Query Variations and their Effect on Comparing Information Retrieval Systems

Guido Zuccon
Queensland University of Technology
Queensland, Australia
g.zuccon@qut.edu.au

Joao Palotti
Vienna University of Technology
Austria
palotti@ifs.tuwien.ac.at

Allan Hanbury
Vienna University of Technology
Austria
hanbury@ifs.tuwien.ac.at

ABSTRACT

We explore the implications of using query variations for evaluating information retrieval systems and how these variations should be exploited to compare system effectiveness. Current evaluation approaches consider the availability of a set of topics (information needs), and only one expression of each topic in the form of a query is used for evaluation and system comparison. While there is strong evidence that considering query variations better models the usage of retrieval systems and accounts for the important user aspect of user variability, it is unclear how to best exploit query variations for evaluating and comparing information retrieval systems.

We propose a framework for evaluating retrieval systems that explicitly takes into account query variations. The framework considers both the system mean effectiveness and its variance over query variations and topics, as opposed to current approaches that only consider the mean across topics or perform a topic-focused analysis of variance across systems. Furthermore, the framework extends current evaluation practice by encoding: (1) user tolerance to effectiveness variations, (2) the popularity of different query variations, and (3) the relative importance of individual topics. These extensions and our findings make information retrieval comparisons more aligned with user behaviour.

1. INTRODUCTION

The current practice in Information Retrieval (IR) evaluation is based on the computation of evaluation metrics (e.g., average precision, cumulated gain, etc.) across a set of topics or information needs[1] and systems are discriminated with respect to the mean effectiveness achieved over the topic set. Recent research has shown that for the same information need, different users pose different queries; these in turn achieve different effectiveness in terms of retrieval measures [2, 13]. Put another way, for the same information need, the quality of a search engine (as established through a traditional evaluation measure) differs across the queries users would issue for that information need (query variations), and thus the search engine effectiveness for an information need as measured by an IR evaluation measure comes with a variance due to the different queries users would pose for that need. However, the current evaluation methodology does not take into account this variance.

Recently, Bailey et al. have argued for the use of query variations in IR evaluation [2]. Previous work in the TREC 8 Query Track also had investigated the use of query variations expressing the same information need for evaluating retrieval systems [5]. The findings of that track was that users posed extremely variable queries (both in essence and effectiveness) and that query variations were as big of a source of variance as system variations. While in the past this example of how query variations could be employed within a formal, Cranfield-like evaluation has not been followed by a significant uptake in practice, recent proposals have put forward new test collections for evaluating IR systems in presence of query variations, for example [16, 11, 3].

A key open issue is how query variations should be used within the evaluation framework and how measurements for different queries and different topics should be combined. For example, in TREC 8, query variations and topics are merged together to compute mean average precision. Thus, system A is deemed better than system B if on average (over all query variations and information needs) A has a higher evaluation measure score than B, regardless of (among other aspects): (i) the popularity across the user population of specific query variations over others, (ii) the stability across different query variations for the same information need. Alternatively, query variations are examined for each topic separately by performing a within-topic (and across systems) ANOVA. This approach unveils whether queries, systems, or both are responsible for effectiveness variations, but cannot be used to draw comparative conclusions about systems. Thus, the lack of understanding regarding how query variations should be used for evaluation poses the risk that the recently proposed new test collections that include query variations [16, 11, 3] lead to inconclusive findings.

In this paper, we aim to address this gap by providing an approach for evaluating and comparing systems over query variations. Our hypotheses are that (1) systems effectiveness depends on the quality of the queries as well as that of systems, and (2) some systems have more stable effectiveness than others (stable systems). Thus, this stability should be accounted for in the evaluation. The intuition is that, in

[1] In the following we use information needs and topics interchangeably, referring to topics as the formulation of users' information needs expressed in the typical TREC topic description fields.

CIKM'16 , October 24 - 28, 2016, Indianapolis, IN, USA

ISBN 978-1-4503-4073-1/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2983323.2983723

different circumstances, users may prefer a more stable but less effective system over a more effective but less stable one, or vice versa. For example, for leisure-related queries (say, searching for a good restaurant) an unstable system may be sufficient, while within enterprise settings, a stable system, even if less effective on average, may be more appropriate.

We use the framework of mean variance analysis, also known as Markowitz's Portfolio Theory [12], as a methodology for evaluating systems both within and across topics. Under our mean variance evaluation (MVE), users express a level of risk they wish to tolerate and this is associated to the variability of system effectiveness. The analysis then ranks systems according to their mean effectiveness across queries as well as their stability, by balancing these properties with respect to the risk the user wishes to tolerate and the popularity or importance of the topic or query variation.

Through the investigation of the proposed evaluation framework across exemplary TREC and CLEF evaluation tasks, we show how this framework can be applied in practice and how it differs from the common approach of considering the mean effectiveness, rather than mean and variance as we propose, when evaluating and comparing IR systems.

2. RELATED WORK

Our evaluation approach is based on the general framework of mean variance analysis (MVA), which is well known in finance, where is often referred to as Markowitz's Portfolio Theory [12] and is used to identifying the optimal allocation of wealth to a portfolio of shares (see Section 3). In IR, the MVA framework has been previously applied as system-side approach to ranking. Wang et al. [26, 27] have used MVA as a criteria for document ranking diversification, where MVA's risk parameter (see Section 3) is used to control the level of diversity required within the ranking, and variance is estimated using cosine similarity. Zuccon et al. [33] have used MVA for rank fusion. Given a query, point-wise relevance estimations for a document, obtained through the systems considered for fusion, are treated as a distribution and its associated mean and variance are computed. Ranking fusion is then performed using MVA.

The analysis of variance among observed measurements is fundamental to statistical tools like ANOVA, which is widely used in IR to determine the portion of effectiveness variation that can be explained by different factors, e.g., variations attributable to systems or topics. Note however that, while the current IR evaluation practice does analyse variance across topics (e.g., when determining statistical significance [22]), it does not separately handle query variations and topics.

Despite the limited amount of evaluations that have considered multiple queries for the same information need, there has been a recent trend in devising collections that include this aspect, e.g. [5, 16, 11, 3, 15].

The TREC 8 Query Track [5] was explicitly aimed at investigating the issue of query variability. They shared our same definition of a topic, i.e., an information need of a user, and they considered many instantiations of each topic in the form of queries. The methodology used in that track for evaluation and system comparison is based on comparison of means across queries and topics, with no explicit exploitation of variability within topics due to query variations: a limitation in the evaluation methodology that we aim to address in our work.

The CLEF 2015 eHealth Lab Task 2 [16] considered query variations in the context of health consumer search. Their further work studied the relevance assessments across many aspects, including query variation [15]. They reported that system ranking correlation was not stable across query variations. As in TREC 8, the evaluation methodology did not differentiate across query variations and topics. In this paper, we use both collections, comparing the findings about systems effectiveness obtained by considering only the mean effectiveness across query variations and topics with that obtained using the proposed MVE approach.

