A model for investigating and calibrating IR systems evaluation

Gabriella Kazai
Vishwa Vinay
Eduarda Mendes Rodrigues
Natasa Milic-Frayling
Chung Tong Lee
Aleksandar Ignjatovic

7 May, 2010

Technical Report
MSR-TR-2010

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

A model for investigating and calibrating IR systems evaluation

Gabriella Kazai‡
v-gabkaz@microsoft.com

Vishwa Vinay‡
vvinay@microsoft.com

Eduarda Mendes Rodrigues‡
eduardamr@acm.org

Nataša Milić-Frayling‡
natasamf@microsoft.com

Chung Tong Lee†
ctlee@cse.unsw.edu.au

Aleksandar Ignjatović†
ignjat@cse.unsw.edu.au

‡Microsoft Research
Cambridge, CB3 0FB, UK

†School of Computer Science and Engineering
University of New South Wales,Australia

ABSTRACT
Benchmarking practices in information retrieval rely on measuring the per-topic performances of systems and aggregating these across the topics in a given test set. For an evaluation experiment, the per-topic scores represent the values in the matrix of the participating systems and the set of topics. In the absence of explicit external reference points indicating the true performance of systems, such a matrix represents a relative view over a sample of the universe of possible system-topic pairs, where a cyclical dependency exists between the systems and the topics. In this paper we develop a unified model for system evaluation by systematically modeling the relationship between topics and systems and by generalizing the way overall system performance is obtained from the individual topic scores with the use of a generalized means function with adaptive weights. We experiment with multiple definitions of the means on TREC evaluation runs and compare our rankings with the standard TREC averages. Our analysis of the different evaluations leads to recommendations for calibrating evaluation experiments.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness).

General Terms
Performance, Reliability, Experimentation

Keywords
System performance, Topic discernment, Performance metrics.

1. INTRODUCTION
A fundamental area of information retrieval (IR) is the evaluation of IR systems. For the past two decades, evaluation practices have largely been shaped by the TREC (Text REtrieval Conference) evaluation initiative ([12][13]). TREC follows a benchmarking design where a community of practitioners agrees on the data sets, search topics, and evaluation metrics in order to assess the relative performance of their systems. Typically, participants submit a set of runs to TREC, each containing, for each topic in the test set, a list of documents retrieved from the provided collection. For a selected subset of the submitted results, the TREC organizers collect relevance judgments. These are then used to calculate the performance metrics for each system and topic pair. The most commonly used metrics are Average Precision (AP) and R_Precision [12]. The overall system performance is usually given by the mean value of the per-topic scores, e.g., the Mean AP value (MAP), which is then used to compare the systems.

An important issue regarding such benchmarking designs is the problem of measuring the extent to which the conclusions of a test can be relied upon. Work on measuring the statistical significance of the comparisons between systems, and related questions of how many topics are needed to establish differences (e.g., [2]) provide a step toward answering this question. More fundamental issues include the questions of how representative sample a given test set is from the universe of all possible topics and how the evaluation is affected by the characteristics of a given topic set. For example, there have been attempts to characterize a *topic's difficulty*, reflected through the performance scores across systems for that topic, and to understand the implications for the design of performance measures that are suitable for comparing and ranking systems ([1][3][7]).

We consider an evaluation experiment a closed world, represented by the matrix of per-topic system performance scores. In the absence of existing reference points to some external ground-truth, such as the true ordering of systems, the matrix characterizes a relative universe in which questions, such as which topics can be trusted to provide the most reliable estimate of a system's true performance, or which systems present the most reliable estimate of a topic's difficulty arise. In addition to these, other questions concern the choice of metric to reflect per-topic system performance as well the choice of measure to aggregate

individual performance scores to arrive at an overall measure of system effectiveness or topic ease.

When it comes to defining an overall measure of effectiveness from the individual topic scores, concerns have been expressed in the literature regarding the use of simple averages to compare and rank systems. For example, the mean is not regarded as a good descriptor when the distribution of measurements is not normal, i.e., bell-shaped ([9], p.176), as it is the case with the AP values across topics. The simple mean also treats all the test topics in the sample equally, which has been shown to skew results [10].

In this paper we develop a unified model for system evaluation by systematically modeling the relationship between topics and systems and by generalizing the way overall system performance is obtained from the individual topic scores. Our framework allows to model concepts such as topic ease and discernment, or system conformity as properties of the test design. To measure overall performance, we use a *Generalized Adaptive-Weight Mean* (GAWM) measure, where the weights are derived from properties of the systems-topics matrix, consisting of per-topic system performance scores. For example, the weights associated with the test topics may reflect the ability of the topics to differentiate among the retrieval systems.

Using our framework, we experiment with multiple definitions of aggregate system performance metrics on TREC evaluation runs and compare our rankings with the standard TREC averages. Our analysis of the different evaluations leads to recommendations for calibrating evaluation experiments.

2. BACKGROUND AND RELATED WORK

2.1 Averaging issues

Taking the mean cross-topic effectiveness measure to be the representative number for a system assumes that all topics have equal weight in the evaluation. There are several reasons why this is not necessarily ideal:

a) *Numerical considerations*: larger numbers, thus easier topics, tend to dominate an averaging process. This is particularly true for measures like AP that are bounded by 0.
b) *Sampling inefficiency*: the topics in a test set represent only a sample from the universe of all possible topics, where different topics may be of varying importance.
c) *Evaluation objective*: an application specific scenario might require preferential treatment of some topics over others, e.g., hard topics in the TREC Robust track.

