On Rank Correlation in Information Retrieval Evaluation

Massimo Melucci
University of Padua
Department of Information Engineering
massimo.melucci@dei.unipd.it

Abstract

Some methods for rank correlation in evaluation are examined and their relative advantages and disadvantages are discussed. In particular, it is suggested that different test statistics should be used for providing additional information about the experiments other that the one provided by statistical significance testing. Kendall's τ is often used for testing rank correlation, yet it is little appropriate if the objective of the test is different from what τ was designed for. In particular, attention should be paid to the null hypothesis. Other measures for rank correlation are described. If one test statistic suggests to reject a hypothesis, other test statistics should be used to support or to revise the decision. The paper then focuses on rank correlation between webpage lists ordered by PageRank for applying the general reflections on these test statistics. An interpretation of PageRank behaviour is provided on the basis of the discussion of the test statistics for rank correlation.

1 Introduction

Ranking is a natural process performed by the IR systems which assign numerical scores to the objects for measuring the presence of a property in each object; for example, the ranking by probability of relevance is an estimation of the unknown, true ranking in which relevant documents are ordered before non-relevant documents. Other instances are term ranking for query expansion purposes, system ranking by Mean Average Precision, webpage[1] ranking by PageRank, or database ranking in distributed retrieval. The order imposed by a system on a set of objects is crucial because it determines the effectiveness of the system and then its success or failure when performing the task for which the system was designed.

When statistics is used for comparing two or more rankings, the rankings compared are either two samples in which each object has been assigned one value, or one sample in which each object has been assigned two values. Rank correlation refers to the suite of statistical methods for measuring the degree to which two rankings are correlated, where correlation signifies the tendency of the values of one ranking to be in the same order as the order of the values of the other ranking. A test statistics is then computed for deciding the correlation between the rankings — a test

[1] A "webpage" is a document identified by an URL in the Web.

statistic is a function of two rankings and provides a value used for deciding about the correlation or similarity between the rankings. As the test statistic is a random variable, a probability distribution permits for calculating the probability that the decision about the correlation between the rankings is correct.

As the assumption of normality of the values of the rankings is hardly applicable in IR, parametric statistics are not generally applicable [1]. As a consequence, nonparametric statistics is the common statistical approach adopted because no normality assumption is needed. The suite of nonparametric rank correlation methods is illustrated, for example, in [2, Chapter 5].

Since order is crucial in IR, it is little surprising that plenty of research papers reported experimental analyses of rank correlation. In particular, an investigation of the relationships between the webpage ranking induced by PageRank when the damping factor or the number of terms of a power series vary is reported in [3]. The correlation was measured by Kendall's τ, which increases when the number of exchanges between the webpages which are necessary for transforming one ranking to the other decreases. A statistical analysis revealed that the correlation between two rankings increased when the size of the rankings increased. In particular, τ became very close to 1 when only a few hundred webpages rankings were compared, even though the high number of exchanges. This behaviour of τ seemed a paradox because a high number of exchanges would indicate low correlation between rankings; for example, if these exchanges affected the top ranked webpages, two rankings would be perceived as different by the end user of a search engine, even when the rankings include several thousand webpages. Of course, one may argue that if the high number of exchanges corresponds to only 10% of the potential exchanges occurs, the two rankings should be considered as correlated when a test statistic suggests that the number of exchanges is not statistically significant.

This paradox can be addressed when the statistical analysis of rank correlation is not only limited to the computation of a single correlation value and of its statistical significance, but also analyses what type of correlation has been observed. This paper argues that the investigation of rank correlation benefits from diverse test statistics which provide different information about the comparison of two rankings. The diversity of information is beneficial because each test give insights into the correlation from different angles or points out different types of correlation.

While the investigation reported in this paper not only helped understand why and how two webpage rankings are correlate in the general IR context, the paper will focus on the specific context of webpage ranking studied in [3] and in related papers thus providing information about the behaviour of PageRank and similar algorithms.

2 Related Work

This section concentrates on some results obtained in IR by using a rank-based correlation method. This is a non-exhaustive sample of all the research conducted in the field. Moreover, more mathematical research works, e.g. [4], are outside the scope of this paper, which rather focuses on the algorithmic properties of the rank correlation measures. The research reported here can be classified into three broad topics: Web IR, Evaluation, Distributed IR. The reader may want to read Appendix A for a quick introduction to the notion of Null Hypothesis and Appendixes B and C for the description of Kendall's τ and Spearman's ρ, which are cited in this section — both

measures range between -1 and $+1$, are as larger as the correlation increases, and tend to 0 in the event of no correlation.

2.1 Rank Correlation in Web IR

In [5] the authors addressed the question whether link-based metrics correlate with human judgements of quality and in general whether link-based metrics are good predictors of webpage quality. Moreover they investigated the correlation between content-based and link-based metrics. Other correlations were also investigated, that is, (1) the extent to which subjects agreed on the quality of webpages and (2) whether any significant differences can be observed within a topic between various link analysis algorithms. The correlation between subjects is a fundamental issue extensively studied in the past in distinct contexts — inter-indexer consistency [6, 7, 8] and inter-linker consistency in automatic hypertext construction [9]. The web sites were ranked by diverse features — one feature at a time — and rank correlation between the rankings were measured by Spearman's ρ and τ. The correlations were very high and were statistically significant. In particular the in-degree and the out-degree were found highly correlated with PageRank, hub and authority scores, thus not confirming the idea that link analysis algorithms would predict quality more effectively than in- or out-link degree. The authors further investigated the problem and found that in- and out-degrees are predictors of more complex algorithms in the event that the set of web sites or pages contains many relevant items.