Bailey et al. [2] also examined query variability. They observed that the common test collection evaluation approach eliminates sources of user variability. By empirically demonstrating that query formulation is critical to query effectiveness, they argued that test collection design would be improved by the use of multiple query variations per topics. This argument was further supported by analysing the implications of query variations on pooling [13]. However, while that analysis highlights that the effect of query variations is as strong as that of system variations, no methodology is proposed to explicitly differentiate between query variations and topic variations when comparing systems.

The recent collections assembled by Bailey et al. [3] and Koopman and Zuccon [11] explicitly consider the availability of query variations; however, both lack of an evaluation methodology that explicitly accounts for and differentiates systems across such variations.

Although not aimed at exploring query variations, a number of other TREC tasks are worth mention because of peripheral commonalities with the problem we are examining.

In the TREC Robust Track, systems were compared by means of geometric mean average precision (GMAP) [18]. GMAP uses geometric mean, rather than arithmetic mean, and rewards the maximisation of average improvements on the most difficult topics (smaller values of AP), thus effectively discriminating some topics over others (i.e., improvements on hard topics more important than on easy topics). Biasing evaluation with respect to some topics or query variations is naturally embedded in the proposed MVE framework, as the experimenter can set the amount of importance a query or topic has over others (akin to assigning different amounts of wealth to shares within the financial Portfolio Theory).

In the TREC 2014 Web Track [8] systems are evaluated using risk-aware evaluation criteria that focus on the absolute difference in effectiveness between a ranking and a baseline for a given query. Systems are compared by counting the number of wins (queries for which a system outperformed a baseline) and the number of losses with respect to the common baseline. The risk-sensitive utility measure (U_{RISK}) is defined as the number of wins weighted by the number of losses and a risk aversion parameter (large values reward more conservative systems which avoid large losses). A more sophisticated variant of U_{RISK} has been proposed which is theoretically grounded in statistical hypothesis testing (T_{RISK}) [10]. This approach is akin to the risk-propensity parameter present in the MVE framework, where there is explicit control of the risk users are willing to take to have effectiveness higher than the mean at the expense of higher effectiveness variability.

Finally, the diversity task of the TREC Web Track in 2009-2014 had settings diametrically opposite to those we consider here. In that task a (ambiguous and underspeci-

fied) query is associated to a number of specific information needs (subtopics), while here we consider the opposite case: multiple queries referring to the same information need.

Our proposal attempts to provide a method to compare retrieval systems across queries (and information needs) alternative to the mere comparison of their mean effectiveness. In this space, prior work by Dincer [9] has used principal components analysis to compare systems effectiveness, which highlights topics interrelations among systems. The work of Zhang et al. [31, 30] shares similar ideas with our MVE framework, however they consider evaluation in ad-hoc settings (no query variations) and apply variance analysis to the error measured with respect to an "ideal" system. In addition, their method assumes the use of average precision and does not account for the user preference towards stability of effectiveness peaks. Nevertheless, we agree with their observation that, as research advances have pushed the effectiveness bounds of IR systems through the years, "the stability [of systems] has become problematic and could have been largely overlooked".

3. MEAN VARIANCE ANALYSIS

We introduce the idea of mean variance evaluation by drawing from a parallel in finance. In finance, investors are often offered a choice of how to invest their wealth across a portfolio of N shares or investment options, i.e., options $1, \ldots, N$. Different portfolios may be proposed to investors. Portfolios are characterised by different levels of wealth investment w to be assigned to each share, i.e., a portfolio may assign a portion of wealth w_i to investment i and a different portion w_j to investment j, where $\sum_{i=1}^{N} w_i = 1$. Each wealth investment is characterised by a rate of return R_i, which is a random variable with a probability distribution over all estimation points (measurements) q_k^i (with $k \in 1, \ldots, M$) such that the return r_k^i is achieved with probability $p(r_k^i)$. Given a specific portfolio P, the portfolio rate of return can be computed as

$$R_P = \sum_{i=1}^{N} w_i R_i \tag{1}$$

The portfolio rate of return has two moments: the mean value μ_P, which represents the expected return of the portfolio, and the variance σ_P^2, which represents the risk associated with the portfolio. Formally, the mean value of portfolio P is calculated as:

$$\mu_P = E[R_P] = \sum_{i=1}^{N} E[w_i R_i] = \sum_{i=1}^{N} w_i \cdot \Big(\sum_{k=1}^{M} r_k^i p(r_k^i) \Big) \tag{2}$$

Similarly, the variance of the portfolio[2] is computed as

$$\sigma_P^2 = var(R_P) = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j cov(R_i, R_j) \tag{3}$$

$$= \sum_{i=j=1}^{N} w_i^2 var(R_i) + \sum_{i \neq j}^{N} w_i w_j cov(R_i, R_j) \tag{4}$$

In finance, Markowitz's Portfolio Theory analyses mean and variance of the rate of return associated to shares in the possible portfolios and determines which portfolio delivers optimal returns. The value of a portfolio (O_P) is:

$$O_P = \mu_P - \alpha \sigma_P^2 = E[R_P] - \alpha var(R_P) \tag{5}$$

[2] Recall that the variance of the sum of N random variables is the sum of the covariances of each pair of random variables.

The optimal portfolio is the one that maximises equation 5. Parameter α represents risk preference[3]: positive values indicate risk aversion (as a large variance would decrease the overall attractiveness of the portfolio), while negative values indicate risk propensity (as a large variance would increase the overall attractiveness of the portfolio).

In finance, portfolio optimisation is typically performed to guide the wealth of investment w that should be allocated to each share, once the risk preference of an investor has been determined. In the evaluation of IR systems, as we shall see next, we consider the wealth of investment as a fixed (or given a priori) quantity, and the objective of applying this type of analysis is not to determine which configuration of wealth of investment should be preferred, but to determine which system, among a number of possible choices, best supports the search requirements of a user (or a user population). In these settings, different systems provide different rate of returns for the same share q_k^i.

4. MEAN VARIANCE EVALUATION

Next, we apply the ideas of mean variance analysis to the evaluation of IR systems, thus developing a mean variance evaluation (MVE) framework for IR. We first consider a general case of the framework (Section 4.1), where variance is due to: (1) multiple topics being considered for evaluation (inter-topics variance), and (2) for each topic, multiple queries are issued to the system (intra-topic variance). These variances correspond to two distinguished sources of variance in system effectiveness: the topics (inter-topics variance) and the user (intra-topic variance). We then further develop two special cases: when multiple queries per topic are present, but only one topic is used (only intra-topic variance is considered – Section 4.2), and when only one query per topic is present, but multiple topics are used (only inter-topic variance is considered – Section 4.3). This latter case corresponds to the common practice in IR evaluation.

4.1 General Evaluation Framework

We first start by remarking that we consider a topic (or information need) as being a high level requirement for information users have on a specific matter. This for example may correspond to a TREC topic such as the TREC 8 Query Track topic 51 "Airbus Subsidies". When a user has one such information need, they may formulate one or more queries to search for relevant information associated with the topic, e.g., q_1^{51} ="recent airbus issues" and q_2^{51} ="Airbus subsidy dispute". Similarly, different users may generate different queries for the same topic: this was the case for some of the works reviewed in Section 2 [2, 5, 16]. In summary, a topic is represented by a number of M queries; and the queries observed by a search engine are generated from a set of N (N may be infinite) topics. Note that different topics may have a different number of associated queries and thus the value of M is variable per topic; however, in the following we simplify this and assume a population of M users is observed, each issuing one and only one query for each topic. This assumption simplifies the development and the notation of the framework. The framework can be applied to situations where the number of queries varies across topics; this is further discussed in Section 7.