A historical example of (c) is [9], where van Rijsbergen suggested weighting the performance scores for a topic based on the topic's *generality*, i.e., the proportion of documents in the corpus that are relevant for the topic. In order to deal with the skewed distribution of performance metrics, Webber et al. [14] investigate the use of *score standardization*. For each system, they adjust the performance score for a topic by the mean and the standard deviation of performance scores for that topic achieved across a sample of systems. The GMAP measure by Robertson [10] can be seen as addressing (a) and (c). Topic differentiation is achieved indirectly through the use of the geometric mean of the AP values, which is sensitive to low values, i.e., to difficult topics. When going from a generic ad hoc task to the goal of the Robust track, the focus is on a subset of topics, specifically the *poorly performing* ones [15]. Given the need for such an evaluation, which is different from the standard TREC-style evaluation across topics, one alternative is to construct a test set of topics with the desired characteristics (i.e., only the *difficult* topics in this case). However, if a candidate set of topics has already been assembled, we might consider weighting the topics differentially to compensate for the lack of sufficient topics with the required characteristics (i.e., (b) above).

In this paper, we provide a method that generalizes the notion of averaging and includes adaptive weighting of performance statistics that is derived from the postulated dependence between *topic* and *system performances*. We begin by reflecting on related work that considers the issues of topic difficulty and coupling of topic characteristics and system performance.

2.2 System and Topic Characteristics

In the interpretation of performance metrics it is often tacitly assumed that some topics are more difficult than others, alluding that this is an inherent property of the topic. In practice, however, topic difficulty is estimated by examining how a given system is able to perform on it. Examples include KL-divergence in Cronen-Townsend et al. [4], the Jensen-Shannon divergence in Carmel et al. [3], document perturbation in Vinay et al. [11], and robustness score by Zhou and Croft [15]. An alternative approach to representing the discriminative value of a topic is based on the notion of *departure from consensus* used by Aslam and Pavlu [1] and Diaz [5]. Aslam and Pavlu [1] assume that those topics for which the systems deviate from the average performance across systems are difficult. Diaz [5] focuses on the system performance and argues that the systems that deviate from others on individual topics are suboptimal. Both papers use evidence from a group of systems to judge the difficulty of a single topic.

In all these cases, topic difficulty is strongly correlated with the AP measure. Thus, a high level of difficulty is attributed to a topic with low performance across systems. At the same time, a system is considered better if it performs better than others on difficult topics. Except in the work by Mizzaro and Robertson [7], this circular definition of topic difficulty and system performance has not been explicitly modeled in the retrieval evaluation. Mizzaro and Robertson relate topic and system performance through a bipartite network model. The network consists of system and topic nodes with edges propagating normalized AP statistics between them. The notions of *system effectiveness* and *topic ease* are then expressed in terms of the hubs and authorities of the system and topic nodes. Calculation of *node hubness* $\mathbf{h}$ and *node authority* $\mathbf{a}$ is facilitated by the system of equations

$$\mathbf{h} = \mathbf{A}\mathbf{a} \quad \text{and} \quad \mathbf{a} = \boldsymbol{A}^{\mathbf{T}}\mathbf{h} \;\rightarrow\; \mathbf{h} = \mathbf{A}\mathbf{A}^{\mathbf{T}}\mathbf{h}$$

that captures the dual relationship of topic ease and system effectiveness. This method is a special case of the approach that we propose.

2.3 Multi-Grader Problem

The need for characterizing both systems and topics fits well a class of multi-grader problems that has been studied by Ignjatović et al. [6]. There, *m* graders, e.g., retrieval systems, are marking *n* assignments, e.g., performing search and recording their performance score for each of the *n* topics, or vice versa. The assigned values represent the graders' affinity for the individual assignments. As it is often the case in practice, graders, may not have uniform criteria or capabilities and thus their scores vary considerably. The objective is to assign the final scores to the assignments (e.g., topics or systems), that capture the differences among the graders. Ignjatović et al. expressed these objectives as

seeking the final scores in the form of the *weighted averages* of the original grades. The weights, in turn, incorporate the difference between the unknown final scores and the original scores. The formulation leads to the Fixed Point for the function representing the weighted averages. It is this framework that we propose to use for modeling the systems' performances and the characteristics of the test topics. It will enable us to derive the ranking of systems and topics based on the newly derived metrics, incorporating the original performance metrics and their variability across systems and topics.

3. MATHEMATICAL MODEL

In this section we develop a mathematical model and describe the generalized performance metrics. The model builds on the system-topic matrix of an evaluation experiment. This matrix represents a sample of the universe of all possible systems and topics, given a collection of documents. Building on this matrix, we model the duality of the relationship between systems and topics and their circular influence on each other. We define operations over the matrix to explicate properties of the system-topic relationship and use these as analysis tools for calibrating an evaluation test. For example, when estimating the difficulty level of a topic as a weighted average of system performance scores, we may give more weight to systems that are good discriminators between topics. We start by defining the system-topic matrix and generalizing the mean function.

3.1 Weighted Means over the System-Topic Matrix

Given a set of *m* systems and *n* topics, we consider a real-valued system-topic matrix $\boldsymbol{\mathcal{P}}$ where the entry $\mathcal{P}[i,j]$ represents the retrieval performance of a system s_i for a topic t_j. Thus, the rows of the matrix correspond to the systems and the columns to the topics. We assume that higher values of $\mathcal{P}[i,j]$ correspond to better performance. We may use any single number measures to quantify performance, such as AP.