Webpage crawling is the topic of [10]. The authors wanted to investigate different crawling algorithms which download the most important webpages early during the crawl. The importance of a webpage was measured by the PageRank algorithm by using the known, complete graph that represents the Web. As an exhaustive crawling is impractical, a sample has to be crawled. Because of the incomplete knowledge of the Web graph, the algorithms tested in [10] had to use partial information about the crawled webpages and the ranking of the crawled webpages were only an approximation of the PageRank-induced ranking. The algorithms tested in the paper were based on either the number of in- or out-links or parameters of the size of the websites. The link-based algorithms were: Backlink-count crawls first the webpages with the highest in-degree, batch-pagerank crawls the webpages with the highest PageRank calculated on the basis of the graph crawled so far, a variant of Batch-pagerank called partial-pagerank, a kind of weighted back-link count called OPIC [11], and a series of variants of PageRank. The website size-based algorithms were: Larger-sites-first crawls the sites with the largest number of webpages still not crawled, and Breadth-first, which crawls in the usual breadth-first ranking. The rank correlation between the rankings were measured by τ. Although the significance analysis of the τ's were not reported in the analysis, our assessment was that all but Backlink-count algorithms are significantly ($p < 0.05$) correlated with PageRank provided that the sample size was $5,000$ as reported in their paper.

In [12] a variant of PageRank based on the aggregation of the webpages of the target graph into a kind of similarity classes was investigated. In the paper, the similarity criterion was quite simple — the classes were the sets of pages on a given host. Therefore, two types of links occur — the inter-host links connect hosts, whereas the intra-host links connect the webpages of the class. Then, the algorithm recognised the difference between the two types of link and assigned distinct probability distributions. Thus, an approximation of PageRank is obtained rather than the exact PageRank. Therefore, the approximated PageRank was computed on the basis of the same

original Web graph, yet the computational load was less than the load for the exact PageRank due to the lower number of hosts than the number of webpages. It has been then natural to compare the ranking induced by the approximated PageRank with the ranking induced by the exact PageRank. In the paper, Spearman's ρ for rank rank correlation was employed. The results report that $\rho = 0.95$.

2.2 Rank Correlation in IR Evaluation

In [13] an investigation of an alternative sampling procedure to the depth pooling method was conducted. The aim of this study was reducing the effort of judging fairly large samples and, at the same time, keeping a good accuracy of the estimation of the performance measures, e.g. average precision. At this aim, the investigation was also based on the measurement of the rank correlation between the rankings of the system produced by the depth pooling methods and by the alternative sampling method. τ was used to measure the rank correlation and high values were found. The rankings compared in this investigation referred to the same runs, that is, the runs were sampled by two distinct sampling methods. Therefore, each run was associated to two distinct performance measures. The authors of [13] claim that τ does "not measure how much the estimated values differ from the actual values." Therefore, even if it indicates "perfectly correlated estimated and actual values, the estimates may still not be accurate. Hence, it is much harder to achieve small RMS errors than to achieve high"[2] τ's. A similar investigation was reported in [14] as regards the rank correlation between pairs of system rankings. The effectiveness of one ranking of the systems was measured by the usual TREC pooling method, whereas the effectiveness of the other ranking was measured by an alternative pooling method — the latter was based on a Machine Learning algorithm. τ was employed for measuring the rank correlation and high values of τ resulted from the investigation. In both the investigations, each system or run was then subjected to a pair of measures, each measure coming from a different sampling method.

In [15] a series of experiments examining three different ways of building test collections was examined by using no system pooling. In particular, the authors investigated whether pooling is necessary to build a relevance assessment (qrel) set or not. The three ways of building qrels were the following ones. First, a collection formation technique combining manual feedback and multiple systems is adapted to work with a single retrieval system. τ was computed between a ranking produced at each iteration and the official TREC ranking. The τ's were from 0.82 to 0.93 on average — the averages were computed over different parameters, the range referred to the five iterations and the relevance assessments came from the same collection. The second method was based on pooling the output of multiple manual searches — the original method reported was [16] is re-examined. The manual searchers were simulated by the manual runs of the interactive track of TREC. The paper reports on the correlation between the ranking of the systems, which were evaluated by MAP and by type of run. The τ's were higher than 0.9 for most of the ranking pairs — this is not surprising since the relevance assessments are about the same topics. Finally, the paper examined an alternative approach where the ranked output of a single automatic search on a single retrieval system is assessed for relevance without any pooling. The automatic ad-hoc runs were ranked by MAP, and τ was computed between these runs and that produced by TREC. These correlations were quite high, too. It is worth noting that the authors of [15] pointed out

[2]RMS: Root Mean Squared errors, i.e., a distance from estimation to actual values.

that "Testing qrel sets on a large number of runs is likely to produce high correlations. The range of values when experimenting with a few run variants from a single system is likely to be smaller." thus acknowledging that there is something imprecise in the use of this rank correlation measure.