[3] In finance, $\alpha \in [0, +\infty]$. Here, however, we consider $\alpha \in [-\infty, +\infty]$; we explore the implications of this choice in Sections 4 and 5, and discuss this difference in Section 7.

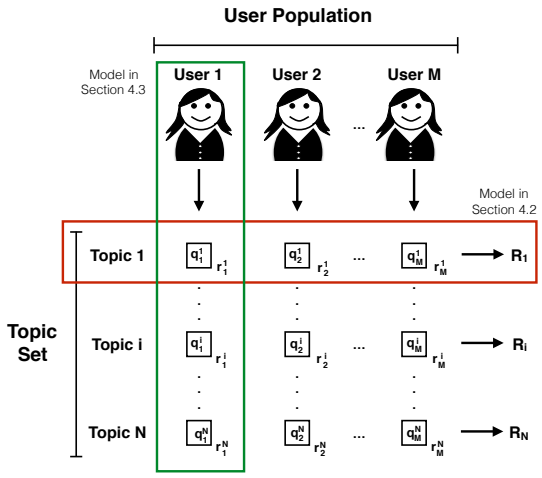

Figure 1: Settings for the general mean variance evaluation framework. Two sources of variance are modelled: the topic set and the user population. The red box highlights the subset of settings considered by the intra-topic MVA evaluation of Section 4.2. The green box highlights the subset of settings considered by the inter-topic MVA evaluation of Section 4.3.

Referring back to the financial parallel, each topic is a share investment[4] and portfolio optimisation acts on the set of N shares (topics in the IR scenario). Certain topics may be more important (e.g., valuable or popular) than others. We represent this importance by associating to each topic an importance value w, which represents the IR counterpart of the financial wealth of investment for a share. A large value of w_i corresponds to ascribing large importance to topic i over other topics in the portfolio. In the empirical experiments of this paper we will assume constant wealth investments across topics; however query traffic information, advertising returns, and expert knowledge may inform the relative level of investment to attribute to each topic.

Each topic is associated with a rate of return R_i. As in the financial scenario, R_i is a random variable with a probability distribution over all estimation points q_k^i; but in the IR scenario, the estimation points q_k^i correspond to the queries that users issue for topic i. Then, by issuing a query q_k^i to the search engine and examining the retrieved documents according to the user model encoded in the measure used to estimate systems effectiveness in the evaluation, the user accesses a rate of return r_k^i. In IR, the rate of return r_k^i corresponds to the value of the evaluation measure for that retrieval task (e.g., average precision, nDCG, etc.) obtained by the system for the search results provided in answer to query q_k^i. For example, $\mathcal{Q}_i = \{q_i^1, q_i^2, \ldots, q_i^M\}$ is the set of queries users would issue to the search engine for topic i and each query produces a rate of return r_k^i. This scenario is pictured in Figure 1.

Overall, given a topic i, the probability of obtaining a rate of return of r_k^i is given by the amount of queries for which the system obtained the effectiveness r_k^i, i.e., $p(r_k^i) = (\#\text{ queries with performance } r_k^i)/|\mathcal{Q}_i|$, where the operator $|.|$ denotes the size of the set. Note that the probability of the rate of return r_i^j for topic i is influenced by how many of the queries associated with topic i produce a search effectiveness of r_i^j. This also includes the fact that some of the M users of the system for topic i may have issued the same query: thus, $p(r_i^j)$ also accounts for the popularity of a query for expressing an information need, as recorded by query usage.

Within these settings, the expected return of the portfolio of topics (equation 2) modulates the effectiveness achieved by the search engine on a query for a topic (rate of return) with its popularity among the queries for that topic (probability of the rate of return), and then it scales this by the importance (level of investment) of each topic.

The variance of the portfolio of topics (equation 4) is instead dependent on the variance in effectiveness recorded among the queries associated to each topic, and on the covariance between each pair of topics.

The covariance is a measure of how much two variables change together: in financial terms this measures how much the rate of return of a share i tracks that of a share j. Positive covariance indicates that two shares increase or decrease in value at the same time, while negative covariance indicates that one share increases in value while the other decreases, and vice versa. In IR, covariance models trends in query effectiveness across topics. Positive covariance across two topics is encountered when those users that issued a good (highly effective) query for a topic, issued a good query also for the other topic, while those that issued a poor query, did so for both topics. Negative covariance across two topics is encountered when those users that issued a good query for a topic did issue a poor query for the other, and vice versa.

Given these settings, two systems can be compared with respect to the value of the portfolio associated to each system following equation 5. Specifically, given systems A and B with associated portfolios P_A and P_B (same topics and queries, thus consequently same wealth distributions, but different rate of returns per query), system A is better than B if and only if

$$\begin{aligned} O_A > O_B \implies \mu_A - \alpha\sigma_A^2 > \mu_B - \alpha\sigma_B^2 \\ \implies E[R_A] - \alpha var(R_A) > E[R_B] - \alpha var(R_B) \end{aligned} \quad (6)$$

Using equation 4, the previous inequality can be rewritten with respect to the individual rates of returns for each query:

$$\begin{aligned} \sum_{i=1}^{N} w_i \cdot \Big(\sum_{k=1}^{M} r_k^{i\,A} p(r_k^{i\,A}) \Big) - \alpha \sum_{i=j=1}^{N} w_i^2 var(R_i^A) \\ -\alpha \sum_{i \neq j}^{N} w_i w_j cov(R_i^A, R_j^A) > \\ \sum_{i=1}^{N} w_i \cdot \Big(\sum_{k=1}^{M} r_k^{i\,B} p(r_k^{i\,B}) \Big) - \alpha \sum_{i=j=1}^{N} w_i^2 var(R_i^B) \\ -\alpha \sum_{i \neq j}^{N} w_i w_j cov(R_i^B, R_j^B) \end{aligned} \quad (7)$$

Since we assumed all topics have the same importance, w_i and w_j can be ignored for rank equivalence reasons. From inequalities 6 and 7, assuming $\alpha > 0$, we can observe that:

- Given two systems with the same mean effectiveness across topics ($\mu_A = \mu_B$), system A is considered better than B if the variance of A is lower than that of B. This variance is low when the effectiveness of different queries for the same topic is similar, i.e., the system has a similar effectiveness for all queries of a topic. In this case, the system with lower variance is more stable than the other for all query variations across each single topic and thus, every

[4] Simply referred to as a share in the following.

user experiences nearly the same effectiveness, regardless of their query. This means that system A is not fitting to a specific user querying, but is consistent across users.