The *i*-th row of $\boldsymbol{\mathcal{P}}$, i.e., $\boldsymbol{\mathcal{P}}[i,*]$, defines the *performance* of system s_i on each topic of the test set. The overall measure of system effectiveness, $E_s[i]$, is a *mean* of the per-topic values in $\boldsymbol{\mathcal{P}}[i,*]$. In practice, it is common to use a simple average, i.e., Mean Average Precision (MAP), which gives uniform weights to all the topics. In contrast, we seek to determine a weight $W_t[j]$ for the individual topics t_j and compute the overall performance measure as:

$$E_s[i] = \mathcal{M}_s(\boldsymbol{\mathcal{P}}[i,*], \boldsymbol{W_t}). \quad (1)$$

where $\mathcal{M}$ denotes a generalized weighted mean function for calculating retrieval performance.

Similarly, we consider the performance scores for a given topic across systems and seek to determine the weights $W_s[i]$ for individual systems s_i:

$$E_t[j] = \mathcal{M}_t(\boldsymbol{\mathcal{P}}[*,j], \boldsymbol{W_s}). \quad (2)$$

In practice, it is common to look at the Average AP (AAP) for a topic across systems [7].

Conceptually, the quantities $E_s[i]$ and $E_t[j]$ are comparable to MAP and AAP (see Figure 1) but, through $\mathcal{M}$, we aim to generalize the form of the mean function and to introduce the non-uniform contribution of individual systems and topics. The choice of mean function in essence describes which values in $\boldsymbol{\mathcal{P}}[i,*]$ contribute more to $E_s[i]$. For example, the geometric mean used for GMAP emphasizes poorly performing topics, and in our setup this would be obtained by setting the function $\mathcal{M}_s$ appropriately.

3.2 Adaptive Weights for Systems and Topics

In the calculation of MAP, the weights $\boldsymbol{W_t}$ are taken to be uniform. In other scenarios, the relative contribution of each topic to the aggregate system performance is controlled by the appropriate choice of weights. In the absence of external inputs

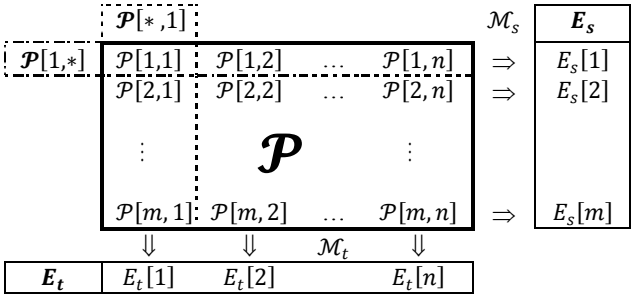

Figure 1. Performance matrix comprises system (rows) performance statistics for individual topics (columns). Aggregation of system statistics $\boldsymbol{E_s}$ reflects overall system performance and the aggregation of topic statistics $\boldsymbol{E_t}$ reflects topic ease.

that induce a preferential ordering over topics, we aim to set the per-topic weights internally (within the closed world) by monitoring the systems' collective performances. We start by defining an abstract function $\mathcal{F}$ that takes the performance matrix $\boldsymbol{\mathcal{P}}$ and maps it to a weight:

$$\mathcal{F}(\boldsymbol{\mathcal{P}}, \boldsymbol{E_s}) \rightarrow \mathbb{R}$$

The second argument of the function is the vector of system scores defined in Equation (1), leading to a circular definition. This is similar to the definition of hubs and authorities, where evidence of being a good authority is obtained by being connected to a good hub. And hubness is in turn dependent on being connected to good authorities. The selection of the function $\mathcal{F}$ depends on the goal of particular evaluation scenario, as will be discussed in the next two sections.

The existence of circular dependence between systems and topics is of course not new. However, making these relationships explicit not only ensures that the evaluation is transparent, but provides more control and configurability through the choice of mean and $\mathcal{F}$ functions.

In our system of equations, i.e., Equations (1) and (2), the only known quantities are the performance scores in matrix $\boldsymbol{\mathcal{P}}$. In order to solve for the inter-connected vectors $\boldsymbol{E_s}$, $\boldsymbol{E_t}$, $\boldsymbol{W_s}$, and $\boldsymbol{W_t}$, we employ an iterative procedure that interleaves the calculation of overall performance measures (represented by the **E** vectors) with updates to the weights (**W** vectors). This iterative process results in an adaptive set of weights, the topic values $\boldsymbol{W_t}$ being used to calculate the system measures $\boldsymbol{E_s}$, and vice versa. The iterative algorithm is defined by the following pseudo-code:

1: Select the $\mathcal{F}$ functions
2: Select the mean functions $\mathcal{M}_s$ and $\mathcal{M}_T$
3: Define a convergence threshold and the maximum number of iterations
4: Repeat until convergence or max number of iterations

5: For each system s_i compute $W_s[i]$
6: For each topic t_j compute $W_t[j]$
7: For each system s_i compute $E_s[i]$
8: For each topic t_j compute $E_t[j]$
9: End.

Our algorithm can be used with any per-topic system performance metrics with suitable preprocessing, e.g. AP, R-precision.

3.3 Evaluating with different viewpoints

The standard method of evaluation gives uniform weights to all topics for determining a system's overall performance. Similarly, when calculating AAP, systems are given uniform weight to contribute to topic ease. Using our framework, we can explicate the underlying assumptions of an evaluation setup and calibrate the evaluation to better reflect the objectives of the evaluation. Here we give two examples.