In [17] the problem of the stability of the system rankings when the set of relevance judgements changes was addressed. The pooling method of TREC states that each participating system, i.e. each submitted run is assessed on the basis the judgements provided for a sample of documents by a group of assessors. The sample of documents is drawn from the top $1,000$ documents listed in every run. The assessed runs are then ranked by MAP. The validity of the pooling method of TREC relies on the reliability of the subset of documents retrieved for assessing the runs. If the system ranking changed when a different set of judgements is used, the pooling method would not be reliable. In the paper the correlation between pairs of system rankings assessed with different set of relevance judgements was investigated. τ was used for measuring the correlation and high values resulted. Although the author pointed out that the sets of relevance judgements were not independent and that the τ's were slightly higher than the correlation that would have been resulted from completely independent sets, it was argued that the correlation indicated that the system rankings were not discordant and that the assessment of the effectiveness of the systems was stable against variations of the sets of relevance assessments. Other analyses were reported in the paper and it was concluded that "the actual value of the effectiveness measure was affected by the different conditions, but in each case the relative performance of the retrieval runs was almost always the same."

In [18] different search engines were compared by investigating their own ability of retrieving highly relevant documents and not only relevant documents — the distinction between highly relevant and relevant documents were made since TREC 1999. The investigation confirmed the idea that some systems work better than others when retrieving only highly relevant documents without searching non-highly relevant documents. The investigation was carried out by forming two system rankings. One system ranking was made by submitting some queries and evaluating the systems by MAP and precision at 10 documents using both relevant and highly relevant documents. The other system ranking was formed using highly relevant documents only. τ was used for measuring the correlation between the two system rankings. In the paper it is claimed that, since the highest correlation was noticeably smaller than the correlations found in an earlier, similar study, one can support the conclusion that different retrieval systems are better at finding the highly relevant documents than those that are better at finding generally relevant documents.

In[19] implicit feedback was addressed. In particular a study on how well implicit feedback models train themselves was measured, where an implicit feedback model was implemented as a list of terms which feed the system for query expansion. In the paper an implicit feedback model trains itself when the list of terms is similar to the list of terms ranked by probability of relevance — in other words, the latter list of terms is the ground truth which is available if explicit relevance feedback is performed by using all the relevance information. τ and Spearman's rank correlation ρ were used for measuring the correlation between list of terms.

2.3 Rank Correlation in Distributed IR

In [20, 21] rank correlation was addressed in distributed information retrieval and precisely in resource or database selection. The selection of the databases from which documents are retrieved

is a crucial step. The selection process requires a set of resource descriptions that accurately represent the contents of each database. In the papers, query sampling is proposed for gathering lists of index terms from the databases. Resource descriptions are created by running queries and examining the documents that are returned. After some documents have been retrieved, words and frequencies from the retrieved documents are returned by the database and added to the resource description until a stopping criterion has been reached. As the resource description is only a sample of the actual description, the paper investigated whether this sample is good enough. As the list of terms are ordered by frequency, the use of rank correlation measures is immediate. Spearman's rank correlation measure was employed for comparing the list of terms produced by the query-based sampling method with the actual description of each resource — some test collections were used for simulating the resources and the queries. Moreover, the lists of terms were built on the basis of the terms common to both resource descriptions. The results showed that the two lists are highly, positively correlated after 250 sampled documents. Note that both lists of terms comes from the same collection.

3 On Rank Correlation

The use of a test statistic for studying rank correlation should be accompanied by an analysis of the underlying assumptions of this use and of the properties of these tests. In fact, distinct tests lead to making different conclusions about the concordance or discordance between rankings. This fact can impact on the conclusions a researcher might draw about the effectiveness or the behaviour of a ranking algorithm — for example, one can argue that by means of one such tests two link-analysis algorithms for webpage ranking yield equivalent webpage rankings in order while they are not if another test is used.

One important feature of a test is the research hypothesis which the test is used for. When the experimenter defines a research hypothesis, the test is selected so as to compute the distribution of probability under the null hypothesis. When the distribution of probability of a test under the null hypothesis of discordance is known, the test can be used to support the research hypothesis of concordance — as said above, neither hypotheses can be disproved, yet the null hypothesis can be rejected. [3] Often, the unique known distribution of probability of a test T is the one under the null hypothesis, while the distribution of T under the research hypothesis cannot be computed. As a consequence, the experimenter should define the null hypothesis when using T and to select another test U for which the null hypothesis of U is the research hypothesis of T. Of course, U cannot be considered as the opposite of T, but two tests may provide a richer information than one test.

Suppose a user is looking at the ranked list of n webpages ranked by, for example, PageRank; suppose also that this is the reference ranking, namely, the true ranking of the n webpages.[4] Without loss of generality $\mathbf{x} = (1, 2, \ldots, n)$ is supposed to be the reference ranking placing webpage i at rank i, namely, $x_i = i$. In this way webpage i follows $i-1$ webpages and precedes $n-i$ webpages. Another ranking algorithm, for example, the PageRank algorithm stopped after k steps, is used

[3] The notion of null hypothesis is illustrated in Appendix A.

