Measuring System Performance and Topic Discernment using Generalized Adaptive-Weight Mean

Chung Tong Lee[†]
ctlee@cse.unsw.edu.au

Vishwa Vinay[‡]
vvinay@microsoft.com

Eduarda Mendes Rodrigues[‡]
eduardamr@acm.org

Gabriella Kazai[‡]
gabkaz@microsoft.com

Nataša Milić-Frayling[‡]
natasamf@microsoft.com

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

[†]School of Computer Science and Engineering
University of New South Wales
Sydney, 2052, Australia

[‡]Microsoft Research
7 JJ Thomson Avenue
Cambridge, CB3 0FB, UK

ABSTRACT
Standard approaches to evaluating and comparing information retrieval systems compute simple averages of performance statistics across individual topics to measure the overall system performance. However, topics vary in their ability to differentiate among systems based on their retrieval performance. At the same time, systems that perform well on discriminative queries demonstrate notable qualities that should be reflected in the systems' evaluation and ranking. This motivated research on alternative performance measures that are sensitive to the discriminative value of topics and the performance consistency of systems. In this paper we provide a mathematical formulation of a performance measure that postulates the dependence between the system and topic characteristics. We propose the Generalized Adaptive-Weight Mean (GAWM) measure and show how it can be computed as a fixed point of a function for which the Brouwer Fixed Point Theorem applies. This guarantees the existence of a scoring scheme that satisfies the starting axioms and can be used for ranking of both systems and topics. We apply our method to TREC experiments and compare the GAWM with the standard averages used in TREC.

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
Benchmarking of information retrieval (IR) systems has been largely shaped by the Text REtrieval Conference (TREC) ([11], [12]), an evaluation initiative organized by the National Institute of Standards and Technology (NIST). The TREC evaluation of IR systems follows a *collective 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.

A crucial step in the benchmarking exercise is the design of tests to measure a particular aspect of the systems. In IR, the tests are expressed as topics, predefined as part of the experiment design, and their outcomes are observed through the measures of retrieval performance that typically reflect the precision and recall achieved by the system for a given topic. Particularly convenient are *single-value retrieval measures*, such as Average Precision, that can be computed for individual topics and then combined across topics to characterize the overall system performance [11].

In recent years, there have been attempts to characterize a *topic's difficulty,* reflected through the retrieval performances 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],[2],[7]). However, this and similar relationships between topics' and systems' characteristics have not been modeled systematically. In this paper we seek a principled way to express the dependence between the topics' properties and the systems' performance and incorporate these into overall performance measures. We develop the *Generalized Adaptive-Weight Mean* (GAWM) measure and present a unified model for system evaluation using GAWM, where the weights reflect the ability of the test topics to differentiate among the retrieval systems.

2. BENCHMARKING IN IR
TREC includes a number of independent tracks which are focused on specific IR tasks and involve the design of appropriate test collections and evaluation measures. As an example, let us consider the *ad hoc query* track and its evaluation procedure:

- Participants are provided a data set, e.g., a collection of documents and a set of test topics. For each topic, they need to return to TREC organizers a list of retrieved documents.
- From each set of submitted retrieval results the TREC organizers select the top N ranked documents to arrive at the pool of documents that will be manually judged.
- The collected relevance judgments are used to calculate the performance metrics for each system and topic pair. The

CIKM'09, November 2–6, 2009, Hong Kong, China.

most commonly used single-valued metrics are Average Precision (AP) and R_Precision [11].

- The overall system performance is typically characterized by the mean value of per-topic performance, e.g., the Mean AP value (MAP), which is then used to compare the systems.

However, there are serious concerns with the use of simple averages to compare and rank systems. First, the mean is not regarded a good descriptor when the distribution of measurements is not normal, i.e., bell-shaped ([8], p.176), as it is the case with the Average Precision values across topics. Second, the simple mean treats all the test topics in the sample equally.

In order to deal with the skewed distribution of performance metrics, Webber et al. [13] investigate the use of *score standardization*. For each system, they adjust the performance score for a given topic by the mean and the standard deviation of performance scores for that topic achieved across a sample of systems. With regards to the topic differentiation, van Rijsbergen [8] 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 the work by Robertson [9], the topic differentiation is achieved indirectly through the use of the geometric mean of the AP values (GMAP), which is sensitive to the low AP values, i.e., the topic difficulty.

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 the *topic discernment* and the *system performance*. We begin by reflecting on related work that considers the issues of topic difficulty and coupling of topic characteristics and system performance.

2.1 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. Several approaches have been taken to estimate topic difficulty and predict system performance. Examples include KL-divergence in Cronen-Townsend et al. [3], the Jensen-Shannon divergence in Carmel et al. [2], document perturbation in Vinay et al. [10], and robustness score by Zhou and Croft [14]. 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. This circular definition of topic difficulty and system performance has not been explicitly modeled in the retrieval evaluation.