Duality of system conformity and topic discernment. A system performing closer to the average or other expressions of consensus across systems may be considered more reliable when assessing the difficulty of a topic. On the other hand, a topic characterized by varied system performances may be considered better discriminator of systems. Based on this relationship, we can define the following two axioms:

A1. The more diverse the systems' performance on a topic, i.e., the higher the *topic discernment*, the more significant the contribution of that topic to the overall system performance assessment.

A2. The closer a system's performance to the performance of others, i.e., the higher the *system conformity*, the more significant its contribution to the judgment of topic difficulty.

To measure topic discernment, we take topic ease $E_t[j]$ as the reference point and compute the dispersion of the system performance scores with respect to E_t:

$$W_t[j] = \Delta_t(\boldsymbol{\mathcal{P}}[*, j], E_t[j]), \quad (3)$$

where Δ_t is a dispersion operator, e.g., Euclidean distance function. Thus, a topic with higher dispersion will have a higher discernment coefficient W_t.

To measure system conformity, we calculate proximity between a given system's row vector and the vector of weighted averages for the topics, i.e., the *topic ease row vector* comprising $E_t[j]$ values for each topic. Using the topic ease vector as a reference point, we compute the system weight W_s as a departure of the system's performance from the topic ease vector:

$$W_s[i] = \Gamma_s(\boldsymbol{\mathcal{P}}[i,*], \boldsymbol{E_t}[*]), \quad (4)$$

where Γ_s is a *proximity* function. Thus, for each topic, we measure how close a given system performs to the weighted average performance across all systems. Another way to interpret this quantity is that it is a reflection of how well the system does on the easy topics.

Duality of topic discernment and system discernment. In an alternative evaluation setup, we may emphasize those systems and topics that have high discriminatory power among systems or topics, respectively:

B1. For topics, the situation is the same as in *A1*.

B2. System weights are based on how non-uniform the system's scores are. The more diverse a system' performance across topics, i.e., the higher the *system discernment*, the more significant the contribution of that system to the overall topic assessment. Thus, the weight $W_s[i]$, associated with individual systems, depends on the distribution of the system performance values $\mathcal{P}[i,*]$ around its mean.

To measure system discernment, reflecting how variedly a given system performs across the set of topics relative to its overall performance, $E_s[i]$, we use:

$$W_s[i] = \Delta_s(\boldsymbol{\mathcal{P}}[i,*], E_s[i]), \quad (5)$$

where Δ_s is a dispersion operator applied to the system's row vector.

The axioms *A1*, *A2*, *B1* and *B2* have been used in prior work, e.g., [8], and have typically also been given semantic meaning. For example, *A1* is used as a proxy for topic difficulty in query performance prediction [1]. Axiom *B2* can be seen as an indicator of system *volatility*. In this paper, we refrain from providing logical interpretations for the quantities represented by our axioms, choosing instead to illustrate the evaluation conclusions that would be reached from accepting a set of axioms to be true. Obviously, a range of combinations of different axioms are possible, each forming the underlying assumptions of an evaluation experiment. In the rest of the paper, we experiment with the axiom pairs *A1* & *A2*, and *B1* & *B2*.

4. EXPERIMENTS

We apply our method to seven TREC tracks to illustrate how the resulting ranking of systems and topics can be used to gain new insights into the nature of measures normally used in IR evaluation.

4.1 Data and Experiment Design

In our experiments we use the TREC performance statistics for the systems participating in the TREC 6-9 Ad hoc tracks and the TREC'04-'06 Terabyte tracks (Table 1).

Table 1. Datasets used in the experiments

Track	No. of runs	No. of topics
Adhoc TREC 6 (ta6)	56	50
Adhoc TREC 7 (ta7)	96	50
Adhoc TREC 8 (ta8)	116	50
Adhoc TREC 9 (ta9)	93	50
Terabyte 04 (tb4)	70	50
Terabyte 05 (tb5)	58	50
Terabyte 06 (tb6)	80	50

For axioms *A1* and *A2*, we use Equations (3) and (4) to derive the weights in Equations (1) and (2). The dispersion function Δ_t and the proximity function Γ_s are based on Euclidean distance:

$$W_t[j] = \sqrt{\sum_{\mathbf{i}} (\boldsymbol{\mathcal{P}}[*, j] - E_t[j])^2} \quad (7)$$

$$W_s[i] = 1 - \sqrt{\sum_{\mathbf{i}} (\boldsymbol{\mathcal{P}}[i,*] - \boldsymbol{E_t}[*])^2} \quad (8)$$

For axioms *B1* and *B2*, we use Equations (3) and (5) to derive the weights in Equations (1) and (2). The dispersion function Δ_s is given as:

$$W_s[i] = \sqrt{\sum_{\mathbf{i}} (\boldsymbol{\mathcal{P}}[i,*] - E_s[i])^2} \quad (9)$$

Unless we state otherwise, $\mathcal{M}$ is a weighted arithmetic mean.

4.2 Comparison with the HITS Algorithms

We first compare our results with the HITS method in [7] since there is an analogy between the authority of the systems A(s) and our system performance measure $\boldsymbol{E_s}$, as well as between the authority of the topics A(t) and our topic ease $\boldsymbol{E_t}$.

As suggested in [10], when constructing the matrix representing a systems-topics graph, we pre-process the data appropriately by subtracting the means of the respective quantities. We compare the A(s) and A(t) values to the equivalents in standard evaluation, i.e., MAP and AAP. The linear correlation, measured by the Pearson coefficient, is shown in the Table 2. Mizzaro and Robertson published results on the TREC 8 dataset, the third row in our table. Note that the number of systems they considered was a few more than ours—we eliminated 8 runs due to incorrect format of their input files. This resulted in the correlations being more than the published values for that dataset.