[4] "True" means, for example, that the ranking has been computed after a large number of iterations so as to make the ranking stable.

n	Value	Decision
10	-1	Discordant
20	0.53	Concordant
30	0.79	Concordant
40	0.88	Concordant
50	0.93	Concordant

(a) Kendall's τ

-

n	Value	Decision
10	0.90	Discordant
20	0.45	Discordant
30	0.27	Discordant
40	0.23	Discordant
50	0.18	Concordant

(b) Kolmogorov-Smirnov's D

Table 1: **Decision suggested by τ and D for increasing sample sizes.**

to rank these webpages and a different ranking may be computed. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be this alternative ranking such that y_i is the rank of webpage i in the alternative ranking.

3.1 Some Drawbacks with τ

τ is sometimes counter-intuitive to measure the correlation between two rankings. A reason why τ may appears as inadequate can be seen by the following example. Let us suppose that $\mathbf{y} = (2, 1, 3, 4, 5, 6, 7, 8, 9, 10)$ is the ranking of $n = 10$ webpages presented to the end user, while $\mathbf{x}$ is the reference ranking. At first sight $\mathbf{y}$ is very similar to $\mathbf{x}$ although only the two top ranked webpages swapped. The end user would not appreciate the difference because the top ranked webpage in $\mathbf{x}$ is very close to the top in $\mathbf{y}$. Indeed $\tau = 0.96$. Let us now suppose that $\mathbf{y}' = (7, 2, 3, 4, 5, 6, 1, 8, 9, 10)$ is the ranking presented to the end user. At first sight $\mathbf{y}'$ is less similar to $\mathbf{x}$ than $\mathbf{y}$ because the top ranked webpage in $\mathbf{x}$ has exchanged with the seventh ranked webpage. The end user would therefore note the difference and webpage 1 would be penalised. A test should lead to a different decision from concordance. However, Kendall's $\tau = 0.51$ which leads to again decide for concordance when $\alpha = 0.05$.

Another inadequacy of τ is due to the sample size. If the sample is not small, the null hypothesis is almost always rejected even though the number of discordant pairs is not very large. This inadequacy can be seen by means of the following example. Let $\mathbf{y}' = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1, \ldots, n)$ is the alternative ranking of n objects such that $y'_i = i, i = 11, \ldots, n$, whereas $\mathbf{x}'$ is the reference ranking such that $x_i = i, i = 1, \ldots, n$. The number of discordant pairs of objects is $n_d = 45$, the number of concordant pairs is $n_c = n_d - 45$. Table 3.1 reports the decision as n varies with $\alpha = 0.05$ when τ is used, while Table 3.1 reports the decision when D is used. This table shows that τ rapidly leads to decide for concordance, that is, reject H_0, as n increases. In fact this test leads to support the research hypothesis that the two rankings are concordant when $n = 20$ even though the top ten objects are inversely correlated and the bottom ten objects are equally ranked. This behaviour is due to the quadratic factor in the denominator of τ which quickly tends to 1 when n increases. This table shows that Kolmogorov-Smirnov's D decreases more slowly than τ increases and the two rankings are still considered as discordant when $n = 40$. They are considered as concordant when $n = 50$ because D is too small to reject the null hypothesis of concordance.

3.2 Kendall's τ and Kolmogorov-Smirnov's D

τ is defined upon the null hypothesis that the two rankings are discordant. Moreover, τ very rapidly tends to accept concordance if the sample size increases without considering whether the increase of the sample size affects the objects with different ranks or not. Furthermore, τ leads to deciding for concordance yet two top but "distant" ranked objects swap, which is rather counter-intuitive. When the rankings are naturally concordant — for instance in the event that PageRank and a similar link-analysis algorithm are studied — τ tends to confirm this natural concordance and is useful for making the experimenter sure that the concordance is not due to chance. It would be very surprising if this test suggested that the two naturally concordant rankings are discordant.

As Kolmogorov-Smirnov's D is on the contrary defined upon the null hypothesis that the two rankings are concordant, the null hypothesis is rejected with small probability of error and the research hypothesis of discordance is supported when enough evidence against it has been observed.[5] As D is less affected by the increase of the sample size than τ, concordance is not rejected while the sample is not large. When the rankings are not naturally concordant — for instance in the event that two radically different retrieval algorithm are studied — D tends to confirm discordance.

At the light of the differences between τ and D, some suggestions can be made about the investigation of the correlation between two rankings. When the rankings are naturally concordant, τ can be used to confirm this concordance — this is a sort of sanity check. However, if the correlation is imperfect and there are some exchanges between two objects, it may be possible that the distance between the ranks as measured by D is due to chance. If this is the case, D would permit to accept the null hypothesis of concordance thus supporting the research hypothesis of τ. When the distance between the ranks, namely, the value of D, is on the contrary so large that is significantly higher than zero, the hypothesis of concordance should be rejected thus contrasting the outcome of τ.