- Given two systems with the same mean effectiveness across topics ($\mu_A = \mu_B$) and the same total variance across topics (i.e., $\sum_{i=j=1}^{N} w_i^2 var(R_i^A) = \sum_{i=j=1}^{N} w_i^2 var(R_i^B)$), system A is considered better than B if the covariance between queries issued by the same user across topics is lower for A than for B. This covariance is low (to the point of being negative) when there is little correlation between users and system effectiveness, i.e., a user issuing a query with high effectiveness on a topic, also issues a query with low effectiveness for another topic. This means that system A is not fitting to a specific user querying, but is consistent across users.
- If the mean effectiveness of system A is lower than that of B ($\mu_A < \mu_B$), system A may still be better than B. This happens when the difference between the variance of the portfolio of topics for the two systems ($\alpha(\sigma_A^2 - \sigma_B^2)$) is more than the difference between the means ($\mu_A - \mu_B$). That is, when A has a lower mean, then it has to have a "comparably" lower variance (being more stable) in order to be preferred to B. The difference in variance between the two systems is modulated by the risk preference parameter α: higher risk aversion ($\alpha >> 0$) requires higher variance difference, and thus a stricter requirement on the stability of A compared to that of B. Of course, setting the right value of α is essential to maintain the quality of the evaluation, as large values of α would annihilate the mean, resulting in systems being assessed only in terms of variance (discarding actual effectiveness).
- Finally, when no risk preference is expressed ($\alpha = 0$), the mean variance evaluation framework resolves to only compare mean effectiveness across queries and topics. This is akin to the current practice in evaluation of IR systems.

These observations can be summarised as: in the presence of risk aversion ($\alpha > 0$) and given the same mean effectiveness, a system which has consistent effectiveness for all users across topics and queries is preferred.

A negative value of risk preference α leads to opposite observations, thus favouring a system over another on the basis of higher variances and covariances. To put it in other terms: in presence of risk propensity ($\alpha < 0$) and given the same mean effectiveness, a system which has inconsistent effectiveness for users across topics and queries is preferred. Expressing a preference towards inconsistent, unstable systems ($\alpha < 0$) may sound counterintuitive. However, note that a system with large variance but low covariance, may be preferred by specific types of users as this would provide them with large gains over the mean effectiveness, if they were able to always (or often) construct highly effective query variations. More discussion about the risk sensitivity parameter is developed in Section 7.

4.2 Intra-Topic Evaluation

We now consider two specialisations of the mean variance evaluation framework. The first specialisation consists of a user or (user population) issuing multiple queries, but for a unique topic only, as illustrated in Figure 1, red box. This case is of interest because allows us to focus on analysing a specific topic; topic-focused analysis of query variations was performed, for example, in the TREC 8 Query Track [5].

Under these circumstances, the variance of the rate of return for the considered topic ($var(R_1)$) can be computed; instead the covariance elements of inequality 7 are null because only one topic is considered. Thus, systems are compared according to inequality 7 but with covariances set to zero: thus only the mean effectiveness and its variance over the topic are considered. Formally, for intra-topic evaluation, the framework for comparing systems simplifies to:

$$w_1 \cdot \Big(\sum_{k=1}^{M} {r_k^1}^A p({r_k^1}^A)\Big) - \alpha w_1^2 var(R_1^A) > w_1 \cdot \Big(\sum_{k=1}^{M} {r_k^1}^B p({r_k^1}^B)\Big) - \alpha w_1^2 var(R_1^B) \quad (8)$$

As observed before, this means that, given the same mean effectiveness and if $\alpha > 0$, a system is favoured over another if its variance is lower, i.e., if the system is more stable across query variations for the same topic. Stability of effectiveness across query variations (and not anymore in addition to that across topics as it was for the general framework) is also the main factor to determine whether A is better than B when A's mean effectiveness is lower than B's.

4.3 Inter-topics Evaluation

The second specialisation consists of a user issuing one query only to the system, but repeating this action for multiple topics (with a different query for each topic), as pictured in Figure 1, green box. These are the common settings for many IR evaluation tasks, e.g., past TREC ad-hoc tracks.

Under these circumstances, the variance associated with the rate of return of the portfolio (σ_P^2) is zero and thus, independent of risk preference, systems are evaluated only with respect to the mean effectiveness. If all topics are assigned the same importance (i.e., constant wealth of investment w), then the mean variance evaluation framework resorts to the common practice of averaging the effectiveness of each system over the topic set, and using these means for systems comparison. When w is not constant across topics, then the framework differs from the common practice of considering means only; yet there is no account for the variance in effectiveness across topics.

Modified Inter-topics Evaluation

To account for the variance across topics expressed by single query variations, the inter-topics evaluation developed according to the general mean variance evaluation framework can be modified. A first option for modifying the framework is to consider each topic as a sub-portfolio, a query as a share, and observe a distribution of measurements for each query variation. This avenue is further discussed in Section 8 and leads to a more complex evaluation framework, the investigation of which is left for future work.

Next, we consider a simpler modification of the framework to account for variance across topics when comparing systems. This consists of adapting the method for intra-topics evaluation by treating topics as query variations instead. According to this, system A is preferred to system B if

$$\sum_{i=1}^{N} w_i {r_1^i}^A p({r_1^i}^A) - \alpha var(w_P^2 R_P^A) > \sum_{i=1}^{N} w_i {r_1^i}^B p({r_1^i}^B) - \alpha var(w_P^2 R_P^B) \quad (9)$$

where $w_P^2 R_P^A$ is the distribution of rates of returns over the portfolio of topics weighted by the wealth invested in each topic. As in the intra-topic case, the covariance component is null and preference towards a system is judged with respect to both mean effectiveness and its variance across topics. This is different from the common approach in evaluating IR systems, which instead only focuses on the mean effectiveness across topics; yet, as noted before, such an approach can be recovered if α is set to zero.

5. EXPERIMENTS AND ANALYSIS

Next we empirically investigate the differences between the mean variance analysis framework and the common practice of using only mean effectiveness. These experiments aim to: (1) study how the proposed framework empirically differs from the current approach, and (2) provide examples demonstrating the use of the framework for IR evaluation.

To do so, we compute ranking of systems according to their mean effectiveness only, i.e., we rank systems in decreasing order of their effectiveness. We then compare this ranking with that obtained using mean variance evaluation. This study is performed across the three different cases we considered: the general mean variance evaluation (from Section 4), the intra-topic evaluation (from Section 4.2) and the (modified) inter-topic evaluation (from Section 4.3).

5.1 Experimental Design and Common Settings

System rankings obtained with the mean variance evaluation framework and the common approach of using the means across queries (and topics) are compared using Kendall's τ and AP rank correlation (τ_{AP}) coefficients [28]. To clarify, we are not comparing document rankings, but orderings of systems based on their effectiveness. Kendall's τ rank correlation coefficient has often been used in comparing the correlations between system rankings obtained by different evaluation measures (e.g., [2]) and methodologies (e.g., [1, 23]). AP correlation is based on average precision and gives more weight to differences at the top of the system rankings [28]. Rank correlation values equal or greater than 0.9 are commonly considered effectively equivalent [25]. The use of correlation to compare systems rankings obtained under different evaluation frameworks is a widely used methodology in IR[5], e.g. [2, 31, 30, 1, 23, 25].