The work by Mizzaro and Robertson [7] relates the 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

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

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

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 [4]. Aslam and Pavlu [1] assume that those topics for which the systems deviate from the average performance across systems are difficult. Diaz [4] focuses on the system performance and argues that the systems that deviate from others on individual topics are suboptimal. Both papers use the departure from the average performance to rank topics and systems, respectively, and aim at predicting retrieval performance for new topics.

2.2 Multi-Grader Problem

We observe that the performance metrics and the need for characterizing both the systems and the 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 the 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 topics, or systems, that capture the differences among the graders. Ignjatović et al [6] 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 system performance 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. We start with the axioms on which we base our model.

3.1 Axiomatic Descriptions

Consider a set of topics and systems. The topics are designed to test the performance of the systems and, thus, differentiate them based on a pre-defined measure of performance. Most systems will process many of the topics with similar success. However, some topics will cause systems to perform very differently, leading to a wider range of performance values. These topics are considered good discriminators and thus desirable from the test design point of view. We attach more weight to these topics:

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.

On the other hand, a system performing closer to the average or other expressions of consensus across systems is more reliable and should get more weight when assessing the difficulty of a topic:

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.

3.2 System Performance and Conformity, Topic Ease and Discernment

Given a set of *n* topics and *m* systems, we consider a real-valued 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 use Average Precision (AP) or R_Precision as entries of $\boldsymbol{\mathcal{P}}$ and assume that higher values of $\mathcal{P}[i,j]$ correspond to better performance.

By considering the *i*-th row vector of $\boldsymbol{\mathcal{P}}$, i.e., $\boldsymbol{\mathcal{P}}[i,*]$, we define the overall *system performance* of the system s_i as a *weighted mean* $E_s[i]$ of the per-topic values in $\boldsymbol{\mathcal{P}}[i,*]$. In practice, it is common to use a simple average, i.e., the 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 a weighted average over the topics:

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

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

Similarly, we consider the performance scores for the topic across systems and seek to determine 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. The higher the value of $E_t[j]$, the better the performance of systems for topic t_j and hence the easier the topic. Thus, $E_t[j]$ can be viewed as a measure of *topic ease*. 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.

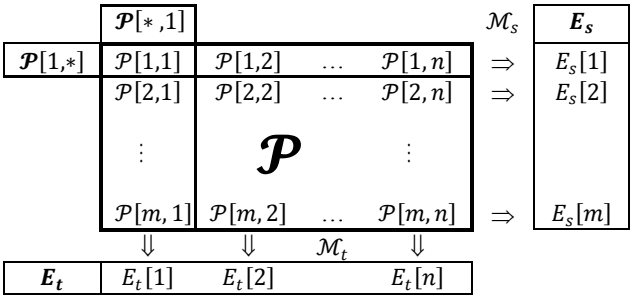

Figure 1. Performance matrix comprises system (rows) performance statistics for individual topics (columns). Aggregation of system statistics $\boldsymbol{E_s}$ reflects the system performance and the aggregation of topic statistics $\boldsymbol{E_t}$ reflects topic ease.

We now use the axiom A1 and A2 to define relationships among the concepts we have introduced. The topic weight $W_t[j]$ in (1), associated with individual topics, is aimed to measure topic discernment among the systems and, thus, its value should depend on the distribution of the topic performance values $\mathcal{P}[*,j]$ across systems. Thus, we take the 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 the dispersion operator. Thus, the topic with higher dispersion will have a higher discernment coefficient W_t.

Consistent with the Axiom A2, we stipulate that the system weights W_s in (2) relate to the system conformity. In order to measure the system conformity we define a *topic ease vector* comprising $E_t[j]$ values for each topic, i.e., a row vector that includes topic ease for each topic. We use the topic ease vector as a reference point and compute the system weight W_s as a departure of the system 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.

By concatenating the system and the topic vectors $\boldsymbol{E_s}$ and $\boldsymbol{E_t}$ to form $\boldsymbol{E}$ and weight vectors $\boldsymbol{W_s}$ with $\boldsymbol{W_t}$ to form $\boldsymbol{W}$, and by combining $\mathcal{M}_s$ with $\mathcal{M}_t$ into $\mathcal{M}$ and Δ_s with Γ_t into Ψ, we arrive at a system of equations

$$\boldsymbol{E} = \mathcal{M}(\boldsymbol{\mathcal{P}}, \boldsymbol{W}) \quad (5)$$

$$\boldsymbol{W} = \Psi(\boldsymbol{\mathcal{P}}, \boldsymbol{E}) \quad (6)$$

that shows the coupling of $\boldsymbol{E}$ and $\boldsymbol{W}$. The equations (5)-(6) are a generalization of the system in [7] with $\boldsymbol{E}$ being the counterpart of the authority and $\boldsymbol{W}$ of the hubness in the Systems-Topics graph.