When τ and Kolmogorov-Smirnov's D yield contrasting outcomes, a great deal of attention should be paid to the interpretation. The use of two tests should not be avoided with the purpose of not falling into contradiction. On the contrary, two tests do allow the experimenter for better investigating the correlation than one test only. The seeming contradiction can be solved when one considers the different role played by each test, namely, the different hypothesis tested by each test. Kendall's τ tests whether $\mathbf{y}$ is one of the random permutation of $\mathbf{x}$, or is $\mathbf{y}$ a particular ranking denoting the same ordering as $\mathbf{x}$'s. Thus, τ is a measure of disorder. On the other hand, Kolmogorov-Smirnov's D detects the occurrence of outliers, namely, pairs of objects which are different ranked to a large extent.

The use of one test or another depends on the interpretation of concordance discordance the experimenter gives about two rankings. In the context of webpage ranking, for instance, the ranking which places a webpage at the seventh rank instead of at the top rank may be considered discordant from the reference ranking. The same ranking may be considered as concordant to the reference one in the context of, say, system ranking. Therefore, there is not any golden rule — testing should be performed by taking different perspectives into account.

When the rankings are fairly large, Kolmogorov-Smirnov's D seems to be more adequate than τ because the denominator of D is n. In fact concordance is decided when there are a significant

[5] An illustration of this statistic is in Appendix D.

number of exchanges only if the sample size is rather large. D should also be preferred when the decision should be sensitive to the extent of differences in order and not only the number of exchanges have to be considered — this is most important when analysing the correlation at the top ranked objects.

4 Webpage Ranking Correlation

In [3] an investigation analysed how the stationary webpage ranking induced by PageRank is correlated to the rankings induced by PageRank either stopped after k steps or tuned by the damping factor d — the higher d, the more the ranking depends on the graph topology. It was found that τ increases as k does and the increment is as more rapid as d tends to 0. The other outcome was that (1) τ got very close to 1 once the ranking size was larger than a few dozens webpages, and that (2) the sub-rankings showed smaller correlations even though the total number of exchanges was the same as the one counted in the whole ranking. Therefore, correlation depends on the ranking size.

In this section, the correlation between PageRank-induced rankings is further investigated and some correlation measures between $\mathbf{x}$ and the ranking computed after $k = 5$ steps and with $d = 0.8$, say, $\mathbf{y}$, are computed and reported in Table 4; correlations are computed after every ten webpages up to 100 webpages. The WT10g test collections was used for this investigation.

Let us consider the second sub-ranking, namely, the webpages from the 11th to the 20th. The value of τ is statistically significant at $\alpha = 0.05$ thus suggesting that the sub-ranking is correlated to the corresponding stationary sub-ranking and that the hypothesis of discordance should be rejected in favour of the hypothesis of concordance. However, D is statistically significant too since is greater than the quantile at α, that is, 0.409 thus suggesting that the sub-ranking differs from the corresponding stationary sub-ranking and that the hypothesis of concordance should be rejected in favour of the hypothesis of discordance.

While ρ is in line with τ, W^2 signals that the overall distance between the rankings is rather small, thus seemingly contradicting D.[6] This means that the two sub-rankings are correlated, even though the ranks of one or two outliers webpages are differs from the ranks of the webpages in reference ranking, as D suggests. As regards the other sub-rankings, W^2 is often significantly different from zero thus suggesting discordance.

Table 4 shows this phenomenon for the two first 30-webpages sub-rankings. The fact that the discrepancy between the outcomes suggested by τ and D occurs at the top ranked webpages should put on guard from concluding anything about the correlation between the rankings induced by two versions of PageRank. While the top ten or thirty webpages are "on average" in the same order, as measured by τ, one can assert that there is *at least* one webpage which is ranked differently with a probability of error less than 5%.

The discrepancy between the decision suggested by τ and that suggested by D is an indicator of the general behaviour of PageRank. Let us first consider τ and the role played as a measure of disorder. The tendency to concordance as the size of the rankings increases means that the ranking of all the webpages, i.e. for n equal to the total number of webpages, should be considered in the final order except for local exchanges, i.e. exchanges between a webpage and its neighbours

[6] W^2 is described in Appendix D.

Ranks	τ	D	W^2	ρ
1–10	0.20	0.60	0.61	0.28
11–20	0.60	0.50	0.26	0.69
21–30	0.33	0.70	0.46	0.46
31–40	0.02	0.90	0.84	0.01
41–50	-0.42	0.80	1.36	-0.60
51–60	0.29	0.60	0.54	0.36
61–70	-0.02	0.70	0.81	0.05
71–80	0.47	0.40	0.28	0.67
81–90	-0.24	0.70	1.24	-0.46
91–100	0.20	0.80	0.67	0.21
1–100	0.72	0.43	0.87	0.90

(a)Measures of correlation between ten-pages sub-rankings.

Ranks	τ	D	W^2	ρ
1–30	0.66	0.47	0.41	0.84
31–60	0.43	0.70	1.05	0.58
61–90	0.21	0.80	1.86	0.26
91–120	0.14	0.73	2.10	0.16
121–150	-0.10	0.93	2.85	-0.14
151–180	-0.12	0.93	2.96	-0.18
181–210	0.29	0.77	1.61	0.36
211–240	0.21	0.67	1.84	0.27
241–270	0.16	0.83	2.02	0.19
271–300	0.35	0.70	1.27	0.49
1–300	0.73	0.70	3.11	0.88

(b)Measures of correlation between thirty-pages sub-rankings.