Mean variance evaluation has two main parameters: w and α. In absence of prior information about the importance of each topic in the evaluation, we set w to a constant and we effectively remove it from the corresponding equations for rank equivalence reasons. Similarly, we do not have information to inform the choice of risk preference; however, rather than fixing the value of α, we explore the variation in system rankings (and corresponding τ and τ_{AP}) obtained by varying α in the range $[-20, 20]$ and in $[-1000, 1000]$ (step 0.1[6]). The former range shows the behaviour of the framework for small risk preference values; the latter shows its "asymptotical" behaviour.

For testing the general framework (Section 4.1) and the intra-topic variant (Section 4.2), we use the TREC 8 Query Track [5] and the CLEF 2015 eHealth Task 2 datasets [16] and consider all runs originally submitted to these tasks. The TREC dataset consists of 50 topics; each topic is associated to 23 queries (1,150 queries in total). The CLEF dataset comprises of three types of queries (*pivot*, *most* similar to the pivot, and *least* similar to the pivot, see [16]) for 21 topics (63 queries in total).

To test the modified inter-topics evaluation variant (Section 4.3), we use the TREC 2013 and 2014 Web Track data sets [7, 8] and the pivot queries from the CLEF dataset. These TREC tasks provide an example of a modern adhoc evaluation setting where each topic is represented by one query only. The CLEF dataset with only pivot queries provides a similar arrangement, and it allows us to explore the effect of topic variance, rather than query variance and topic covariance (investigated in the general evaluation).

To measure system effectiveness, we use precision at 10 (P10) and average precision (AP); P10 is the primary measure for the CLEF task, while AP is the primary measure for the TREC 8 task. While P10 is known to be unstable and AP is often preferred, our use of P10 is justified due to how systems in CLEF 2015 were pooled [16]. While we also report AP for CLEF, the measure may be unreliable on this dataset due to the small, shallow pools. Nevertheless, these observations do not influence the main goal of this empirical investigation nor do they influence its findings.

All empirical results and scripts that implement the different settings of the MVE framework are made available at http://github.com/ielab/MeanVarianceEvaluation. This also includes the τ correlations for the experiments in Sections 5.3 and 5.4: they have no noticeable difference when compared to τ_{AP} for these experiments and thus, for clarity, are not reported in the plots in the relevant sections. We also make available an interactive system to explore system rankings according to MVE for variable settings of α.

5.2 General Mean Variance Evaluation

This set of experiments compares system rankings obtained using inequality 7 and those using mean effectiveness only ($\alpha = 0$). That is, these experiments fully consider the general MVE framework proposed in this paper.

Figure 2 shows τ and τ_{AP} correlations between the two system rankings for CLEF 2015 and TREC 8 with varying values of α, for P10 and AP as measures of effectiveness.

The following observations can be made.

There are many cases where, for a topic, a system is better than another for a query variation but worse for another. For example, for Topic 1 in CLEF, run `FDUSGInfo_EN_Run.5` has a P10 of 0.9 for the *most* query and 0 for the *least* query; while `baseline_run.1` has P10 of 0.8 and 0.7, respectively. Systems `CUNI_EN_Run.6` and `KISTI_EN_RUN.1` are instead an example of systems with the same mean (0.3761 – a tie if only mean is considered) but different variance (5.5×10^{-3} and 6.5×10^{-3}, respectively), thus resulting in a clear ranking difference in the mean variance evaluation framework. Finally, systems `KISTI_EN_RUN.3` and `CUNI_EN_Run.5` provide an example where the first system has a higher mean than the other (0.3730 and 0.3682, respectively), but is ranked below the second in the mean variance framework when[7] $\alpha \geq 0.9$ because their variance is 8.9×10^{-3} and 3.0×10^{-3}, respectively.

These previous cases involving runs 5 and 6 of `CUNI_EN`

[5] Note that another property of an evaluation measure that is often investigated is discriminative power. However, discriminative power is a property of evaluation measures and not of evaluation frameworks such as the MVA framework proposed here and thus its investigation in our case is of little significance. In addition, methods like bootstrap to compute discriminative power are not applicable to the MVA framework because "the most basic assumption that it [the bootstrap method] relies on is that the original topics of the collection are independent and identically distributed samples from the population" of possible topics [19].

[6] Only 51 steps (11 for intra-topic) are reported in the result figures for clarity; trends with all steps are equivalent.

[7] It can be analytically derived that the rank inversion occurs for $\alpha \geq 0.83$.

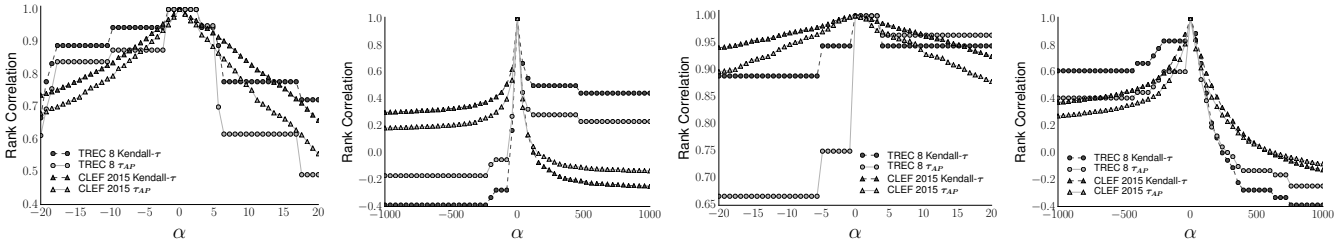

Figure 2: Correlations for CLEF 2015 and TREC 8 using P10 (left pair) and AP (right pair) for $-20 \leq \alpha \leq 20$ (left plot of each pair) and $-1000 \leq \alpha \leq 1000$ (right plot of each pair).

and run 1 and 3 of `KISTI_EN` are compelling examples showcasing that the MVE framework can individuate (and demote) systems that are more sensitive to query variations.

System rankings are comparable when the risk sensitivity parameter α is close to zero as correlations are > 0.9. However, as α positively increases or negatively decreases, rankings become less correlated and thus the two evaluation practices favour different systems. When P10 is used, rankings of CLEF systems do effectively differ for $\alpha \geq 8$ and $\alpha \leq -6$. For TREC 8 systems, the actual values of α for which rank correlations are less than 0.9 do differ, although are comparable in magnitude. When considering a wider range for α values, we can observe that there are always values of α for which systems rankings largely differ.

We can also observe that rank correlations plateau beyond certain α values: this is because at those points the variance component of inequality 7 is larger than the mean component and thus dominates the choice of system rankings, effectively resulting in ranking systems in decreasing order of variance. Note that when evaluating systems, these values of α should be avoided as in these circumstances actual effectiveness is ignored in favour of variance only.