We see that our $\boldsymbol{E_s}$ and $\boldsymbol{E_t}$ values, which are the adaptive averages for system and topic performance, respectively, based on axioms *A1* and *A2*, are highly correlated with MAP and AAP, more so than the authority values A(s) and A(t).

Table 2. Correlation of system performance measures and weights, and corresponding topic-related quantities

Set	Pearson (MAP,E_s)	Pearson (MAP,A(s))	Pearson (AAP,E_t)	Pearson (AAP, A(t))	Pearson (MAP,W_s)	Pearson (AAP,W_t)
ta6	0.988	0.958	0.999	0.999	0.029	0.781
ta7	0.998	0.995	1	0.999	0.048	0.815
ta8	0.996	0.996	1	0.999	0.750	0.890
ta9	0.984	0.972	1	0.996	0.116	0.897
tb4	0.999	0.997	1	0.998	-0.223	0.933
tb5	0.999	0.999	1	0.999	0.565	0.768
tb6	0.998	0.991	1	0.998	0.579	0.758

What makes our averaging results different? Let us look at the quantities that we are mainly interested in, $W_s[i]$ and $W_t[j]$, because these influence $\boldsymbol{E_t}$ and $\boldsymbol{E_s}$, respectively. A topic with high $W_t[j]$, is meant to be indicative of one that helps distinguish amongst systems, while a system with high $W_s[i]$ will be indicative of doing well on easy topics (conformant). Note that

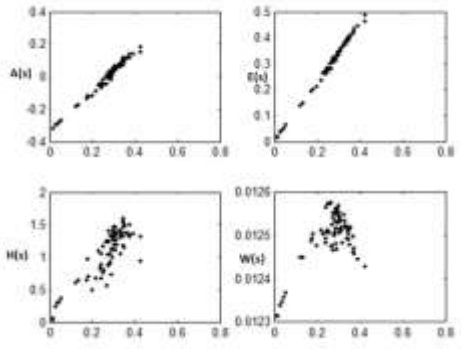

Figure 2a. Correlation of system authority and hubness (left) and system performance and conformity (right) with MAP (X axis) on the tb06

this follows directly from our axioms, which serve only as examples of possible evaluation setups.

Hubness in the systems-topics graph is highly correlated with MAP and AAP [7], see also Figures 2a and 2b. While our topic weights correlate with AAP, our system weights behave rather differently. Logically, a good system would get high values and a bad system would get low scores for most topics. Thus, looking at their scores is unlikely to help differentiate between the topics. It is the 'mid-range' systems that are therefore going to flag the topics as being easy or difficult.

The high correlation between AAP and $W_t[j]$, see also Figure 2b, indicates that 'easier topics', i.e., those for which all systems had comparatively higher AP values, are also those that get high $W_t[j]$ values. This may or may not be desirable, depending on a point of

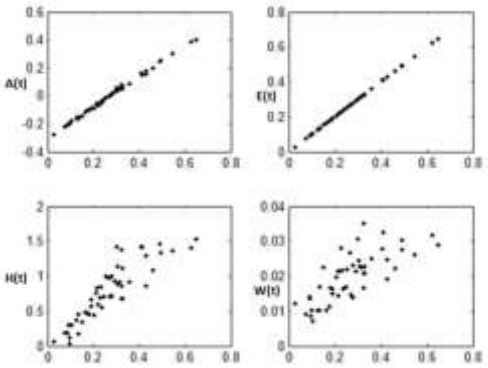

Figure 2b. Correlation of topic authority and hubness (left) and topic ease and discernment (right) with AAP (X axis) on tb06

view. Mizzaro and Robertson deal with this by changing the metric for system effectiveness, going from MAP to GMAP. In the next section we show how to address this through the appropriate selection of the mean function $\mathcal{M}$.

Figure 3 illustrates the relationship between the raw AP values of the systems-topics matrix and the derived system and topic weights. The matrix of AP scores was arranged to have the best (according to MAP) system at the top and the easiest (according to AAP) topic on the left. The upright column bars in color are the system oriented measures - MAP, A(s) and $\boldsymbol{E_s}$ calculated using the arithmetic and geometric means. In the plots, red indicate

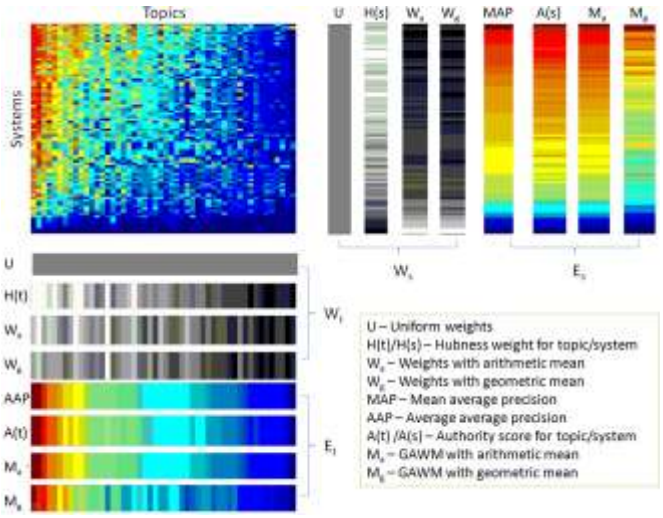

Figure 3. Heatmaps of the raw system-topic matrix and the GAWM weights and means

larger values than blue. The MAP and AAP monotonically move from red to blue, while the shuffling of colors for the other quantities indicate that they diverge from MAP and AAP respectively. The corresponding weight vectors, $\boldsymbol{W_s}$ and $\boldsymbol{W_t}$, provided in grayscale, use white to indicate systems/topics that contribute more to the performance scores of the topics/systems correspondingly. For comparison, we have also provided the uniform weighting used for MAP and AAP computations. Again, the shuffled greyness indicates the absence of a direct relationship between MAP & AAP and the corresponding weights.