Table 2: **Webpage ranking correlation results — the last line reports the correlation measures computed between the complete rankings).**

in the ranking — the last line of both tables is an indicator of that. This signifies that, when a ranking is compared to the reference ranking and τ is close to 1 after a few iterations, the webpages ranked on the top will be still on the top, and the webpages ranked on the bottom will be still on the bottom after a few iterations are performed. If some variations in order are observed, these are concentrated within small sub-rankings. In other words, the order induced by PageRank is greatly determined by the global structure of the graph which represents the sample of the Web studied. Let us now consider D. For many sub-rankings, in particular for those in Table 4, the value is significant thus suggesting that the disorder indicated by τ may be due to exchanges between distant webpages in the sub-ranking and not only with the closest neighbours in the sub-ranking. This is confirmed when the values of W^2 are high, otherwise the disorder is due to outliers. In summary, PageRank fixes the order of the webpages at a "global" level after a few iterations.

5 Conclusions

On the basis of the general reflections on rank correlation measures and the empirical investigation of PageRank-induced webpage ranking correlation, some general conclusions can be drawn. First, one aspect is related to the sample size. When τ is used, one should consider the fact that the null hypothesis of discordance will be likely rejected when the sample size is large. Because the samples are often large in IR, it is not surprising that the use of τ often support concordance. An example is PageRank-induced webpage ranking, yet other examples can be observed in Distributed IR or System Evaluation. Second, the nature of correlation should be analysed: If the two rankings are drawn from two populations which are naturally correlated, it is not surprising that a test statistic rejects the hypothesis of discordance. For example, if the experimenter wants to compare two link analysis-based webpage rankings produced by two algorithms which are both based on in-degree, the rankings will be likely correlated. In these cases, a test statistic such as τ or ρ is useful for confirming the natural correlation, while what should be investigated is the type of discordance which may occur between some sub-rankings. Finally, a great deal of attention should be paid to the null hypothesis on which the decision about the correlation depends. A possible interpretation is that the null hypothesis should state the default outcome of the test, whereas the research hypothesis should state the surprising outcome. For example, two radically different webpage ranking algorithms should produce discordant rankings — this is the default outcome and the null hypothesis should state the discordance. Therefore, the research hypothesis would state the concordance between the two rankings, which is the surprising outcome of the test. A test statistic would tend to not reject the null hypothesis of discordance, but, when it did, the support of the research hypothesis of concordance would be surprising and would provide a lot of information. In contrast, a test statistic which confirms a natural correlation, and then reject the null hypothesis, provides little information.

References

[1] D.A. Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 329–338, Pittsburgh, PA, USA, 1993.

[2] W.J. Conover. *Practical Nonparametric Statistics*. Wiley, 1999.

[3] M. Melucci and L. Pretto. PageRank: When order changes. In *Proceedings of the European Conference on Information Retrieval Research (ECIR)*, LNCS. Springer, 2007.

[4] A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas. Link analysis ranking: Algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, 5(1):231–297, February 2005.

[5] B. Amento, L. Terveen, and W. Hill. Does authority mean quality? Predicting expert quality ratings of web documents. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 296–303, 2000.

[6] K. Sparck Jones. *Automatic keyword classification*. Butterworths, 1971.

[7] H. Borko. Inter-indexer consistency. In *7th Cranfield Conference*, 1979.

[8] C.W. Cleverdon. The significance of the Cranfield tests on index language. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 3–12, Chicago, MC, October 1991.

[9] D. Ellis, J. Furner-Hines, and P. Willett. On the creation of hypertext links in full-text documents: Measurement of inter-linker consistency. *Journal of Documentation*, 50(2):67–98, 1994.

[10] R.A. Baeza-Yates, C. Castillo, M. Marín, and A. Rodríguez. Crawling a country: better strategies than breadth-first for web page ordering. In *Proceedings of the World Wide Web Conference*, pages 864–872, 2005.

[11] S. Abiteboul, M. Preda, and G. Cobena. Adaptive on-line page importance computation. In *Proceedings of the World Wide Web Conference*, pages 280–290, 2003.

[12] A. Z. Broder, R. Lempel, F. Maghoul, and J.O. Pedersen. Efficient pagerank approximation via graph aggregation. *Journal of Information Retrieval*, 9(2):123–138, 2006.

[13] J.A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 541–547, 2006.

[14] J.A. Aslam, V. Pavlu, and R. Savell. A unified model for metasearch, pooling, and system evaluation. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pages 3–8, 2003.

[15] M. Sanderson and H. Joho. Forming test collections with no system pooling. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 25–29, 2004.

[16] G.V. Cormack, C.R. Palmer, and C.L.A. Clarke. Efficient construction of large test collections. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 282–289, Melbourne, Australia, August 1998. ACM Press, New York.

[17] E.M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 315–323, Melbourne, Australia, August 1998. ACM Press, New York.

[18] E.M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 74–82, 2001.

[19] R.W. White, I. Ruthven, J.M. Jose, and C.J. van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3):325–361, 2005.

[20] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19(2):97–130, 2001.

[21] J. Caverlee, L. Liu, and J. Bae. Distributed query sampling: a quality-conscious approach. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 340–347, 2006.