It is interesting to also contrast rank correlations obtained with P10 and AP. While asymptotically the differences between evaluation approaches using these two measures are similar, the values of correlations at which they converge is different, both across measures and collections. In addition, the two measures result in different values of correlations being observed for α close to zero, especially for CLEF 2015, e.g., while system rankings obtained with P10 are distinguishable when $\alpha \geq 8$ ($\tau < 0.9$), for AP rankings they are not distinguishable when $-15 \leq \alpha \leq 10$. This is because differences in P10 are much larger than differences in AP, e.g., the gain of a relevant topic within the top ten ranks is 0.1 for P10 and 0.1 divided by the number of relevant documents for AP. This suggests that α values should be tailored to users, measures and contexts (datasets): in particular, measures act on different parts of the scale of possible values and α needs to be set accordingly.

5.3 Intra-topics Evaluation

This set of experiments compares system rankings obtained using the intra-topic evaluation method (inequality 8) and those using mean effectiveness only ($\alpha = 0$). This analysis focuses on query variations only (ignoring that there may be topics other than the one at hand). This setting has been considered for example by Buckley and Walz for TREC 8 [5] (for selected topics) to determine how effectiveness varies due to both system differences and query differences through an ANOVA analysis.

Figure 3 shows the τ_{AP} correlations between system rankings evaluated with P10 for TREC 8 and CLEF 2015 obtained for each topic (reported in the z-axis of the plots).

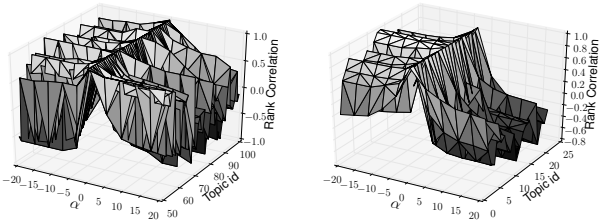

Figure 3: Correlations for TREC 8 (left) and CLEF 2015 (right) using P10 for $-20 \leq \alpha \leq 20$ in the intra-topic evaluation setting. Topics are analysed individually and are displayed on the z-axis.

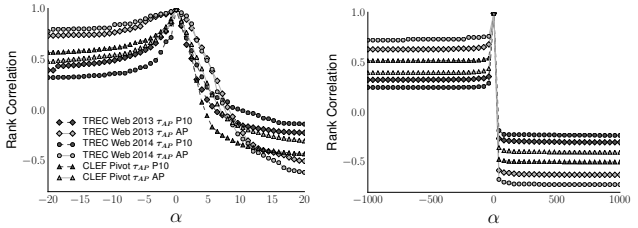

Figure 4: Correlations for TREC Web Track 2013 and 2014 (adhoc) and CLEF 2015 (pivot queries only) using P10 for $-20 \leq \alpha \leq 20$ (left) and $-1000 \leq \alpha \leq 1000$ (right).

The figures show that apart from $\alpha = 0$, differences in systems rankings between the two evaluation approaches depend both on α and the topic. For some topics systems rankings cannot be effectively distinguished ($\tau_{AP} \geq 0.9$), e.g., for Topic 12 of CLEF 2015. However, for other topics systems rankings largely differ, and this is so regardless of the value of α. As for previous evaluation settings, the figures also show that correlations reach asymptotical behaviour with $\alpha >> 0$ or $\alpha << 0$.

5.4 Modified Inter-Topics Evaluation

This set of experiments compares system rankings obtained using the modified inter-topics evaluation method (inequality 9) and those using the mean effectiveness only ($\alpha = 0$). In this case there are no query variations, but variance is instead produced by topic variations.

Figure 4 shows the τ_{AP} correlations between system rankings for TREC Web Track 2013 and 2014 and CLEF 2015 (pivot queries only).

As for other evaluation settings, the results show that when variance is of little importance ($\alpha \to 0$) then the two comparison approaches show similar findings. However, when variance has more influence, then system rankings greatly diverge. Rank correlations exhibit a similar behaviour as a function of α to the general evaluation settings; however the transition between correlated and not-correlated system rankings appears sharper for intra-topics

evaluation than for the general evaluation settings with both query and topic variations (compare Figures 2 and 4). This suggests that, in the mean variance evaluation framework and with real IR collections, query variations have a greater impact for determining systems rankings than topic variations. Thus, in general, systems are more robust across topics than across query variations, for the considered datasets.

6. PILOT USER EXPERIMENT

The previous section has examined the differences between the MVE framework and the use of mean alone using TREC data and exploring a range of intervals for α. Overall, the results showed that taking into account variance produced different system rankings. Next, we aim to empirically confirm this by means of a pilot user experiment that studies whether users consider only mean effectiveness or they also are sensible to effectiveness variance when asked about systems preference. We consider the settings of the inter-topic evaluation (Section 4.3). This pilot experiment does not pretend to be comprehensive, but it reports a methodology and initial results to support a more in depth study of user preferences about systems stability.

We recruited 15 domestic undergraduate IT students from the Queensland University of Technology with no specific expertise in IR. We developed a side-by-side evaluation interface like that in [24] that showed 5 search result snippets (title and 4 lines of text) of two systems (A (left) and B (right)); snippets were randomly ranked. We showed all assessors the same snippets for a first batch of 10 queries from the TREC 2014 Web Track. Known-relevant snippets were generated from a document marked relevant in the qrels using the default Elastic Search snippet generation algorithm; subsequent manual intervention by the authors of the paper modified each snippet to improve its quality. The same method was used to generate known-irrelevant snippets from qrels, including manual intervention. Shown snippets were either known relevant or irrelevant.

We produced two systems with the same mean precision at 5 (P5) of 0.4. We assigned to System S_1 2 known-relevant snippets per query (variance $\sigma_{S_1} = 0$); to S_2 we assigned 4 relevant snippets for 5 randomly selected queries and 0 relevant snippets for the remaining 5 queries ($\sigma_{S_2} = 0.1\bar{7}$).

Assessors were instructed the aim of the searches was to satisfy recall-oriented tasks and they considered each of the 10 queries consecutively. Furthermore, they had to mark the snippets they felt were relevant (94% agreement with the known-relevant/irrelevant TREC assessments). After all queries were examined assessors were asked to indicate which system between A or B they preferred, or whether they were equivalent. To avoid position bias, the assignment of S_1 and S_2 to A and B was randomised across users.

Overall 9 users expressed a preference towards S_1 (stable system), 3 towards S_2 and 3 felt the systems were equivalent. The first group is thus characterised by $\alpha > 0$, the second by $\alpha < 0$, and the third by $\alpha = 0$. The experiment was repeated by showing the assessors further 10 queries, but varying S_2 to increase its mean effectiveness to 0.46, such that it retrieved 0 relevant for 3 queries, 1 relevant for 2 queries, 4 relevant for 3 queries, and 5 relevant for 1 query ($\sigma_{S_2} = 0.17$). Under this setting, one assessor from the first and third group each flipped their preference from S_1 (stable) or equivalent to S_2 (unstable, but now more effective than S_1); other preferences remained unchanged. According to these preferences, $\alpha > 0.35$ for the first group, $\alpha < 0.35$ for the second group, and $\alpha = 0.35$ for the third group.