4.3 Impact of the Mean Function

Using the weighting functions W_t and W_s (here we use w) defined in Equations (3) and (4), we now experiment with five forms of the generalized mean function $\mathcal{M}$:

1) $\mathcal{M}(\vec{x},\vec{w}) = Min(x_i)$, **minimum value** in the given set,
2) $\mathcal{M}(\vec{x},\vec{w}) = \frac{1}{\sum_{i=1}^{N} \frac{w_i}{x_i}}$, the **harmonic mean**,
3) $\mathcal{M}(\vec{x},\vec{w}) = \prod_{i=1}^{N} {x_i}^{w_i}$, the **geometric mean** (where $\sum_{i=1}^{N} w_i = 1$),
4) $\mathcal{M}(\vec{x},\vec{w}) = \sum_{i=1}^{N} x_i w_i$, the **arithmetic mean** as used before,
5) $\mathcal{M}(\vec{x},\vec{w}) = Max(x_i)$, **maximum value** in the given set.

Strictly speaking, the minimum and maximum are not averaging methods. We include them as they represent an interesting testing criterion: should a system be represented by its best or worst performance, rather than its average? From the pseudo-code in section 3.2, we see that the mean function controls $\boldsymbol{E_s}$ and $\boldsymbol{E_t}$. Due to the circular definition, the dispersion quantities Δ_t and Γ_s will also change to reflect the altered pivot point with respect to which distance it is calculated.

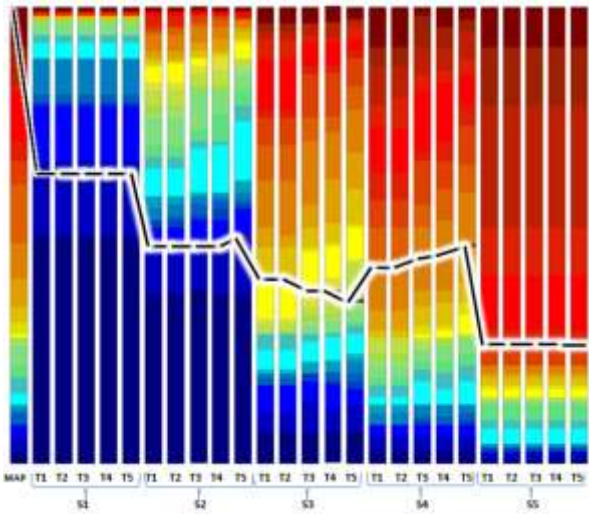

Figure 4. Tracking the leader system (according to MAP) across different evaluation settings. The heatmaps represent overall system performance for different combinations of the mean functions 1) to 5), used to obtain $\boldsymbol{E_s}$ (S_1 to S_5) and $\boldsymbol{E_t}$ (T_1 to T_5).

Figure 4 illustrates how the overall performance of a system changes across different evaluation settings, when different mean functions are used for topics and systems. Each column is a ranking of all the systems based on their overall performance scores, $\boldsymbol{E_s}$, for a given evaluation setup. For example, the third column corresponds to the systems rankings obtained by our GAWM method, based on axioms *A1* and *A2* and using the minimum for the system mean ("S1") and the harmonic mean for topics ("T2"). The first column on the left is the system rankings according to MAP. The colour in the heatmap reflects the absolute values of the overall performance scores, scaled by the maximum in a given column: dark red is highest value, dark blue is lowest. The black line traces the rank position of the system that was declared as best by MAP across the different combinations of mean functions (5×5) in our GAWM method. Predictably, different ways of defining 'the best system' (i.e., the mean function for systems) gives different 'best systems'. The mean function for the topics seems less important. Interestingly, the worst system according to MAP doesn't go up more than 5 positions in any list.

Table 3. System and topic weights for different mean functions. Results shown on the tb06 dataset

Data	Generalised Mean for Topics	Pearson (MAP, W_s)	Pearson (AAP, W_t)
tb6	Minimum	-0.995	**0.992**
tb6	Harmonic Mean	-0.982	0.838
tb6	Geometric Mean	-0.346	0.705
tb6	Arithmetic Mean	0.579	0.758
tb6	Maximum	**0.997**	0.096

Next we concentrate our analysis on $W_t[j]$ and $W_s[i]$, because they are the weights that indicate how discriminatory or conformant the corresponding objects are. Since the HITS algorithm does not allow for different means, we only present the GAWM metrics in Table 3. Note that using the geometric mean makes our experiments comparable with [7] when they took GMAP as the metric for system effectiveness. We only show results for the tb06 dataset and for the setting when the system mean is fixed as the arithmetic mean function (corresponding to the S_4 heatmaps in Figure 4). While the results do vary depending on the chosen system mean, they remain fairly similar across the different datasets for the same combinations of the topic and system means. This fact is influenced by the proportion of values in the performance matrix $\boldsymbol{\mathcal{P}}$ that are close to zero. Some combinations of means can have little effect on the results, e.g. the use of the minimum and maximum breaks the cyclical dependence between system and topic ($\boldsymbol{E_s}$,$\boldsymbol{W_t}$) and ($\boldsymbol{E_t}$,$\boldsymbol{W_s}$) vectors.