[22] R.A. Fisher. *The design of experiments*. Hafner Publishing Company, 1971.

[23] M. Kendall. *Rank Correlation Methods*. Charles Griffin & Co. Ltd., fourth edition, 1975.

[24] W.R. Knight. A computer method for calculating Kendall's τ with ungrouped data. *Journal of American Statistical Association*, 61(314):436–439, 1966.

[25] A.N. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933. In Italian. The English title is "On the empirical definition of a distribution function".

[26] T.W. Anderson. On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159, 1962.

A The Null Hypothesis

When an experimenter has to decide about the similarity of two rankings, he has first to define the research hypothesis, namely, H_1. In the context of the analysis of rank correlation in IR, H_1 states either the discordance or the concordance between two rankings, that is, the objects of one ranking either tend or do not tend to be ranked in the same order as the objects of the other ranking. H_0 is the negation of H_1 and is stated with the aim of rejecting it. If H_1 states, for example, that two judges has different favourite movies, H_0 states that the two judges rank their favourite movies in the same order.

To support the research hypothesis, the statistical science dictates that the null hypothesis, namely, H_0, has to be falsified or rejected. Therefore, H_1 is not proved directly but it is supported indirectly by the rejection of H_0. In other words, concordance cannot be proved directly, but is indirectly supported when rejecting discordance. The reason is that concordance cannot be proved by the enumeration of all the events that two rankings are concordant, but can be supported by the occurrence of only one event so that two rankings cannot be considered as discordant. Of course, the converse holds too — H_0 is never proved, but is possibly disproved. This principle of falsifiability is due to R.A. Fisher [22].

Following the principle of falsifiability, what is important is that the distribution of probability related to the test statistic has to be computed under H_0. Therefore the decision taken about the rejection or acceptance of the null hypothesis, and therefore the computation of the error probabilities, are based on the distribution of probability under the null hypothesis. As a consequence, when the experimental outcome suggests to reject H_0 with low error probability under the null hypothesis, it is not implied that H_1 is accepted with low error probability. To accept H_1 with low error probability, the probability distribution under H_1 should be computable, but it is often unknown.

A test statistic is defined for investigating one aspect of rank correlation, while another test may help highlight other aspects. Therefore, in the context of rank correlation in IR, when a test suggests to reject the null hypothesis of discordance between two rankings when error probabilities are computed under the null hypothesis, *another test* may suggest to reject the hypothesis of concordance between two rankings when error probabilities are computed under the research hypothesis, and yet another test may suggest to reject the hypothesis of concordance under the null hypothesis. What is important is that each test has different limitations and potentialities to which a great deal of attention should be paid.

B Kendall's τ

A test statistic named τ was introduced in [23] and often used for comparing two rankings. τ is as follows. n objects are ranked as $\mathbf{x} = (1, 2, \ldots, n)$, that is, $x_i = i$, and then are assigned the y_i's. A pair of objects i, j is concordant if they are in the same order in both rankings, otherwise the pair is discordant. Let n_c, n_d be the number of concordant pairs and the number of discordant pairs, respectively. The quantity $S = n_c - n_d$ reaches its maximum when the two rankings completely agree, i.e. when $S = N = \frac{n(n-1)}{2}$ and reaches its minimum when the two rankings completely disagree, i.e. when $S = -N$. τ is defined as

$$\tau = \frac{S}{N} \ . \tag{1}$$

When, for example, $n = 10$ objects are ranked as $\mathbf{x} = (1, 2, \ldots, 10)$ and then are assigned $\mathbf{y} = (2, 3, 1, 4, 5, 7, 6, 8, 9, 10)$, then $\tau = 0.87$, which denotes concordance. If, on the contrary, the objects are assigned $\mathbf{y} = (10, 9, 8, 1, 6, 5, 4, 2, 7, 3)$, then $\tau = -0.51$, which denotes a negative correspondence. If τ is near to zero, then the two rankings are considered as discordant, if τ is near to 1 (-1) the two rankings are considered as positively (negatively) correlated. Knight proposed in [24] an $O(n \log n)$ algorithm based on the Merge Sort algorithm to compute S.

As the distribution of probability of τ is known only under the null hypothesis of discordance, τ can be used for testing that $\mathbf{x}$ and $\mathbf{y}$ are discordant against the research hypothesis the two rankings tend to be concordant. For small n's the exact probability distribution of τ under the null hypothesis can be computed. As n gets larger, τ converges to a Normal random variable with mean 0 and variance dominated by $\frac{2}{9}\frac{n+\frac{5}{2}}{N}$ [23]. The null hypothesis is rejected when τ is greater than the $1-\alpha$ quantile in the null distribution, where α is the significance level of the test, namely, the maximum probability of rejecting a true null hypothesis. For example, when $\tau = 0.87$, $n = 10$ and $\alpha = 0.05$, then the null hypothesis is rejected because τ is greater than the 0.95 quantile in the null distribution.