3.3 Fixed Point Theorem

The circular definition of $\boldsymbol{E}$ and $\boldsymbol{W}$ can be viewed as a mapping of the Euclidean *k*-space $\mathbb{R}^k$ into itself where $k = m + n$. By substituting (6) into (5) we note that $\boldsymbol{E}$ is, in fact, a fixed point of the mapping $\mathcal{F}: \boldsymbol{E} \mapsto \mathcal{M}\big(\boldsymbol{\mathcal{P}}, \Psi(\boldsymbol{\mathcal{P}}, \boldsymbol{E})\big)$, i.e., $\boldsymbol{E} = \mathcal{F}(\boldsymbol{E})$.

Brouwer Fixed Point Theorem (BFPT) guarantees the existence of fixed point of a *continuous* mapping of a *closed*, *bounded convex* set in $\mathbb{R}^k$ into itself [5]. The space is obviously closed and bounded as these properties are inherited from $\mathcal{P}$. Since in our application the values of $\boldsymbol{\mathcal{P}}$ are bounded, we have the hypercube $[\min(\boldsymbol{\mathcal{P}}), \max(\boldsymbol{\mathcal{P}})]^{m+n}$., which is a convex set. The choice of the functions $\mathcal{M}$ and Ψ can ensure that the mapping is continuous. As a result, $\mathcal{F}$ is continuous on this *closed*, *bounded*, *convex* set and we can apply BFPT. Fixed point existence is guaranteed.

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

We compare our results with the HITS method [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 $\boldsymbol{E_t}$. For $\mathcal{M}$ we use a weighted arithmetic mean. The dispersion function Δ_t and the proximity function Γ_s are based on Euclidean distance $\| \ \|_2$, with a real-value spreading factor q:

$$W_t[j] = \|\boldsymbol{\mathcal{P}}[*, j] - E_t[j]\|_2 \tag{7}$$

$$W_s[i] = \left(1 - \frac{\|\boldsymbol{\mathcal{P}}[i,*] - \boldsymbol{E_t}[*]\|_2}{\sum_{i=1}^{m} \|\boldsymbol{\mathcal{P}}[i,*] - \boldsymbol{E_t}[*]\|_2}\right)^q \tag{8}$$

We present the results of the hub and authority algorithms alongside, our GAWM performance measure (Table 2).

4.2 Comparison with the HITS Algorithms

Following the procedure in [7], we pre-process the data by subtracting the means of the respective quantities and construct the matrix representing the Systems-Topics graph. We compare the A(s) and A(t) values to the equivalent standard metrics, i.e., MAP and AAP, respectively. The linear correlation, measured by the Pearson coefficient, is shown in Table 2. Mizzaro and Robertson [7] published results on the *trec8* dataset. The third row in Table 2 (ta8) shows our HITS results for the same runs. The number of systems considered in [7] was slightly larger since we had to eliminate eight systems due to incorrect format of files with performance data. For comparison, we show correlations of $\boldsymbol{E_s}$ and $\boldsymbol{E_t}$ with MAP and AAP, noting that the Pearson coefficients are high but lower than that for A(s) and A(t).

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

Data	Pearson (MAP,$\boldsymbol{E_s}$)	Pearson (MAP,A(s))	Pearson (AAP,$\boldsymbol{E_t}$)	Pearson (AAP, A(t))
ta6	0.904	0.958	0.867	0.999
ta7	0.952	0.995	0.956	0.999
ta8	0.959	0.996	0.882	0.999
ta9	0.899	0.972	0.737	0.996
tb4	0.972	0.997	0.925	0.998
tb5	0.941	0.999	0.924	0.999
tb6	0.971	0.991	0.963	0.998

5. SUMMARY AND FUTURE WORK

Benchmarking tests in IR compare systems performance across a set of test topics. The standard TREC measures involve simple arithmetic means of the system performance according to a pre-defined measure across the test topics. However, it has been observed that topics vary and are more or less difficult depending on how well the systems performed 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 two 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 $\boldsymbol{E_s}$ and the system conformity weight $\boldsymbol{W_s}$ to characterize systems, and the topic ease $\boldsymbol{E_t}$ and the topic discernment weight $\boldsymbol{W_t}$ to characterize topics. These quantities are obtained by computing the fixed-point of a well behaved function. The topic and the system weights thus adapt to the set of experiments that are included in the evaluation.

Based on the mathematical formulations, we find similarities with the HITS method proposed in [7]. The GAWM subsumes HITS as a special case. It also 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), where a variety of different metrics are used to assess the quality of document summaries. 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.

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

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

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

[5] Griffel, D. H. *Applied Functional Analysis*. Dover Publications, Jun 2002.

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

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

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

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

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

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

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