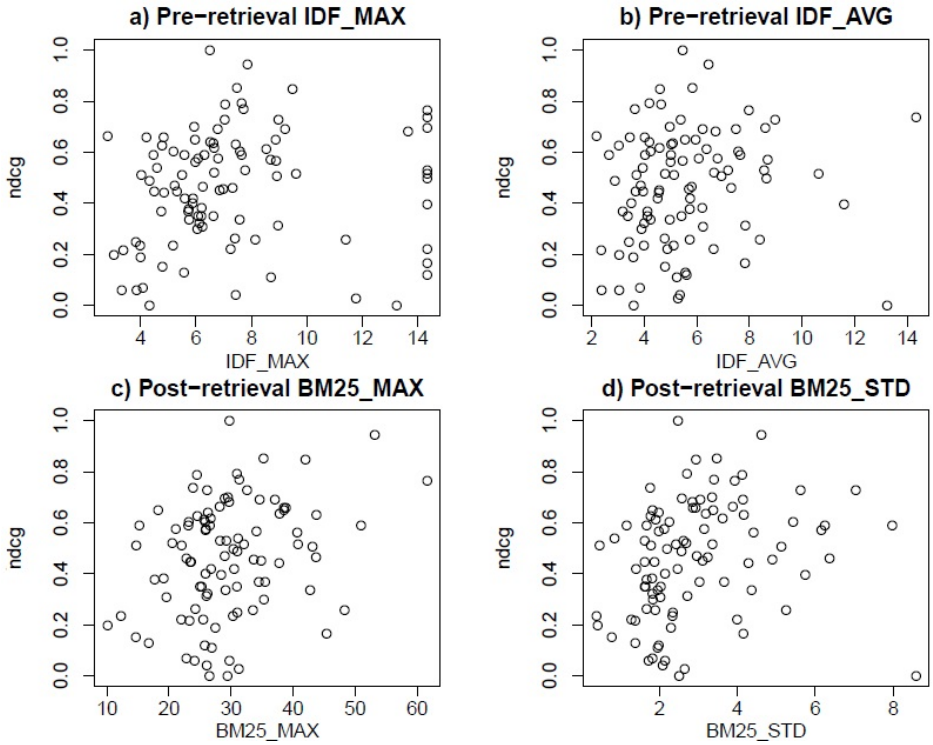

Figure 2: Topics visualisation considering either IDF pre- or BM25 post-retrieval predictor (Xaxis) and NDCG (Yaxis) values.

observing the plots in Figure 2, we can see that a topic (#463) in the right-side bottom corner of Figure 2 d. is an "outlier" (like the outlier from the 3rd Anscombe's data set). If we remove this outlier and calculate the correlation again, we obtain the first group of rows in Table 4. Indeed, when removing this single topic from the collection, the Pearson correlation from BM25_STD increases of about 46% (from 0.232 to 0.339) and becomes higher than BM25_MAX, while the later is stable (0.294).

In the same way, when considering IDF_AVG, the numerical results indicates that the independence cannot be rejected (cor=0.127 and p-value=0.2125). Removing topic 463 from the collection when analysing IDF_AVG, the correlation is doubled, but more importantly, while it was not significant initially, the independence can be rejected with quite high confidence now (the p-value 0.027 is lower than the commonly used 0.05 value).

We believe that the coefficients should be used with caution when comparing different predictors.

5. Conclusion and future work

In this paper we point out the need of discussions on the use of correlation coefficient for query performance prediction. We illustrated the possible misinterpretation of correlation measures. This is a challenge when comparing several variables with regard to their link with a target

variable.

The influence of outliers has been little studied in the case of correlation coefficient.

In the case of Principal Component Analysis, which is also used to analyse variable relationships when a large number of variables are involved, Kriegel *et al.* proposed an approach to increase the robustness. They suggested to use weighted covariance in order to make PCA less sensitive to outliers [12]. In the case of regression, Huang *et al.* proposed the Robust regression [13] also to make the method less sensitive to outliers. To the best of our knowledge, nothing similar has been proposed for correlation. Considering the popularity of this method; it would be worth investigating this problem.

References

[1] S. D. Ravana, P. Rajagopal, V. Balakrishnan, Ranking retrieval systems using pseudo relevance judgments, Aslib Journal of Information Management 67 (2015) 700–714.

[2] A. Al-Maskari, M. Sanderson, A review of factors influencing user satisfaction in information retrieval, Journal of the American Society for Information Science and Technology 61 (2010) 859–868.

[3] A. M. Krasakis, M. Aliannejadi, N. Voskarides, E. Kanoulas, Analysing the effect of clarifying questions on document ranking in conversational search, in: Proc. of the ACM SIGIR Intern. Conference on Theory of Information Retrieval, 2020, pp. 129–132.

[4] D. Carmel, E. Yom-Tov, Estimating the query difficulty for information retrieval, Synthesis Lectures on Information Concepts, Retrieval, and Services 2 (2010) 1–89.

[5] S. Mizzaro, J. Mothe, Why do you think this query is difficult?: A user study on human query prediction, in: Proc. of the 39th Inter. ACM SIGIR conference on Research and Development in Information Retrieval, 2016, pp. 1073–1076.

[6] C. Hauff, D. Hiemstra, F. de Jong, A survey of pre-retrieval query performance predictors, in: Proc. of the 17th ACM Conference on Information and Knowledge Management, ACM, 2008, pp. 1419–1420.

[7] S. Datta, D. Ganguly, M. Mitra, D. Greene, A relative information gain-based query performance prediction framework with generated query variants, ACM Transactions on Information Systems 41 (2022) 1–31.

[8] R. K. Pearson, Exploring process data, Journal of Process Control 11 (2001) 179–194.

[9] G. Casper, C. Tufis, Correlation versus interchangeability: The limited robustness of empirical findings on democracy using highly correlated data sets, Political Analysis 11 (2003) 196–203.

[10] F. J. Anscombe, Graphs in statistical analysis, The american statistician 27 (1973) 17–21.

[11] J. Mothe, Analytics methods to understand information retrieval effectiveness—a survey, Mathematics 10 (2022) 2135.

[12] H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, A general framework for increasing the robustness of pca-based correlation clustering algorithms, in: International Conference on Scientific and Statistical Database Management, Springer, 2008, pp. 418–435.

[13] D. Huang, R. Cabral, F. De la Torre, Robust regression, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2016) 363–375.