While these pilot experiments are admittedly underpowered (both in topics and users numbers) to allow to generalise these observations, they provide evidence that effectiveness variance influences overall system preference and that different users have different preferences towards levels of system instability (α). These experiments also provide an example of how the evaluation framework could be used and serve as an initial protocol for further empirical tests and validations of the proposed framework in future work.

7. DISCUSSION AND LIMITATIONS

Mean variance evaluation compares IR systems on the basis of both the systems' mean effectiveness over queries and the associated variance. This goes beyond the current practice, where systems are compared only on the basis of means, while variance is only used to establish the significance of differences observed in mean effectiveness (through statistical tests). Next, we summarise and discuss the core features of the framework and its current limitations.

Query variations. To the best of our knowledge, MVE represents the first, comprehensive attempt to explicitly modelling query variations for IR systems evaluation.

The framework assumes that it is possible to clearly identify which queries are associated to a topic. While this is possible in controlled evaluation environments such as TREC, CLEF and similar initiatives, it is unclear if this would be possible in real world, industrial environments. While evaluation per se is usually associated to some type of approximation of the real world, we believe that this problem could be often partially mitigated by methods that attempt to predict query intent, e.g., [4, 17]: thus, queries with the same or similar intent could be grouped together to form sets of query variations for the common underlying information need.

Wealth of investment w. The wealth of investment controls the relative importance of each topic in the evaluation and comparison of systems effectiveness. This can be set according to the importance of each topic as perceived by users; e.g., preference to work-related over leisure-related topics. Alternatively, this could be informed by query traffic or advertisement returns from the search engine provider; e.g., assigning more importance to information needs that are more popular or that generate more profits in terms of referral and advertisement. The empirical comparisons in this paper have considered fixed wealth of investment across topics; however, future work will empirically explore the impact of such features on the evaluation of systems.

While previous NTCIR diversity tasks (and associated measures) have considered the popularity of query intents in the evaluation of IR systems [20], we are not aware of other general evaluation approaches that explicitly control for the importance of some information needs over others.

Risk sensitive parameter α. In the framework, the value of the portfolio of queries/topics is influenced by the parameter α, which controls the level of variance in effectiveness that is tolerated by the user. The results in Figure 2 highlighted that the same value of α produces different systems rankings depending on the measures used to compute rate of returns (effectiveness). This is because different measures act on different parts of the scale and the impact of a relevant document is modelled differently by each measure. This implies that α needs to be set individually per

measure. This is in line with the intuition that different measures are grounded in different user models (e.g., the user model of P10 is different from that of AP) and thus each user model reacts differently to different magnitudes of variations in effectiveness. Ascertained that α depends upon users and measures (and thus, tasks and contexts), it is yet unclear how α could be estimated a priori.

When instability is preferred. In finance, only positive values of the risk sensitive parameter are allowed ($\alpha \in [0, +\infty]$) because it is assumed that investors are risk averse (although people may seek risk for the sake of risk, e.g., casino gamblers). In developing the mean variance evaluation framework, instead, we have not made this assumption and have allowed $\alpha \in [-\infty, +\infty]$. As seen in Section 4.1, when α is positive, systems with smaller variance are preferred and thus systems that are stable over query and topic variations are preferred over more unstable ones. However, if α is negative, then unstable systems are preferred. We did not restrict the values of α to the positive space because we do not know a priori in which circumstances (if at all) information seekers exhibit behaviour that may be modelled by risk propensity ($\alpha < 0$). Indeed, it is unclear when users would prefer unstable systems, although this may happen (as in the pilot user experiments). Intuitively this may be the case for less valuable search tasks where the user may prefer risky information seeking strategies, thus preferring systems that are more variable in terms of effectiveness, hoping to striking higher gains earlier. These circumstances need further empirical exploration and validation; and overall, the framework itself needs to be further validated against users' preferences and behaviour to ascertain the exact role of variance and risk in assessing system preferences.

Covariance. In the framework, the covariance is responsible for tracking the comparative effectiveness of queries issued by different users on different topics. This is a novel aspect of the MVE framework, and in fact covariance is null when settings akin to those used in traditional IR evaluation are considered (inter-topics evaluation).

Variable number of queries per topic. In the development of the framework, we assumed an equal amount of query variations for each topic. Without this assumption, covariance between the rate of returns for two topics cannot be computed, as the covariance between two random variables of different dimensionality is undefined (and would have no meaning otherwise). This issue could be addressed in a number of ways.

Firstly, downsampling could be applied, by removing the extra queries from those topics with larger amount of query variations. This may however not be ideal, as the evaluation would ignore a number of potentially valuable variations.

Secondly, inference could be used to estimate the likely effectiveness of a missing query for a topic, had the user issued it. This effectiveness could be inferred by using the queries the user issued for other topics. This solution has the drawback that it assumes the effectiveness of queries for some topics can be used to estimate the hypothetical effectiveness the user would have obtained on other topics.

Thirdly, the intra-topic evaluation framework could be used. This would evaluate each topic separately; then systems could be compared by studying their portfolio values O_P on each topic, possibly considering both the mean and variance of such distributions. Doing so, the computation of covariances would not be required, thus addressing the issue that covariances are undefined.

8. CONCLUSIONS AND FUTURE WORK

We have proposed the mean variance evaluation (MVE) framework for IR that explicitly models and separates query variations from topic variations. The framework goes beyond the use of mean system effectiveness across topics or queries for comparing IR systems. Instead, the framework bases systems rankings also on the variance in effectiveness across query variations and topics. In addition, it directly models the preference or aversion of users towards unstable systems and the relative importance of different topics within the evaluation. Note that the proposed evaluation approach does not replace the use of statistical testing to assess whether systems are distinguishable.

The empirical results analysed in this work highlighted that MVE ranks systems differently from the current practice in IR and they established the effect of query variations and risk sensitivity when comparing systems.

This work contributes to providing the foundations for building Cranfield-like, controlled and repeatable IR evaluations more aligned with user behaviour by explicitly modelling query variations. Clearly, there are a number of ways in which we can develop this work further, from refining the proposed framework to applying the theory in practice. Directions for future work include:

Framework refinements. MVE could be further refined by considering each topic as a separate sub-portfolio. This would mean that each query variation will be modelled as if it were a share, rather than a point-wise estimation of the value of a share. This approach requires that each query in a topic (sub-portfolio) corresponds to a number of measurements which are used to estimate the mean value and the variance of the share. While introducing a new requirement, this extension also introduces the opportunity of exploiting query-based measurements to model uncertainty in the effectiveness of a system's answer to a query for a user. In other words, while given a query and a search result ranking, the current approach computes a point-wise value for one evaluation measure, the envisioned extension may consider a user model such as that of RBP [14], where the user has a probability of assessing a document at a specific rank position. By considering such probabilities, the evaluation may use the gain distribution provided by the documents in the result ranking as a way to compute the rate of return for the query. Such gain would have a mean and a variance, which is determined also by the likelihood of the user assessing the documents at each rank position. This extension would also offer an alternative to the three solutions identified in Section 7 to address the presence of a variable number of queries per topic. Similar ideas have been put forward by Carterette et al. who attempted incorporating variability of user behaviour into system evaluation [6]. This line of thoughts has deeper implications beyond the framework proposed here. The framework strives to "reintroducing" a stronger notion of user into the Cranfield paradigm by questioning the assumption of a single query/user for an information need. Similarly, the role of a single relevance assessment for a query-document pair could be questioned: by providing a single relevance assessment for each document, the Cranfield paradigm "collapses" all users in this single dimension as well. The presence of query and user variations

appear instead to justify considering variations with respect to relevance assessments. Relevant to this research avenue, some initial work has explored variations in relevance assessments, e.g. [21], and evaluation measures have emerged where relevance measurements are tailored to characteristics of single users, e.g., the level at which they understand the retrieved information [34, 32]. Similarly, Yilmaz et al. have shown that the choice of topic intent descriptions affect relevance assessment and evaluation [29], thus reaffirming the need for a holistic framework like that proposed in this paper to tackle information need and query variations.