We see that using the Maximum gives the highest correlation between MAP and $\boldsymbol{W_s}$, while using the minimum gives the best correlation between AAP and $\boldsymbol{W_t}$. This suggests the following principles:

a) Comparing systems to the best tells us which systems to weight/trust more,
b) Comparing topics to the hardest tells us how much we should value performance on each topic.

4.4 Confidence in the GAWM Ranking of Systems

If we view TREC as a competition where a winning system is selected, it is desirable that this winning system is picked in a

reliable manner. The current practice is to order systems based on MAP, where the system with the highest value is declared the best. The HITS method and our GAWN method both provide alternate averaging criteria. How would the result of the TREC competition differ if these new approaches were adopted?

An initial indication of the answer to this question is already available from the experiments of the previous section (see Figure 4). We see that the best system, chosen according to MAP, drops to a position as low as rank 85 when using different criteria (combination of mean and distance functions) to declare winners. We need, thus, to verify that we are confident in our choice of the best system.

To do this, we follow on from the analysis of Webber et al in [14]. They found that due to variations in a given system's performance across topics, an ordering of systems based on MAP does not conclusively separate out the best system from the rest. We attempt to verify the corresponding behavior in the context of systems chosen by GAWM.

Each system S_i is characterized by a Gaussian distribution with mean $\tilde{s}_i$ and variance $\tilde{v}_i$ calculated from their AP scores. Given two such systems, S_i is better than S_j is given by

$$p(S_i > S_j) = p(S_i - S_j > 0)$$

$$= \int_0^\infty \mathcal{N}\left(s;\ (\tilde{s}_i - \tilde{s}_j), (\tilde{v}_i + \tilde{v}_j)\right) ds$$

We choose three reference systems – the best, the worst, and the median – as illustrative examples. We calculate the probability that the system deemed to be the best is actually better than the system picked as the worst, and similarly, we compare between the best system and the median.

Figure 5 shows the results for the TREC 8 Ad hoc track, when using weighted arithmetic mean for topics, and varying the system mean function. As it can be seen, across all choices of mean functions, using adaptive weights provides a more confident choice of best system than using MAP.

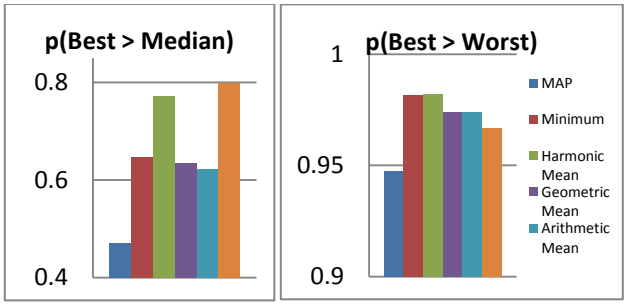

Figure 5. Probability that the best system picked by MAP and GAWM (for 5 mean functions) is actually better than the median system (left) or the worst system (right).

4.5 Stability Analysis

As described so far, GAWM internally calibrates two quantities whose interaction is represented by scores in a matrix. In the context of IR evaluation and the TREC initiative, the objective is to obtain a standardized evaluation setup (i.e., including a corpus, a topic set, and relevance judgments) for future evaluations. Ideally, no bias should be introduced at the time of test collection creation, because this might disadvantage those systems that did not participate in the original competition. In this section, we conduct experiments to measure the sensitivity of our topic and system metrics. In order to instantiate a specific evaluation setup, we show results when following axioms *B1* and *B2* (see Section 3.4) and use all seven datasets (see Table 1).

4.5.1 Leave One Topic Out Experiment

For these experiments, we interpret MAP and our E_s measure as a system's best guesses of AP for an unseen query. We test the predictive power of MAP and E_s calculated based on different setups. From the initial AP matrix of *m* systems and *n* topics, we remove one topic. We calculate both MAP and E_s for each system as usual, but over (*n-1*) topics. We measure how well these orderings agree with that induced by the held-out topic alone.

We set the topic mean function to be the arithmetic mean and experimented with different system mean functions. The results, calculated as an average over the correlations obtained by holding out each topic in turn, and over all seven datasets, are given in Figure 6. In summary, we find that our measures are *not* as predictive as MAP for future effectiveness. Thus, the calibration methods described in this paper seem better suited for *closed worlds*, i.e., when all systems and topics are available.

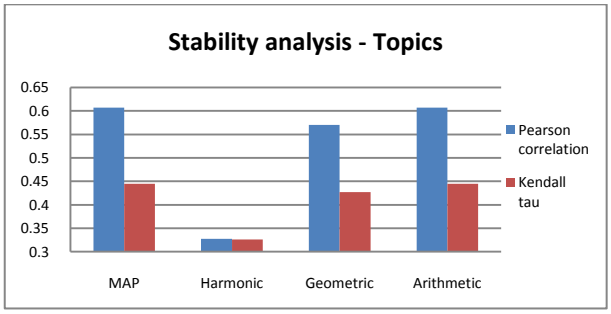

Figure 6. Average correlation over *n* runs with (*n-1*) topics, and 7 datasets (?).

4.6 Experimental Findings

Due to the way the TREC testing process is setup, there are likely to be inefficiencies. For example, the topics are first established before participating teams provide systems that generate retrieval results from the frozen document collection. It is therefore conceivable that none of the participating systems provides satisfactory results from the data collection for a given query, not because the systems are ineffective but potentially because there was a mismatch between the query and the documents in the collection.