C Spearman's ρ

Spearman's rank correlation measure called ρ is another test statistic which is often used instead of τ.

$$\rho = \frac{\sum_{i=1}^{n} \bar{x}_i^2 + \sum_{i=1}^{n} \bar{y}_i^2 - \sum_{i=1}^{n} (x_i - y_i)^2}{2\sqrt{\sum_{i=1}^{n} \bar{x}_i^2 \sum_{i=1}^{n} \bar{y}_i^2}} \tag{2}$$

where $\bar{z}_i = z_i - \bar{z}$ and $\bar{z}$ is mean rank in the ranking $\mathbf{z}$. These test statistics are related by $-1 \leq 3\tau - 2\rho \leq +1$. As both employ all the sample data, they tend to reject the null hypothesis of discordance at the same significance level when the *null* hypothesis holds. However, they may give different information about the concordance when the *research* hypothesis, namely, concordance, holds. It is also useful noting that Eq. 2 yields correct values even when ties are present.

D Kolmogorov-Smirnov's D

Another test of the correlation or concordance between two rankings is described in the following. At this aim let us consider the following function:

$$F_{\mathbf{z}}(i) \quad = \quad \frac{\#\text{ objects not following } i \text{ in ranking } \mathbf{z}}{n} \ .$$

When $\mathbf{x} = (1, 2, \ldots, n)$, it can be easily seen that $F_{\mathbf{x}}(i) = \frac{i}{n}$ because object i occupies rank i, therefore $i - 1$ is the number of objects preceding it. If, for example, $\mathbf{y} = (2, 1, 3, 4, 5, 6, 7, 8, 9, 10)$ is the alternative ranking, the rank of object 1 is 2, the rank of object 2 is 1, and all the other ranks are equal to those of $\mathbf{x}$. As a consequence $F_{\mathbf{y}}(1) = \frac{2}{10}, F_{\mathbf{y}}(2) = \frac{1}{10}$ and $F_{\mathbf{y}}(i) = \frac{i}{10}$ for all $i > 2$

One can see that $F_{\mathbf{x}}$ looks like a uniform distribution function. This analogy leads to consider the Kolmogorov-Smirnov's goodness-of-fit test. This test was introduced by Kolmogorov [25] to see when two random samples are governed by the same unknown distribution or when a random sample is from a specified distribution. That is, the null distribution specifies some distribution function F and the empirical distribution of the random sample is compared to F.

In the event of two rankings one logical way of comparing two rankings is by means of the "uniform" distribution induced by $\mathbf{x}$ and the empirical distribution induced by $\mathbf{y}$, that is, by calculating the number of objects not following i in $\mathbf{y}$ for every i. Kolmogorov-Smirnov's test is defined as

$$D = \max_i |F_{\mathbf{x}}(i) - F_{\mathbf{y}}(i)| \tag{3}$$

which is the maximum vertical distance between the graphs corresponding to the distributions. If, for example, $\mathbf{y} = (2, 1, 3, 4, 5, 6, 7, 8, 9, 10)$ is the alternative ranking, then $D = \frac{1}{10}$.

D measures the agreement between the two distributions — in our context, it measures the concordance between the two rankings. The null hypothesis of D is concordance and therefore the research hypothesis is discordance. In this sense, Kolmogorov-Smirnov's D is the opposite of τ, that is, if there is not good agreement between $F_{\mathbf{x}}$ and $F_{\mathbf{y}}$, namely, the empirical distribution of $\mathbf{y}$ is different from the "uniform" distribution induced by $\mathbf{x}$, then the null hypothesis can be rejected and the research hypothesis of discordance can be supported.

The exact probability distribution of D can be computed under the null hypothesis of concordance and for small values of n. For n large, the distribution can be approximated [2]. When α is the significance level of the test, the null hypothesis is rejected when D is greater than the $1 - \alpha$ quantile in the null distribution. For example, $D = \frac{1}{10}$ is less than the 0.95 quantile and therefore the null hypothesis — the rankings are concordant — is accepted when $\alpha = 0.05$.

Let us suppose $\mathbf{y} = (7, 2, 3, 4, 5, 6, 1, 8, 9, 10)$ be the alternative ranking, then $D = \frac{6}{10}$. In fact the number of pages not following object 1 is 7, therefore, $F_{\mathbf{y}}(1) = \frac{7}{10}$ and the number of pages not following 7 is 1, therefore, $F_{\mathbf{y}}(7) = \frac{1}{10}$, whereas $F_{\mathbf{y}}(i) = \frac{i}{10}$ for the other i's. Because $F_{\mathbf{x}}(1) = \frac{1}{10}$ and $F_{\mathbf{x}}(7) = \frac{7}{10}$, it follows that $D = \left|\frac{7}{10} - \frac{1}{10}\right| = \frac{6}{10}$. When $\alpha = 0.05$, the 0.95 quantile is 0.409, the null hypothesis is rejected, and therefore the hypothesis that the two rankings are discordant can be supported.

One property of D is that the hypothesis of concordance is rejected if one pair of objects swap their rank. In some contexts, this property may be little appropriate when the experimenter wants to assess the overall deviation from the reference ranking. A test statistic which uses the information from all, and not just the largest, deviations is the Cramér-von Mises goodness-of-fit test statistic [2, 26], which is defined as:

$$W^2 = \frac{1}{2} \sum_{i=1}^{n} \left(F_{\mathbf{y}}(i) - F_{\mathbf{x}}(i)\right)^2 \ . \tag{4}$$

The quantiles can be found in [26].