Framework extensions for diversity task. MVE may be applied to model evaluation in the context of diversity being required in the search result sets [20]. To proceed with this, we suggest that each topic intent is modelled as a different query variation. Thus, for each topic, this approach would produce a distribution of rate of return over topic intents; this would not require the use of diversity measures. Whereas, standard evaluation measures could be used; the probability of a query variation may be modelled using the probability of an intent, if available [20].

Framework extensions for comparisons across tasks. MVE may be used to evaluate IR systems across tasks. This would require adding a further source of variance, the task. The increased complexity of the framework would allow comparing systems based on their effectiveness and stability across different tasks, each possibly characterised by their own evaluation measures, which, however, are required to be mapped to a common rate of return scale.

Comparison with user preferences. Aside from the pilot user experiment, the empirical results reported in the paper mainly relied on correlation to demonstrate to what extent and under which settings MVE differs from mean effectiveness only. While correlation is a widely adopted indicator when studying measures and evaluation approaches in IR [2, 31, 30, 1, 23, 25], these results do not necessarily demonstrate that the systems ranking produced with the proposed framework are actually preferred to that with mean alone. Nevertheless, previous work and empirical evidence strongly support the idea that system stability is an important aspect in assessing the effectiveness of systems and this may be reflected by user preferences [30, 31], as our pilot results showed. Future work needs to extend the pilot experiment to investigate whether system preferences obtained by MVE correlate with those expressed by users, and for which settings of α. This user-based study would also allow for an **estimation of the risk aversion parameter** α for different tasks, user categories, contexts and evaluation measures.

9. REFERENCES

[1] J. A. Aslam, V. Pavlu, and R. Savell. A Unified Model for Metasearch, Pooling, and System Evaluation. In *CIKM*, 2003.

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. User Variability and IR System Evaluation. In *SIGIR*, 2015.

[3] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV: A Test Collection with Query Variability. In *SIGIR*, 2016.

[4] D. J. Brenes, D. Gayo-Avello, and K. Pérez-González. Survey and Evaluation of Query Intent Detection Methods. In *WSCD*, 2009.

[5] C. Buckley and J. A. Walz. The TREC-8 Query Track. In *TREC*, 1999.

[6] B. Carterette, E. Kanoulas, and E. Yilmaz. Incorporating variability in user behavior into systems based evaluation. In *CIKM*, 2012.

[7] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. Clarke, and E. M. Voorhees. TREC 2013 Web Track Overview. In *TREC*, 2013.

[8] K. Collins-Thompson, C. Macdonald, P. Bennett, F. Diaz, and E. M. Voorhees. TREC 2014 Web Track Overview. In *TREC*, 2014.

[9] B. T. Dinçer. Statistical Principal Components Analysis for Retrieval Experiments. *JASIST*, 2007.

[10] B. T. Dinçer, C. Macdonald, and I. Ounis. Hypothesis Testing for the Risk-sensitive Evaluation of Retrieval Systems. In *SIGIR*, 2014.

[11] B. Koopman and G. Zuccon. A test collection for matching patient trials. In *SIGIR*, 2016.

[12] H. Markowitz. Portfolio Selection. *The journal of finance*, 1952.

[13] A. Moffat, F. Scholer, P. Thomas, and P. Bailey. Pooled Evaluation Over Query Variations: Users are as Diverse as Systems. In *CIKM*, 2015.

[14] A. Moffat and J. Zobel. Rank-biased Precision for Measurement of Retrieval Effectiveness. *ACM TOIS*, 2008.

[15] J. Palotti, G. Zuccon, J. Bernhardt, A. Hanbury, and L. Goeuriot. Assessors agreement: A case study across assessor type, payment levels, query variations and relevance dimensions. In *CLEF*, 2016.

[16] J. Palotti, G. Zuccon, L. Goeuriot, L. Kelly, et al. CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving information about medical symptoms. In *CLEF*, 2015.

[17] F. Radlinski, M. Szummer, and N. Craswell. Inferring Query Intent from Reformulations and Clicks. In *WWW*, 2010.

[18] S. Robertson. On GMAP: And Other Transformations. In *CIKM*, 2006.

[19] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *SIGIR*, 2006.

[20] T. Sakai and R. Song. Diversified Search Evaluation: Lessons from the NTCIR-9 INTENT task. *Inf Ret*, 2013.

[21] M. Sanderson, F. Scholer, and A. Turpin. Relatively relevant: Assessor shift in document judgements. In *ADCS*, 2010.

[22] M. D. Smucker, J. Allan, and B. Carterette. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *CIKM*, 2007.

[23] I. Soboroff, C. Nicholas, and P. Cahan. Ranking Retrieval Systems without Relevance Judgments. In *SIGIR*, 2001.

[24] P. Thomas and D. Hawking. Evaluation by Comparing Result Sets in Context. In *CIKM*, 2006.

[25] E. M. Voorhees. Evaluation by Highly Relevant Documents. In *SIGIR*, 2001.

[26] J. Wang. Mean-variance Analysis: A New Document Ranking Theory in Information Retrieval. In *ECIR*, 2009.

[27] J. Wang and J. Zhu. Portfolio Theory of Information Retrieval. In *SIGIR*, 2009.

[28] E. Yilmaz, J. A. Aslam, and S. Robertson. A New Rank Correlation Coefficient for Information Retrieval. In *SIGIR*, 2008.

[29] E. Yilmaz, E. Kanoulas, and N. Craswell. Effect of intent descriptions on retrieval evaluation. In *CIKM*, 2014.

[30] P. Zhang, L. Hao, D. Song, J. Wang, Y. Hou, and B. Hu. Generalized bias-variance evaluation of TREC participated systems. In *CIKM*, 2014.

[31] P. Zhang, D. Song, J. Wang, and Y. Hou. Bias-Variance Decomposition of IR Evaluation. In *SIGIR*, 2013.

[32] G. Zuccon. Understandability biased evaluation for information retrieval. In *ECIR*, 2016.

[33] G. Zuccon, L. Azzopardi, and K. van Rijsbergen. Back to the Roots: Mean-variance Analysis of Relevance Estimations. In *ECIR*, 2011.

[34] G. Zuccon and B. Koopman. Integrating understandability in the evaluation of consumer health search engines. In *MedIR*, 2014.