Such effects will be factored into any post-averaging process that attempts to infer the effectiveness of a query by looking across systems. We could therefore be more lenient when it comes to punishing systems for low AP values on such topics. Both the HITS method and ours provides such an alternative.

A major advantage of our methodology is that the underlying assumptions of the evaluation experiment are made explicit through the choice of axioms. This in turn defines the choice of mean and distance/dispersion functions. It is important to recognize that there is unlikely to be a general purpose best-practices list that can be used for evaluations in all contexts. Having a framework that allows for a wide range of applications is advantageous, and danger of incorrect usage is mitigated by the fact that all components used need to be individually declared.

To illustrate the framework, we formalized some commonly found IR heuristics ([1][5]). Analyzing historical TREC data through these rules indicates that mutually calibrated topic and system scores could provide more stable and reliable effectiveness measures. And, incorporating new/other heuristics, is trivial, as long as they can be defined in our framework. However, a shortcoming of the methodology is that it relies on a closed system, i.e., it provides limited evidence of future behavior of topics and systems. This should be evident given that all definitions of system and topic effectiveness were internal.

5. SUMMARY AND FUTURE WORK

Benchmarking tests in IR compare systems' performances across a set of test topics. The standard TREC measures involve simple arithmetic means of the system performance according to a predefined measure across the test topics. However, it has been observed that topics vary and are considered more or less difficult depending on how well the systems perform on them. There have been several attempts to incorporate topic characteristics into the evaluation and comparison of systems. However, none of these efforts managed to provide a coherent and generalized framework that subsumes the standard methods and covers a broad class of evaluation measures.

Starting with a pair of axioms that postulate the relationship between the performance of systems and topics, we define the Generalized Adaptive-weight Mean (GAWM) as a unified model which incorporates the system performance vector $\boldsymbol{E_s}$ and the system conformity weights $\boldsymbol{W_s}$ to characterize systems, and the topic ease vector $\boldsymbol{E_t}$ and the topic discernment weights $\boldsymbol{W_t}$ to characterize topics. These quantities are obtained by solving the system of equations iteratively. Thus, both topic and system weights adapt to the raw evaluation data.

Based on the mathematical formulations, we find similarities with the HITS method proposed in [7]. The GAWM subsumes HITS as a special case. It enables us to vary the form of the generalized mean function and therefore specify different criteria for system comparison and improvement.

The GAWM approach is generic and can be used in other evaluation contexts, such as TAC (Text Analysis Conference: www.nist.gov/tac/). Furthermore, the GAWM framework can be applied directly to the ranking and scoring of retrieved results and formulated to capture the characteristics of systems, topics, and documents. Generally, the method opens new possibilities for modeling cyclical relationships in closed systems where values and measurements are defined in a relative rather than absolute sense.

In principle, we believe in the work of Webber et al [14], being able to track system evaluations of over time and across collections is the only way to reassure ourselves that ranking methods are getting becoming more reliable. Our averaging methods are agnostic of the contents of the original raw score matrix, therefore it could well be populated with the standardized scores. Cross-collection calibration can then be achieved, provided there is some overlap (in either systems or topics). Extending our method to achieve this aim is left to future work.

6. REFERENCES

[1] Aslam, J. and Pavlu, V. 2007. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Proceedings of ECIR*, 198-209.

[2] Buckley, C. and Voorhees, E. 2000. Evaluating evaluation measure stability. In *Proc. of SIGIR*, 33-40.

[3] Carmel, D., Yom-Tov, E., Darlow, A., and Pelleg, D. 2006. What makes a query difficult?. In *Proc. of SIGIR*, 390-397.

[4] Cronen-Townsend, S., Zhou, Y., and Croft, W. B. 2002. Predicting query performance. In *Proc. of SIGIR*, 299-306.

[5] Diaz, F. 2007. Performance prediction using spatial autocorrelation. In *Proc. of SIGIR*, 583-590.

[6] Ignjatović, A., Lee, C. T., Kutay, C., Guo H. and Compton, P. 2009. Computing Marks from Multiple Assessors Using Adaptive Averaging. In *Proc. of ICEE & ICEER*.

[7] Mizzaro, S. and Robertson, S. 2007. HITS hits TREC: exploring IR evaluation results with network analysis. In *Proc. of SIGIR*, 479-486.

[8] Mizzaro, S. 2008. The Good, the Bad, the Difficult, and the Easy: Something Wrong with Information Retrieval Evaluation? In *Proc. of ECIR'08*.

[9] Rijsbergen, C. J. van. Information Retrieval, Butterworths, London, 1979.

[10] Robertson, S. 2006. On GMAP: and other transformations. In *Proc. of CIKM*, 78-83.

[11] Vinay, V., Cox, I. J., Milic-Frayling, N., and Wood, K. 2006. On ranking the effectiveness of searches. In *Proc. of SIGIR*, 398-404.

[12] Voorhees, E. M. and Harman, D. K. 2005 *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. The MIT Press.

[13] Voorhees, E. M. 2006. The TREC 2005 robust track. *SIGIR Forum* 40(1), 41-48.

[14] Webber, W., Moffat, A., and Zobel, J. 2008. Score standardization for inter-collection comparison of retrieval systems. In *Proc. of SIGIR*, 51-58.

[15] Zhou, Y. and Croft, W. B. 2006. Ranking robustness: a novel framework to predict query performance. In *Proc. of CIKM*, 567-574.